Trustworthy Super-Resolution of Multispectral Sentinel-2 Imagery With Latent Diffusion

Simon Donike, Cesar Aybar, Luis Gómez-Chova, *Senior Member, IEEE*, and Freddie Kalaitzis

***Abstract*—Remote sensing super-resolution aims to enhance the spatial details of satellite images by introducing meaningful high-frequency features while avoiding hallucinations and spectral distortions. High-resolution imagery is usually not publicly available, whereas low-resolution imagery is freely available with a much higher revisit rate, such as the Sentinel-2 multispectral imaging mission. Cross-sensor super-resolution has the potential to bridge this gap, providing high spatial and temporal resolution imagery which are otherwise unavailable for many remote sensing users and applications. With the recent advancements in diffusion models, many methodologies have emerged which take advantage of their generative power to perform super-resolution. We propose an adapted latent diffusion approach, since image diffusion is computationally prohibitive to be applied to large Earth observation datasets. Contrary to standard latent diffusion, we encode the low-resolution image to condition the diffusion process, forcing better spectral consistency with the input imagery. The model includes visible and near-infrared bands. To ensure trustworthy results, we utilize the probabilistic nature of diffusion models to generate pixel-level uncertainty maps. This confidence metric is crucial for real-world applications, such as environmental monitoring, land cover classification, and change detection, where accurate surface feature reconstruction and spectral consistency are essential. The uncertainty map allows users to evaluate the reliability of the product for these tasks. The proposed model super-resolves Sentinel-2 imagery at 10 to 2.5 m and is the first multispectral remote sensing (RS) super-resolution diffusion model efficient enough to process large-scale RS datasets, as well as the only model providing a pixelwise uncertainty metric.**

***Index Terms*—Deep learning, latent diffusion, model uncertainty, remote sensing (RS), Sentinel-2, super-resolution (SR).**

I. Introduction

EARTH observation (EO) represents the collection and analysis of information about the Earth's physical, chemical, and biological systems, as well as the anthropological influences on these systems. Remote sensing data acquired from remote sensing (RS) satellites and aerial platforms play a crucial role in monitoring environmental changes, supporting disaster response and climate research, and surveilling and predicting agricultural yield [1]. EO data and the insights gained from their analysis provide policy makers and the private sector with the knowledge needed to make well-founded decisions [2]. The European Space Agency (ESA) Sentinel-2 (S2) satellites provide freely available optical imagery at 10 m for RGB-NIR bands with a high revisit rate of up to 5 days [3]. However, privately owned satellites, such as SPOT or Pléiades Neo acquire very high-resolution (HR) data (1.5-m pansharpened and 0.3 m, respectively), but are expensive to acquire and have a much lower revisit rate. Both the spatial and temporal resolution of EO images are fundamental bottlenecks in downstream applications [4], [5], [6], [7] with spatial resolution limiting the scale of analysis [8]. A spatial super-resolution (SR) workflow that approximates the spectral and spatial HR image characteristics based on a low-resolution (LR) image can alleviate these issues, providing researchers and organizations with free HR data with high temporal resolution [9]. The problem can be formalized as $x = \Phi(y;\theta)$, where the LR image x is defined as the degradation of the HR image y by the function Φ with the degradation parameters θ. The goal is to reverse the effects of Φ as accurately as possible to reconstruct $\hat{y} = \Phi^{-1}(x;\theta)$. Due to the inherent lack of information, this inverse problem is ill-posed, which means that many plausible $\hat{y}$ reconstructions exist [10]. However, some reconstructions are more plausible than others, which is why it is crucial to draw the most likely reconstruction from an appropriate probability distribution. The term SR is also applied to spectral SR, where spectral instead of spatial information is added [11], [12], [13], [14], or more broadly to other resolution enhancement methodologies, such as deep learning-based pan-sharpening [15].

We argue that an impactful RS SR model should:

1) provide state-of-the-art performance in order to enhance the original imagery's capabilities;
2) extend beyond the RGB bands;
3) be applicable to freely available imagery in many geographical contexts;
4) come with metrics that allow the user to estimate the accuracy of the product.

