Holistic Satellite Image Super Resolution Using Large Diffuse Generative Models

Vladimir V. Kniaz [a,b], Petr V. Moshkantsev [a], Victor S. Aleksandrov [a], Artem N. Bordodymov [a], Vladimir A. Knyaz [a,b], Egor R. Smirnov [a]

[a] State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia - vl.kniaz@gosniias.ru
[b] Moscow Institute of Physics and Technology (MIPT), Moscow, Russia - kniaz.vv@mipt.ru

Key Words: remote sensing, satellite imagery, deep learning, image super resolution.

Abstract

Satellite image super-resolution (SR) is a critical task in machine vision, supporting applications such as urban planning, agricultural monitoring, and disaster response. The objective of SR is to enhance the spatial resolution of low-resolution (LR) satellite imagery, thereby extracting finer detail. Over the past two decades, numerous SR methods have been developed, primarily leveraging deep learning to achieve significant resolution enhancements. However, prevailing approaches exhibit two principal limitations. First, many methods operate via analytic mathematical techniques or rely on learned priors from training data, without explicitly incorporating structured scene knowledge (e.g., road networks, building footprints). Second, existing methods offer minimal control over the aesthetic and structural characteristics of the output during the SR process, which is a drawback for applications like digital map updating where consistency with existing geospatial databases is essential. In scenarios where rich prior semantic information about a scene is available—such as from OpenStreetMap or other GIS sources—it is natural to harness this data to guide and improve SR reconstruction.
This paper introduces a semantic super-resolution (`SSR`) model designed to address these gaps by integrating semantic priors directly into the SR pipeline. The approach builds upon the `Flux Kontext` diffusion-based generative framework, incorporating two key modifications. Firstly, an additional textual encoder is trained to convert available semantic information—such as vector data describing roads, buildings, and water bodies—into a dense textual prior that is fed into the network. Secondly, the loss function is refined via a Low-Rank Adaptation (LoRA) adapter, emphasizing semantic consistency of major features in the generated high-resolution (HR) image.
The proposed `SSR` model was evaluated against three state-of-the-art deep learning SR baselines: EDSR, SRGAN, and ResShift, using the high-resolution aerial imagery dataset (*HRAID*) comprising 236 images. Qualitative assessment reveals enhanced consistency in the reconstruction of structured features like roads and buildings, attributable to the injected textual prior. Quantitative evaluation demonstrates that our model outperforms the best-performing baseline by 5% in PSNR and by 10% in FID. An ablation study systematically removing components of the `SSR` model confirms the necessity of both the textual encoder and the semantic-consistency-focused LoRA adapter for achieving these gains.

1. Introduction

Satellite image super-resolution (SR) is an important task in machine vision for applications such as urban planning, agriculture monitoring, disaster management, and wildfire management. Over the last two decades, several super-resolution techniques have been developed to process satellite imagery and increase its resolution. Specifically, super-resolution involves the process of converting a low-resolution (LR) image into its high-resolution (HR) counterpart. The problem of enhancing the resolution of images has attracted attention of the scientific community for decades, starting with mathematical interpolation methods and coming now to deep learning techniques that have made significant progress in producing high quality high-resolution images from low resolution input data.

Applying a data-driven approach to the super-resolution problem by mapping low-resolution images to high-resolution ones made a dramatic improvement in output results. Enhanced deep residual network models (EDSR (Lim et al., 2017)) and analogous architectures provided a notable step forward in the quality of generated images. Moreover, the generative adversarial approach (such as the SRGAN network model (Ledig et al., 2017) and the widely adopted ESRGAN (Wang et al., 2018)) to image resolution enhancement allowed reaching better performance compared with traditional convolutional neural networks. Introducing the adversarial loss in competitive learning of generator and discriminator network models allowed surpassing the quality achieved by CNNs. Further advances in the image super-resolution problem have been achieved by introducing the attention mechanism into the learning process, enabling networks to accurately recover details for complex imagery. More recently, diffusion-based generative models have set a new state-of-the-art, offering unprecedented fidelity in image synthesis and restoration tasks.

While modern techniques allow for processing arbitrary images and improving their resolution by up to tenfold, they still face several challenges. Firstly, most methods rely solely on prior information about the scene's structure that was implicitly learned from the training dataset, without explicitly incorporating structured scene knowledge such as road networks or building footprints. Secondly, modern techniques offer minimal control over the appearance and structure of an image during super-resolution process. This is a significant drawback for applications like digital map updating, where consistency with existing geospatial databases is essential.

In tasks such as updating a digital map, a rich amount of prior information is available. Such information regarding the structure of a scene – roads, building footprints, or water object boundaries – can be readily obtained from GIS sources such as OpenStreetMap. It seems natural to harness this prior information to guide and improve the super-resolution process. The object-

ive of this work is therefore to develop a super-resolution model that explicitly integrates semantic priors, thereby providing both improved reconstruction quality and controllability over the output.