We present an SR model named Latent Diffusion Super Resolution for Sentinel-2 (LDSR-S2) that is trained to super-resolve Sentinel-2 RGB-NIR imagery by a factor of $4\times$ from 10 to 2.5 m. Our model is based on the latent diffusion approach introduced by [21], adapted to RS SR by adding a fourth band, using a Wasserstein autoencoder (AE) [22], and additionally encoding the diffusion conditioning. The generative nature of this model

Received 22 October 2024; revised 24 December 2024 and 27 January 2025; accepted 4 February 2025. Date of publication 14 February 2025; date of current version 13 March 2025. This work was supported by the European Space Agency (ESA)-Explainable AI: Application to Trustworthy Super-Resolution (OpenSR). The work of L. Gómez-Chova was supported by the Spanish Ministry of Science and Innovation (project PID2023-148485OB-C21) under Grant MCIN/AEI/10.13039/501100011033. *(Corresponding author: Simon Donike.)*

Simon Donike, Cesar Aybar, and Luis Gómez-Chova are with the Image Processing Laboratory (IPL), University of Valencia, 46010 Valencia, Spain (e-mail: simon.donike@uv.es).

Freddie Kalaitzis is with the University of Oxford, OX12JD Oxford, U.K.

Code is available at https://github.com/ESAOpenSR/opensr-model.

Digital Object Identifier 10.1109/JSTARS.2025.3542220

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

TABLE I
COMPARISON OF RS-SR MODEL CAPABILITIES

Feature	RSRCAN [16]	SR4RS [17]	Satlas SR [18]	Evoland [19]	EDiffSR [20]	LDSR-S2 (ours)
Factor ≥ 4	✓	✓	✓	✗	✓	✓
Multispectral	✗	✓	✗	✓	✗	✓
S2 Inference	✗	✓	✓	✓	✗	✓
Confidence Metric	✗	✗	✗	✗	✗	✓
Code, Weights, Data public	✗	✗	✓	✓	✗	✓
Data	Aerial	S2	S2	S2	Aerial	S2
Method	Attn.-based	GAN	GAN	CNN	Diffusion	Lat. Diffusion

provides the capability to introduce realistic high-frequency textures and objects. Since the model is probabilistic, we can estimate the confidence of the SR product pixel- and regionwise through an analysis of the variance of multiple samples of the same location [23]. This enables the generation of pixelwise uncertainty metrics, which play a pivotal role in downstream tasks by offering insights into the confidence of SR outputs. These explainability metrics allow users to assess whether the reconstructed data can be reliably used for decision-making in their specific downstream applications, thus building trust among end users in the validity of the product. This trust in SR products is crucial to further the adoption and usage of SR products by the wider EO community.

The rest of this article is organized as follows. Section II reviews the SR models adopted in RS. Section III details the selected datasets and the training strategy. Section IV describes the architecture of the latent diffusion model (LDM) and the experimental setup. Sections V and VI present quantitative evaluation and qualitative insights on diffusion models. The discussion is provided in Section VII. Finally, Section VIII concludes this article.

II. RELATED WORK

A. Adoption of SR Models for Remote Sensing

The fields of medical imaging, astronomy, video and photo editing, as well as RS, all draw in part from the same body of knowledge, in particular in recent years, from the deep learning models and architectures proposed by the broader computer vision community [24]. These models are typically designed for RGB images taken by end-user cameras depicting everyday objects, trained on very large natural image datasets with artificial degradations to create synthetic LR–HR pairs. The adoption of these models proves especially challenging for RS applications due to the very different nature of the signals involved [6], [7]. RS data are based upon the physical reflectance, captured in a variety of spectral ranges, depicting information with very different spectral and spatial properties. The presence of more spectral bands than the standard three RGB bands additionally provides a challenge when trying to adapt these models to multispectral RS scenarios [7], [25]. Many proposed solutions perform 3-band SR even for RS imagery, but this fails to create a valuable product for downstream EO applications [16], [26], [27]. Some of the proposed models also do not provide a final operational product, but instead are trained on specific datasets, such as DOTA or UC Merced [28], [29], competing for small improvements in pixel-based fidelity metrics. Other methodologies implement SR techniques within application-specific workflows [30]. In this context, diffusion models have also been successfully used for RS 3-band SR [20], [31], [32], [33], [34].

Table I summarizes the characteristics and capabilities of the RS-SR models used in the experimental results to compare the proposed model against the state-of-the-art SR methods (see Section V for details). In the context of our outlined goals, most existing models meet some, but not all, of the key requirements. Although some models are applicable to Sentinel-2 imagery, they often exclude the NIR band. Furthermore, many do not provide publicly available code or checkpoints, while none provide metrics for accuracy estimation. Other diffusion models for SR exist, but these are usually not applicable to S2, use inefficient pixel-space diffusion, do not use multispectral data, and do not provide confidence metrics.

B. SR Diffusion Models

Diffusion models are a class of probabilistic and generative models that iteratively refine noise or conditioning vectors to recreate the original data distribution, originally inspired by nonequilibrium thermodynamics [35], [36], [37]. This type of models can be conditioned on many modalities and therefore can perform text-to-image generation, unconditional or conditional image generation, in-painting and SR [38], [39]. Diffusion models are characterized by their ability to gradually transform noise into structured data, mimicking the reverse of the diffusion process employed during training. When conditioning on an LR image, diffusion models can generate state-of-the-art SR performance [40], but their iterative process slows down the inference. Performing the diffusion process in a compressed latent space significantly speeds up the inference time but adds complexity to the training process.

III. DATASETS AND TRAINING STRATEGY

Our workflow should ideally rely on a comprehensive multispectral RGB-NIR 4-band dataset, consisting of 10-m LR and corresponding 2.5-m HR image pairs, spanning diverse geographic regions, without spectral of geographic distortions. Since the introduction of both a fourth band and the encoded conditioning mechanism requires retraining from scratch, a large amount of high-quality data are essential.

The *WorldStrat* dataset [41] was initially considered, offering approximately 12 000 image pairs (in 512×512 pixel patches) from SPOT6/7 HR and S2 LR images. However, the cross-sensor

nature of this dataset introduces georeferencing errors, spectral differences, and temporal discrepancies, complicating direct model training [42], [43]. This challenge prompted the creation of a tailored dataset and training strategy that integrates data from multiple sources to meet the high data demands of our AE and diffusion models.

Our training strategy begins with a large-scale natural image dataset from OpenImages [44], comprising 500-k images of at least 512 × 512 pixels. We generate LR versions through bicubic downsampling and Gaussian blurring, appending the fourth band as intensity. This dataset enables the AE and denoising UNet to learn general features for SR, even though the spatial and spectral characteristics differ from those of the target S2 data. To refine the model further, we employ a dataset of 250k LR–HR pairs derived from random S2 locations worldwide. The native 10-m S2 imagery serves as the HR version, with LR versions created by degrading 40-m interpolations [25]. This phase helps the model adapt to S2 spectra and learn RS specific features, balancing spatial and spectral properties. Finally, the SEN2NAIP dataset [45], [46] is used for the last stage of training. This dataset, based on NAIP aerial imagery, includes more than 300 k 512 × 512 pixel pairs, with various degradation models applied to simulate S2-like imagery. The SEN2NAIP dataset effectively addresses common challenges in RS SR by ensuring spectral consistency with S2 imagery, accurately replicating its spatial characteristics, and avoiding geographic and temporal discrepancies typically seen in cross-sensor datasets [42], [43], [47].

IV. Methodology

The LDM utilized in this work builds on the foundational work of [21]. Although the original model is versatile, capable of dealing with tasks, such as text-to-image generation, unconditional image generation, in-painting, and SR, our approach significantly modifies the architecture and training procedure to specifically address the challenges of super-resolving the four 10-m bands of S2 imagery in the context of RS.

Our methodology involves a three-phase process: first, training the AE until it achieves satisfactory performance and produces a latent space significant for SR. Then, the AE is frozen and the denoiser is trained using the latent space provided by the AE. Finally, the super-resolved image is obtained in the diffusion inference step.

A. LDSR-S2 Autoencoder

As opposed to pixel-space diffusion, latent diffusion performs the denoising in a previously created latent space. The creation of a meaningful latent space is crucial to the overall quality of the SR reconstruction. This model follows the general outline of denoising in the latent space [21] by compressing the original image space of 4 × 512 × 512 down to a latent representation of size 4 × 128 × 128. Since this dimensionality is the same as the LR, this compression factor allows a concatenation of the encodings with the DDPM noise vectors.