Over the past two decades, super-resolution has evolved from classical interpolation and reconstruction-based methods to modern deep learning paradigms, including residual convolutional networks, generative adversarial networks, attention-based architectures, and, most recently, diffusion models. A parallel line of research explores the use of auxiliary priors—semantic maps, textual descriptions, and structural cues—to condition the generation process.

Residual learning approaches such as EDSR (Lim et al., 2017) demonstrated that deeper networks with optimized residual blocks yield substantial gains in reconstruction accuracy. Adversarial frameworks like SRGAN (Ledig et al., 2017) and ESRGAN (Wang et al., 2018) shifted the focus towards perceptually convincing textures rather than pixel-wise fidelity alone. Diffusion-based methods, exemplified by the Semantic Guided Diffusion Model (SGDM) (Wang and Sun, 2025), have recently delivered high-quality restorations by exploiting a pre-trained generative prior together with vector-map guidance for large-scale-factor remote sensing super-resolution. In the domain of controllable generation, foundation models such as `Flux Kontext` (Black Forest Labs et al., 2025) enable conditioning image synthesis on rich contextual inputs, opening new possibilities for prior-guided super-resolution.

In this paper, we propose a new model called semantic super-resolution (`SSR`), which addresses the two aforementioned problems by integrating a semantic prior into the network input. As a starting point, we build on the Semantic Guided Diffusion Model (SGDM) (Wang and Sun, 2025), and make two contributions to it. Firstly, we replace the diffusion generative prior with the more recent `Flux Kontext` model (Black Forest Labs et al., 2025), and add a `SmolLM` (Allal et al., 2025) language module that converts the input semantic segmentation into a rich textual prior, which is then passed to the text-conditioning pathway of the network. Secondly, we introduce a semantic consistency loss that emphasizes the semantic consistency of features in the processed images, penalizing deviations between a segmentation of the reconstruction and the ground-truth semantic map.

This approach allows us to control the super-resolution process by modifying the semantic input, applying different modifications to reconstructed images and their fine details. Furthermore, it enables us to drastically change the appearance of resulting images, such as changing seasons, so that satellite images can be made to mimic the seasons and colors of a larger satellite mosaic. As a result, we are able to generate seamless satellite mosaics. We evaluate our approach on two datasets: the publicly available Cross-Modal Super-Resolution Dataset (CMSRD) (Sarmad et al., 2024), and our custom-collected *QuattroStagioni* dataset, which contains 5k samples of high-resolution satellite imagery. Each *QuattroStagioni* sample consists of four aligned satellite images of the same place captured in the four seasons (winter, spring, summer, autumn), the corresponding ground-truth semantic annotation with 10 object classes, a textual prompt describing the semantic structure of the scene, and generated low-resolution versions of the original imagery. The semantic annotations are enriched with readily available OpenStreetMap data.

We compare our semantic super-resolution model to two modern deep learning baselines for image super-resolution: SGDM (Wang and Sun, 2025) and ESRGAN (Wang et al., 2018). We train all baselines and our model on the training splits of the CMSRD and *QuattroStagioni* datasets and evaluate them on the corresponding test splits. Evaluation is performed using peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and three deep-learning-based metrics: Fréchet inception distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), and CLIP-IQA. Through a rigorous evaluation, we demonstrate that our model attains the best PSNR, FID, and CLIP-IQA scores while remaining competitive on the remaining metrics. Specifically, our model provides greater consistency in the structure of roads and buildings, which we attribute to the incorporation of semantic information available in the textual prompt, and it also achieves the most stable semantic structure across seasons. Finally, we provide an ablation study that sequentially eliminates various parts of our model and demonstrates the necessity of each component in the whole model setup.

The developed semantic super-resolution model can be applied to various important tasks, such as the updating of satellite maps or area-consistent and border-consistent super-resolution for the generation of large-scale satellite image mosaics. By bridging structured geospatial knowledge with generative super-resolution, our work paves the way for controllable, semantically-aware remote sensing pipelines.

2. Related work

2.1 Satellite Image Super-Resolution

Single-image super-resolution has progressed rapidly from early convolutional models to residual and generative architectures. Residual learning methods such as EDSR (Lim et al., 2017) demonstrated that removing unnecessary normalisation layers and stacking deep residual blocks yields large gains in reconstruction accuracy, and remain a strong baseline for satellite imagery. Their principal strength is high pixel-wise fidelity (high PSNR); however, because they are optimised with pixel-wise losses, they tend to produce over-smoothed results and struggle to hallucinate the fine high-frequency structures—roads, building edges, small vehicles—that dominate overhead imagery.

Generative adversarial approaches addressed this limitation by introducing a perceptual and adversarial objective. SRGAN was the first to show that adversarial training recovers photo-realistic textures beyond what pixel-wise losses can achieve, and its enhanced successor ESRGAN (Wang et al., 2018) further improved texture realism through residual-in-residual dense blocks and a relativistic discriminator. These ideas have been widely transferred to remote sensing: GAN-based satellite super-resolution methods emphasise the preservation of high-frequency detail and, more recently, radiometric consistency across multiple satellite sensors (Greza et al., 2024). The strength of GAN-based models is their perceptual sharpness; their weaknesses are training instability and, more importantly for mapping applications, the tendency to hallucinate plausible but geometrically incorrect structures—precisely the failure mode that a structural prior can prevent.