The AE architecture is similar to that of many variational AEs [48], [49], but, as opposed to the original LDM implementation, we select to regularize the latent space through the

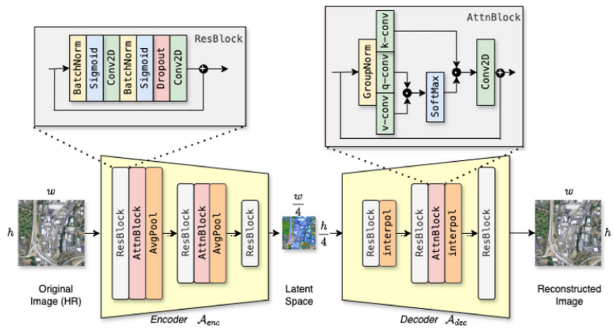

Fig. 1. Autoencoder training strategy.

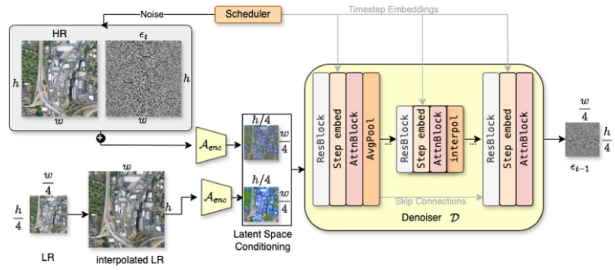

Fig. 2. Denoising UNet training strategy, conditioned on the encoded LR representation as well as the encoded HR image with noise ϵ_t.

Wasserstein Distance (WD) [50], [51]. This regularizes the probability distribution between the latent space created by the AE and the prior distribution, which is the original HR image (Fig. 1). Wasserstein AEs have been used in generative AI models because of their ability to maintain a well-structured latent space that reflects the underlying data distribution more effectively, leading to more stable training and improved generalization [22], [52].

Reconstruction quality after encoding and decoding is measured via a weighted loss of multiple components. We modify the standard LDM discriminator model to accept 4 bands, which is trained in parallel to the AE after a warm-up period. The GAN part of the discriminator leads to realistic reconstructions. The Learned Perceptual Image Similarity Index (LPIPS) [53] with a VGG backbone [54] recreates the perception of the human visual system, but the pretrained versions only accept 3-band input. We randomly select three of the 4 bands at each training step to calculate this metric.

The total loss $\mathcal{L}_{\text{total}}$ (1) is created by weighting the different components, forcing both a meaningful latent space and a high quality of reconstruction

$$\mathcal{L}_{\text{total}} = \lambda_{\text{WD}} \cdot \mathcal{L}_{\text{WD}}(z) + \lambda_{\text{MAE}} \cdot \mathcal{L}_{\text{MAE}}(\hat{x}_{\text{HR}}) + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}}(\hat{x}_{\text{HR}}) + \lambda_{\text{LPIPS}} \cdot \mathcal{L}_{\text{LPIPS}}. \quad (1)$$

B. LDSR-S2 Denoiser

The training process for the denoiser (Fig. 2), conditioned on the encoded representation of a low-resolution image, involves the following steps. First, the low-resolution image x_{LR} is interpolated to the target resolution x'_{LR} and then encoded into the latent representation z via the encoder. The x_{HR} is encoded

as well

$$z = \mathcal{A}_{\text{enc}}(x'_{\text{LR}}) \tag{2}$$

$$x^{\text{enc}}_{\text{HR}} = \mathcal{A}_{\text{enc}}(x_{\text{HR}}) \tag{3}$$

where $\mathcal{A}_{\text{enc}}$ denotes the encoder.

The forward diffusion process gradually adds noise to the encoded HR target representation $x^{\text{enc}}_{\text{HR}}$ over a sequence of timesteps $t = 1 \dots T$, creating a series of increasingly noisier images $x_1, \dots, x_T$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \tag{4}$$

where β_t is the variance schedule that controls the noise level at each timestep t. The training objective is to learn the reverse process, a denoising model $p_\theta(x_{t-1}|x_t, z)$, that predicts the noise at x_{t-1} from the noisier image x_t and the encoded LR representation z, which conditions the denoising process, parameterized by θ:

$$p_\theta(x_{t-1}|x_t, z) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, z), \Sigma_\theta(x_t, t, z)). \tag{5}$$

μ_θ and Σ_θ are functions learned by the model to predict the mean and variance of the noise distribution over x_{t-1}. The model is trained by minimizing an L2 loss, which is shown to be functionally equal to the negative log-likelihood by [55]. Saharia et al. [56] suggested that an L2 loss leads to more sampling diversity than L1, which is preferred for this application. The loss is calculated over the refined encoding x_{t-1} and the target $\hat{x}_{t-1}$ at the same noise level at timestep $t-1$, *refined encoding* and *encoded HR*).

The interpolation and encoding of x_{LR} prevents the model from having to learn the mapping from the RGB-NIR space to the encoded space. In a traditional 3-band SR LDM, the RGB-space x_{LR} is used as conditioning. Since the decoder $\mathcal{D}$ expects the denoiser output in the latent space domain of z, the translation must occur during denoising. Feeding the denoising with the previously encoded conditioning z removes this task from the denoiser, leading to closer spectral coherence of $\hat{x}_{\text{HR}}$ to x_{LR} and therefore x_{HR}, which is desirable in RS applications [57].

C. LDSR-S2 Inference Sampling

Given the low-resolution image x_{LR}, we interpolate it to the target resolution x'_{LR} and then encode it using the AE $\mathcal{A}_{\text{enc}}$ to get its latent representation

$$z = \mathcal{A}_{\text{enc}}(x'_{\text{LR}}). \tag{6}$$

Contrary to the standard DDIM sampling process [37], we condition the generation on the latent representation z of LR instead of its original RGB-NIR state (Fig. 3). The process iteratively refines the prediction of the HR image starting from a noise distribution

$$x_T \sim \mathcal{N}(0, I) \tag{7}$$

$$\hat{x}_{t-1} = \mathcal{D}(x_t, t, z; \theta) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_t) \tag{8}$$

where T is the total number of inference timesteps, $\mathcal{D}$ is the denoising function parameterized by θ that conditions on z, and ϵ_t is the noise added at timestep t with covariance Σ_t. The latent representation of the high-resolution image $\hat{x_0}$ obtained after the iterative refinement process, is then decoded to produce the final HR image

$$\hat{x}_{\text{HR}} = \mathcal{A}_{\text{dec}}(\hat{x}_0) \tag{9}$$

where $\mathcal{A}_{\text{dec}}$ represents the decoder part of the AE, transforming the final latent representation $\hat{x}_0$ into the HR prediction $\hat{x}_{\text{HR}}$. Fig. 3 shows a graphical representation of this workflow.

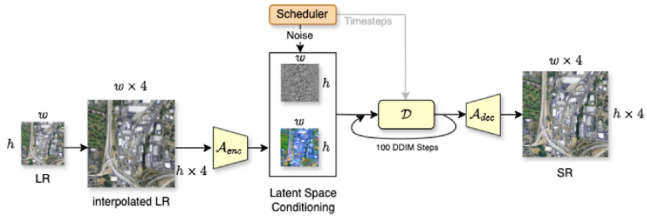

Fig. 3. Inference step, where the interpolated and encoded LR image is concatenated with the noise ϵ_t in order to provide the conditioning for the DDIM sampling process. The refined encoding containing the introduced details is then decoded and spatially increased to produce the SR image.

TABLE II
AE PERFORMANCE METRICS COMPARISON FOR SEN2NAIP HR AND S2 INTERPOLATED TO HR SPACE, AVERAGED OVER 512 × 512 × 4 IMAGE PATCHES OF VALIDATION DATASETS

Data Type	PSNR ↑	SSIM ↑	LPIPS ↑
SEN2NAIP	37.08	0.918	0.88
S2 - LR int.	45.88	0.994	0.99

V. RESULTS AND ANALYSIS

Although the main training is performed on the SEN2NAIP dataset [45], [46], the intention is to generalize these checkpoints to real S2 imagery. Since there is no actual HR ground truth imagery for the S2 images, the results on the synthetic and the real-world cross-sensor datasets are both presented. The quality metrics are calculated in the test set of SEN2NAIP and some selected S2 tiles to provide a global baseline.

A. Autoencoder Results

The AE shows robust performance in capturing and compressing the synthetic input data. Through extensive experimentation, the AE is able to consistently achieve high fidelity of reconstruction across a variety of spectral and geographical features. This shows that the spatial reduction by a factor of 4 can be reversed with only a minor loss in quality, validating the band permutation approach described in Section IV-A. The good reconstruction quality ensures that the AE's decoder can recover the detail added by the denoiser after the DDIM sampling step, converting the image back into the RGB-NIR space and simultaneously upsampling it to the desired resolution (see Fig. 4).