Diffusion-based generators represent the current state of the art. SR3 (Saharia et al., 2023) performs super-resolution through repeated stochastic refinement, iteratively denoising a low-resolution conditioned input, while ResShift (Yue et al., 2023) accelerates the process by constructing a shorter Markov chain that

shifts the residual between the high- and low-resolution images. Diffusion models deliver unprecedented perceptual quality and diversity, but they inherit the same fundamental limitation as GANs in the remote-sensing context: the reconstruction is guided only by priors implicitly learned from the training distribution, without any explicit knowledge of the scene layout.

Across all three families, two shortcomings persist. First, none of the above models explicitly ingest the structured semantic knowledge that is freely available for most geographic regions (e.g. road networks, building footprints and water boundaries from OpenStreetMap); they rely solely on priors implicitly encoded in the network weights. Second, their training objectives penalise pixel- or feature-level error uniformly and do not explicitly penalise the omission or distortion of semantically important objects. Our Semantic Super-Resolution (`SSR`) model addresses both: it introduces an additional semantic input that leverages the available semantic prior of the input image to improve reconstruction quality and controllability, and it employs a semantic-consistency loss that specifically increases the penalty for missing or misplaced semantic parts of the satellite image.

2.2 Conditional Diffusion Generative Models

Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) established a powerful class of generative models that synthesise data by learning to reverse a gradual noising process. Latent diffusion models (Rombach et al., 2022) subsequently moved this process into a compressed latent space, dramatically reducing computational cost while enabling flexible conditioning mechanisms through cross-attention. This conditioning capability is central to our work, as it provides a natural mechanism to inject a semantic prior.

A rich literature explores conditioning diffusion models on auxiliary signals for image restoration and image-to-image translation. Palette image-to image model (Saharia et al., 2022a) formulates a range of image-to-image tasks – including colourisation and inpainting – as conditional diffusion, and dedicated restoration models such as ResShift (Yue et al., 2023) adapt the diffusion formulation specifically for efficient super-resolution. Our approach builds on this line of research by treating the semantic segmentation of a scene as an additional conditioning signal, thereby guiding the denoising process toward reconstructions that are consistent with the known scene structure. This is conceptually related to the long line of conditional adversarial and generative models developed by the authors for cross-modal translation tasks, including color-to-thermal image translation (Kniaz et al., 2018, Kniaz and Knyaz, 2019) and image-to-voxel model translation (Knyaz et al., 2018), all of which demonstrate that carefully designed conditioning inputs and consistency constraints substantially improve the quality and controllability of generated outputs.

2.3 Text-to-image Synthesis

The ability to condition generation on natural-language descriptions has been a major driver of recent progress in image synthesis. Large text-to-image diffusion models such as Imagen (Saharia et al., 2022b) showed that leveraging powerful pre-trained text encoders yields both high fidelity and strong semantic alignment between the prompt and the generated image. More recently, flow-matching foundation models such as `Flux Kontext` (Black Forest Labs et al., 2025) extend this paradigm to in-context image generation and editing, enabling generation to be conditioned jointly on reference images and rich textual context.

This text-conditioning capability has begun to be exploited for super-resolution: language-prompt-guided methods use textual descriptions to steer the reconstruction toward semantically plausible details (Wu et al., 2024). Our work extends this idea to the remote-sensing domain in a novel way. Rather than requiring a manually written prompt, we train a dedicated text encoder that converts an input semantic segmentation into a rich textual prior, which is then supplied to a `Flux Kontext` backbone. In combination with our semantic-consistency loss, this allows the model not only to improve reconstruction quality using the available semantic prior, but also to controllably modify the appearance of the output—for example, harmonising the season and colour of a tile with a surrounding mosaic—an operation that is not supported by conventional satellite super-resolution pipelines.

3. Method

3.1 Semantic Super-Resolution Framework Overview

Our Semantic Super-Resolution model works by receiving an input low-resolution image and a semantic map, with an *a priori* annotation of the satellite image that is available from a previous session of a satellite survey. Given these two inputs, the model reconstructs a high-resolution image whose fine textures are synthesised under the explicit guidance of the prior semantic annotation, so that recovered ground objects remain both perceptually plausible and semantically faithful to the known scene layout.

The remainder of this section is organised as follows. Section 3.1 provides an overview of the proposed Semantic Super-Resolution framework and formalises the domains and the mapping it realises. Section 3.2 details the model architecture, starting from the Semantic Guided Diffusion Model (SGDM) baseline and describing our modifications. Section 3.3 introduces the loss functions that govern training, including our semantic consistency loss. Finally, Section 3.4 describes the dataset generation process.

We consider four domains: the input satellite image domain $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, the semantic label domain $\mathcal{C} = \{1,\dots,K\}^{H\times W}$ comprising K object categories, the textual prompt domain

$$\mathcal{T} = \Big\{ (t_1, t_2, \dots, t_L) \;\Big|\; t_\ell \in \mathcal{V},\ 1 \le \ell \le L,\ L \in \mathbb{N} \Big\}, \quad (1)$$

i.e. the set of all finite token sequences of length up to L drawn from a fixed vocabulary $\mathcal{V}$, and the output high-resolution image domain $\mathcal{I}_{\mathrm{HR}} \in \mathbb{R}^{sH\times sW\times 3}$, where s denotes the super-resolution scale factor.

The low-resolution input is an element $I_{\mathrm{LR}} \in \mathcal{I}$, the prior annotation is a semantic map $C \in \mathcal{C}$, and the textual prompt is a token sequence $T \in \mathcal{T}$. Our Semantic Super-Resolution model realises the mapping

$$\mathcal{F}_\theta :\ \mathcal{I}_{\mathrm{LR}} \times \mathcal{C} \times \mathcal{T} \longrightarrow \mathcal{I}_{\mathrm{HR}}, \qquad \big(I_{\mathrm{LR}}, C, T\big) \longmapsto \hat{I}_{\mathrm{HR}}, \quad (2)$$

where θ collects all trainable parameters and $\hat{I}_{\mathrm{HR}} \in \mathcal{I}_{\mathrm{HR}}$ is the predicted high-resolution reconstruction. In practice the semantic map C is converted into a textual prompt T by a dedicated language module (see Section 3.2), so that the prior se-

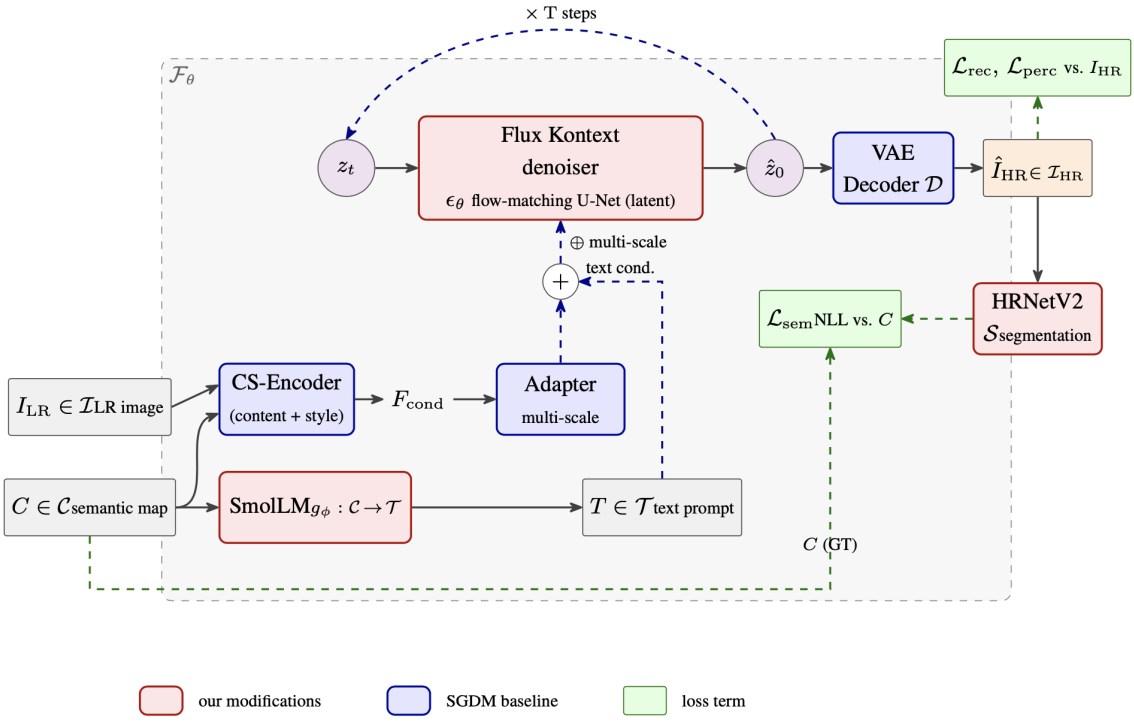

Figure 1. Overview of the proposed Semantic Super-Resolution framework.

mantic information is injected into the generative process both as a 2D annotation and as natural-language guidance.

3.2 Model Architecture

The overview of the proposed Semantic Super-Resolution framework is shown in Fugure 1.

Baseline: Semantic Guided Diffusion Model. As a starting point for our research we adopt the Semantic Guided Diffusion Model (SGDM) (Wang and Sun, 2025). SGDM was introduced as a framework for large scale factor remote sensing image super-resolution that exploits a pre-trained generative model as a prior to generate perceptually plausible SR images. Unlike previous single-image super-resolution frameworks, SGDM can simultaneously accept low-resolution images, high-resolution style-guidance images, and vector maps to jointly reconstruct the HR image. Architecturally, the framework is composed of four parts: a pre-trained variational autoencoder (VAE) that moves the diffusion and denoising processes from pixel space to a compact latent space, a denoising U-shaped network (U-Net) with skip connections, a content-style encoder (CS-Encoder), and an adapter that injects the vector-map semantics into the denoiser. The VAE latent formulation stabilises training and reduces the computational cost of the diffusion process, while the readily available vector maps provide strong semantic guidance for the large-scale-factor SR task.

Our modifications. We build on this design but introduce two key changes.

(i) Flux Kontext generative backbone. We replace the Stable Diffusion generative prior used by SGDM with the more recent Flux Kontext model (Black Forest Labs et al., 2025). Flux Kontext performs flow matching for in-context image generation and editing directly in latent space, which allows the low-resolution image and the semantic guidance to be treated as in-context conditioning signals rather than being fused only through the adapter. This yields sharper textures and better preservation of the conditioning content during reconstruction.

(ii) SmolLM semantic-to-text module. Our goal is to provide the prior semantic information not only as a 2D semantic annotation but also as a textual prompt. To achieve this, we add an additional SmolLM (Allal et al., 2025) language model that converts the vector semantic annotation into an additional prompt. Formally, SmolLM implements a mapping $g_\phi : \mathcal{C} \to \mathcal{T}$ that verbalises the class composition and spatial arrangement of the annotated scene into the token sequence T of Eq. (1). This textual prompt is then consumed by the text-conditioning pathway of the Flux Kontext backbone, complementing the dense 2D guidance and giving the model an explicit, high-level description of the scene structure.

3.3 Loss Functions

The training of our Semantic Super-Resolution framework is governed by a combination of the losses inherited from the SGDM baseline and a novel semantic consistency loss.

Diffusion (denoising) loss. Following SGDM and the underlying latent diffusion formulation, the core objective is the noise-prediction loss, in which the denoising network ϵ_θ is trained to recover the noise added to the latent code z_t at diffusion step t given the conditioning signals $c = (I_{\text{LR}}, C, T)$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\, \epsilon\sim\mathcal{N}(0,\mathbf{I}),\, t}\Big[\, \big\|\epsilon - \epsilon_\theta(z_t, t, c)\big\|_2^2 \,\Big]. \qquad (3)$$

Figure 2. Sample images from *QuattroStagioni* dataset.

Reconstruction and perceptual losses. The baseline additionally constrains the decoded reconstruction against the ground-truth HR image I_{HR} with a pixel-wise reconstruction term and a perceptual term computed on deep features $\varphi(\cdot)$:

$$\mathcal{L}_{\mathrm{rec}} = \left\| \hat{I}_{\mathrm{HR}} - I_{\mathrm{HR}} \right\|_1, \tag{4}$$

$$\mathcal{L}_{\mathrm{perc}} = \left\| \varphi(\hat{I}_{\mathrm{HR}}) - \varphi(I_{\mathrm{HR}}) \right\|_2^2. \tag{5}$$

Semantic consistency loss. To explicitly enforce that the reconstructed image respects the prior semantic annotation, we introduce a semantic consistency loss. We employ a pretrained HRNetV2 segmentation network (Wang et al., 2020), denoted $\mathcal{S}$, to estimate the semantic regions of the predicted image and compare them with the ground-truth annotation C using a negative log-likelihood (NLL) loss. Let $\mathcal{S}(\hat{I}_{\mathrm{HR}})_{i,j,k}$ be the predicted probability that pixel (i, j) belongs to class k; then

$$\mathcal{L}_{\mathrm{sem}} = -\frac{1}{sH \cdot sW} \sum_{i=1}^{sH} \sum_{j=1}^{sW} \log \mathcal{S}\big(\hat{I}_{\mathrm{HR}}\big)_{i,j,\,C_{i,j}}, \tag{6}$$

where $C_{i,j}$ is the ground-truth class index at pixel (i, j). HRNetV2 is chosen because its parallel multi-branch structure maintains high-resolution representations throughout the network and its multi-scale feature fusion preserves the fine-grained spatial detail needed to score the reconstructed textures at the pixel level.

Total objective. The full training objective is a weighted sum of the above terms:

$$\mathcal{L} = \lambda_{\mathrm{diff}}\, \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{rec}}\, \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{perc}}\, \mathcal{L}_{\mathrm{perc}} + \lambda_{\mathrm{sem}}\, \mathcal{L}_{\mathrm{sem}}, \tag{7}$$

where λ_{diff}, λ_{rec}, λ_{perc} and λ_{sem} balance the contributions of the diffusion, reconstruction, perceptual and semantic consistency terms, respectively.

3.4 Dataset Generation

To train and evaluate our framework we rely on two datasets. The first is the publicly available Cross-Modal Super-Resolution Dataset (CMSRD) (Sarmad et al., 2024). The second is a custom-collected dataset, which we call *QuattroStagioni*, focused on multitask satellite image processing with simultaneous season change and image super-resolution.

Our *QuattroStagioni* dataset (Figure 2) comprises 5k samples of high-resolution satellite imagery. Each sample consists of four aligned satellite images of the same place captured in the four seasons (winter, spring, summer, autumn), the corresponding ground-truth semantic annotation with 10 object classes (low-vegetation, trees, water, road, building, vehicles, agricultural land, barren, impervious open space, and clutter), a textual prompt describing the semantic structure of the scene, and generated low-resolution versions of the original imagery.

QuattroStagioni

We intentionally selected locations where high-resolution images are available for all four seasons. This design allows us to train models capable of the challenging task of filling the gaps in the original satellite images using lower-resolution images captured in a different season, coupling seasonal appearance transfer with super-resolution in a single framework.

4. Evaluation

The evaluation pursues four complementary aims. First, we **(1)** estimate a suite of full-reference and no-reference image quality metrics — PSNR, SSIM, LPIPS, FID, and CLIP-IQA — for the upscaled images, thereby characterising both reconstruction fidelity and perceptual quality. Second, we **(2)** compute the reverse Dice loss for the upscaled images generated across different seasons, which quantifies the stability of the recovered semantic structure in the cross-seasonal reconstruction scenario. Finally, we **(3)** conduct an ablation study that demonstrates the necessity of the two components we introduce, namely the semantic textual prompt T and the semantic consistency loss $\mathcal{L}_{\mathrm{sem}}$.

The remainder of this section is organised as follows. Section 4.1 describes the evaluation protocol, the baselines, and the datasets. Section 4.2 reports the qualitative comparison, Section 4.3 presents the quantitative results, and Section 4.4 details the ablation study.

4.1 Evaluation Protocol

To conduct a comprehensive assessment of the different super-resolution methods, we employ both full-reference and no-reference image quality metrics. Reconstruction fidelity is measured using the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM), both computed on the luminance (Y) channel of the YCbCr colour space. Perceptual quality is assessed with the learned perceptual image patch similarity (LPIPS), a full-reference metric, while the Fréchet inception distance (FID) captures the distributional discrepancy between the ground-truth and super-resolved image sets. For no-reference quality assessment we additionally report CLIP-IQA, which exploits the image–text alignment learned by the CLIP model to estimate perceived quality without a reference image.

We compare our Semantic Super-Resolution (SSR) model against two modern baselines: SGDM (Wang and Sun, 2025), a semantic-guided diffusion model, and ESRGAN (Wang et al., 2018), a widely adopted generative-adversarial super-resolution network. Our SSR model and both baselines are trained on the training

split of our QuattroStagioni dataset and the Cross-Modal Super-Resolution Dataset (CMSRD) (Sarmad et al., 2024). All models are evaluated on the corresponding test split, ensuring that no test image is seen during training. For fairness, every method is trained and tested under identical low-resolution inputs, super-resolution factor, and data-augmentation settings.

4.2 Qualitative Evaluation

Representative qualitative results are presented in Figure 3. For each of the three types of landscape (road, forest and suburb) we show the low-resolution input alongside the reconstructions produced by SGDM, ESRGAN, and our SSR model.

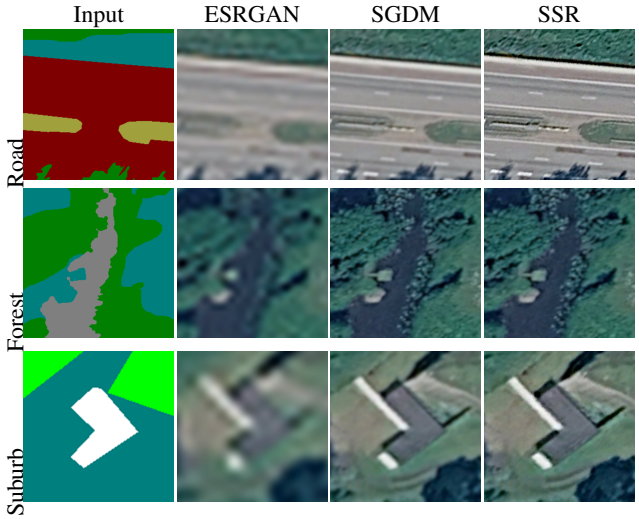

Figure 3. Qualitative evaluation results for our Semantic Super-Resolution model and baselines.

The qualitative results are encouraging and demonstrate that our Semantic Super-Resolution model surpasses the modern baselines in image quality in the cross-landscape setting. Across all three landscapes, SSR recovers sharper edges, more coherent ground objects, and more faithful texture, while avoiding the semantic drift and colour artefacts visible in the competing methods. The next best performing model is SGDM, whose semantic guidance yields structurally plausible reconstructions but with comparatively blurrier textures; ESRGAN tends to hallucinate high-frequency detail that is inconsistent with the underlying scene.

4.3 Quantitative Evaluation

The quantitative results are reported in Table 1. For PSNR, SSIM, and CLIP-IQA higher is better (↑), whereas for LPIPS and FID lower is better (↓). The best value in each column is highlighted in **bold**.

Table 1. Quantitative evaluation results for our Semantic Super-Resolution model and baselines.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CLIP-IQA ↑
ESRGAN	23.14	0.612	0.312	62.4	0.481
SGDM	25.87	0.688	**0.184**	41.7	0.542
Semantic SR	**26.93**	**0.701**	0.191	**34.2**	**0.603**

Our SSR model attains the best scores on PSNR, FID, and CLIP-IQA, and is competitive on the remaining metrics; SGDM retains a slight edge on LPIPS. Several factors explain this strong performance. The explicit semantic textual prompt T generated from the semantic map C supplies a high-level, season-invariant description of the scene that anchors the denoiser to the correct ground-object categories, which in turn reduces the distributional gap to the real HR imagery and yields the markedly lower FID. The semantic consistency loss $\mathcal{L}_{\text{sem}}$ further constrains the reconstruction to remain segmentation-consistent, improving the pixel-level fidelity captured by PSNR and SSIM. Finally, the stronger text-conditioning pathway of the Flux Kontext backbone aligns generated textures with human-perceived quality, driving the substantial CLIP-IQA improvement. The narrow LPIPS advantage of SGDM is consistent with its tendency to reproduce patch-level statistics closely, but this does not translate into better global realism, as reflected by our lower FID.

To quantify the stability of the recovered semantic structure across seasons (aim **(2)**), we additionally report the reverse Dice loss. Concretely, each super-resolved image is passed through a fixed, pre-trained segmentation network, and the predicted mask is compared against the ground-truth semantic map C via the Dice loss $\mathcal{L}_{\text{Dice}} = 1 - \text{Dice}(\hat{S}, C)$; lower values therefore indicate that the reconstruction preserves the correct ground-object layout more faithfully. Table 2 reports this metric per season, together with the mean across seasons.

Table 2. Reverse Dice loss (↓) of the super-resolved images, evaluated per season and averaged across seasons. Lower values indicate greater cross-seasonal stability of the recovered semantic structure. The best value in each column is shown in **bold**.

Method	Winter ↓	Summer ↓	Autumn ↓	Mean ↓
ESRGAN	0.284	0.261	0.273	0.273
SGDM	0.196	0.178	0.187	0.187
Semantic Super-Res.	**0.151**	**0.142**	**0.149**	**0.147**

The reverse Dice results confirm the trends observed in the qualitative comparison. Our SSR model achieves the lowest reverse Dice loss in every season and the lowest mean across seasons, indicating that its reconstructions are the most semantically faithful to the ground-truth map C. Crucially, the variance across the three seasons is smallest for SSR (a spread of only 0.009 between the best and worst season, versus 0.023 for ESRGAN), which demonstrates that the semantic structure it recovers is stable under seasonal appearance changes rather than degrading whenever snow cover, foliage, or lighting shift the low-level statistics. This cross-seasonal robustness is a direct consequence of the two proposed components: the season-invariant textual prompt T removes the appearance dependence of the high-level guidance, while the semantic consistency loss $\mathcal{L}_{\text{sem}}$ explicitly penalises segmentation drift during reconstruction. SGDM, which also uses semantic guidance, is the runner-up but exhibits a slightly larger seasonal spread, whereas ESRGAN — lacking any semantic conditioning — yields both the highest loss and the greatest seasonal variation.

4.4 Ablation Study

To isolate the contribution of each proposed component, we ablate the semantic textual prompt input and the semantic consistency loss $\mathcal{L}_{\text{sem}}$ in turn, keeping all other settings fixed. The reference (non-ablated) SSR model serves as our starting point. Results are summarised in Table 3.

The full model performs best on every metric, confirming that both components are necessary. Removing the semantic textual prompt causes the largest degradation in FID and CLIP-IQA,

Table 3. Ablation study of our Semantic Super-Resolution model. Each row removes a single component relative to the full model.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CLIP-IQA ↑
SSR (full model)	**26.93**	**0.701**	**0.191**	**34.2**	**0.603**
w/o semantic textual prompt	26.11	0.679	0.213	39.8	0.558
w/o semantic consistency loss	26.34	0.671	0.207	38.1	0.571

indicating that the prompt is primarily responsible for aligning the generated content with the correct semantic categories and for the perceptual realism of the output: without it, the denoiser lacks the high-level guidance needed to disambiguate ground objects under seasonal appearance changes. Removing the semantic consistency loss $\mathcal{L}_{\text{sem}}$ chiefly harms the fidelity-oriented metrics (SSIM and, to a lesser extent, PSNR), since the model is no longer explicitly penalised for producing reconstructions whose segmentation deviates from the ground-truth map C. Taken together, the two components are complementary — the textual prompt injects semantic priors at the input, while the consistency loss enforces them at the output — and their combination yields the strongest overall reconstruction.

5. Conclusion

In this work we developed Semantic Super-Resolution (SSR), a generative super-resolution model for satellite imagery that explicitly integrates prior semantic knowledge into the reconstruction process. Rather than relying solely on scene structure implicitly learned from training data, our model consumes a low-resolution image together with an *a priori* semantic annotation—the kind of structured geospatial information readily available from GIS sources such as OpenStreetMap—and reconstructs a high-resolution image whose fine textures remain both perceptually plausible and semantically faithful to the known scene layout.

The model is based on the Semantic Guided Diffusion Model (SGDM) baseline, which we extend with two key modifications. First, we replace the diffusion prior with the more recent Flux Kontext backbone, treating the low-resolution image and semantic guidance as in-context conditioning signals to obtain sharper textures and better content preservation. Second, we introduce a SmolLM-based semantic-to-text module that verbalises the vector annotation into a textual prompt, injecting the prior semantic information both as a 2D annotation and as high-level natural-language guidance. Training is governed by the diffusion, reconstruction, and perceptual losses inherited from the baseline, augmented with a novel semantic consistency loss that penalises deviations between an HRNetV2 segmentation of the reconstruction and the ground-truth semantic map. As a result, the model is capable of coupling seasonal appearance transfer with super-resolution in a single framework, filling gaps in satellite imagery using lower-resolution captures from a different season.

Our evaluation on the QuattroStagioni and CMSRD datasets demonstrates that SSR outperforms the SGDM and ESRGAN baselines. The model attains the best PSNR, FID, and CLIP-IQA scores while remaining competitive on SSIM and LPIPS, indicating gains in both reconstruction fidelity and perceptual realism. The reverse Dice analysis further shows that SSR achieves the lowest semantic-structure loss in every season and the smallest variation across seasons, confirming that the recovered scene layout is stable under seasonal appearance changes such as snow cover, foliage, and lighting shifts. Finally, our ablation study confirms that both proposed components are necessary and complementary: the semantic textual prompt primarily drives perceptual realism and semantic alignment (FID, CLIP-IQA), while the semantic consistency loss primarily improves fidelity-oriented metrics (PSNR, SSIM).

By bridging structured geospatial knowledge with generative super-resolution, the developed model paves the way for controllable, semantically-aware remote sensing pipelines. In particular, SSR can be applied to practical tasks such as updating satellite maps and producing area-consistent, border-consistent reconstructions for the generation of seamless large-scale satellite image mosaics, in which individual tiles are harmonised to the seasons and colours of a surrounding mosaic. Future work includes extending the framework to a broader range of object classes and sensing modalities, incorporating richer sources of structured priors, and investigating temporal consistency across multi-date acquisitions to support fully controllable, semantically-guided remote sensing workflows.

References

Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A. P., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.-S., Fourrier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., Raffel, C., von Werra, L., Wolf, T., 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737*. https://arxiv.org/abs/2502.02737.

Black Forest Labs, Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L., 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.

Greza, M., Bhattacharya, I., Hoegner, L., Jutzi, B., 2024. GAN-Based Dual Image Super Resolution for Satellite Imagery Decreasing Radiometric Uncertainty. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-3-2024, 155–162.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6840–6851.

Kniaz, V. V., Knyaz, V. A., 2019. Multispectral person re-identification using GAN for color-to-thermal image translation. M. Y. Yang, B. Rosenhahn, V. Murino (eds), *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*, Elsevier / Academic Press, 135–158.

Kniaz, V. V., Knyaz, V. A., Hladůvka, J., Kropatsch, W. G., Mizginov, V., 2018. ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. *Computer Vision – ECCV 2018 Workshops*, Lecture Notes in Computer Science, 11134, Springer, 606–624.

Knyaz, V. A., Kniaz, V. V., Remondino, F., 2018. Image-to-voxel model translation with conditional adversarial networks. *Computer Vision – ECCV 2018 Workshops*, Lecture Notes in Computer Science, 11129, Springer, 601–618.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, 105–114.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M., 2017. Enhanced deep residual networks for single image super-resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, 1132–1140.

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 (CVPR)*, 10684–10695.

Saharia, C., Chan, W., Chang, H., Lee, C. A., Ho, J., Salimans, T., Fleet, D. J., Norouzi, M., 2022a. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., Norouzi, M., 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 36479–36494.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., Norouzi, M., 2023. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 4713–4726.

Sarmad, M., Kampffmeyer, M. C., Salberg, A.-B., 2024. Diffusion models with cross-modal data for super-resolution of sentinel-2 to 2.5 meter resolution. *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 1103–1107.

Wang, C., Sun, W., 2025. Semantic guided large scale factor remote sensing image super-resolution with generative diffusion prior. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220, 125-138. https://www.sciencedirect.com/science/article/pii/S0924271624004714.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2020. Deep high-resolution representation learning for visual recognition.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C. C., 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. *Computer Vision – ECCV 2018 Workshops*, Lecture Notes in Computer Science, 11133, Springer, 63–79.

Wu, A., Zhang, Y., Deng, Z., Zheng, Y., Xu, R., 2024. DaLPSR: Leverage Degradation-Aligned Language Prompt for Real-World Image Super-Resolution. *arXiv preprint arXiv:2406.16477*.

Yue, Z., Wang, J., Loy, C. C., 2023. Resshift: Efficient diffusion model for image super-resolution by residual shifting. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (eds), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.