Since we use the AE to encode the S2 conditioning input image at inference time, the performance on interpolated S2 data must be validated. Because the interpolated S2 image contains much less information than the HR ground truth, the model has no issue reconstructing the input image. This is evident from the reconstruction metrics in Table II and Fig. 4. Both the visual reconstruction quality and the spectral reconstruction are satisfactory, showing that the AE is capable of performing

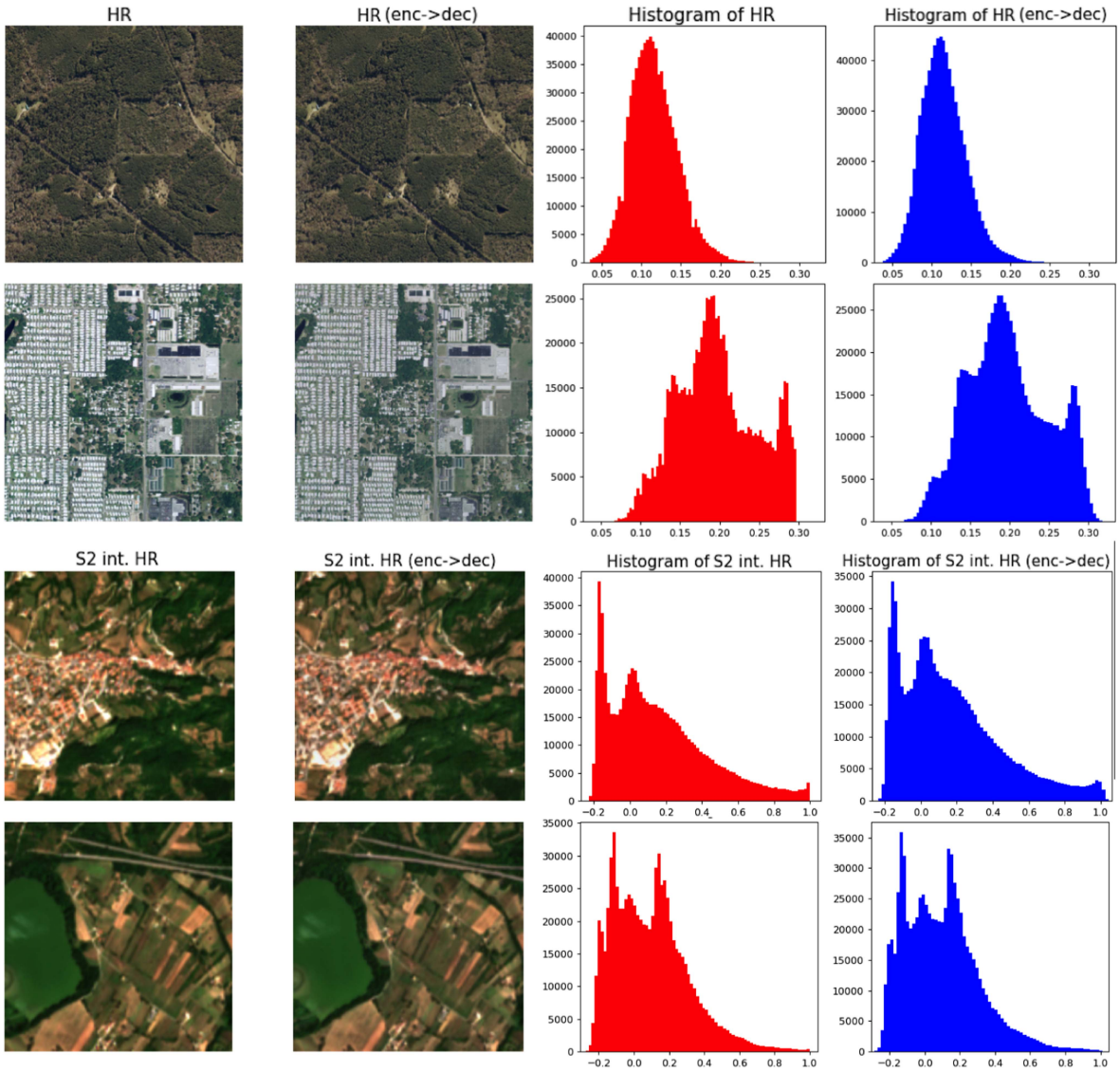

Fig. 4. RGB bands of the HR images, and it's reconstruction after encoding and decoding; also including histograms to show spectral faithfulness. Shown for both the SEN2NAIP synthetic test dataset (upper panel), and interpolated Sentinel-2 images (bottom panel).

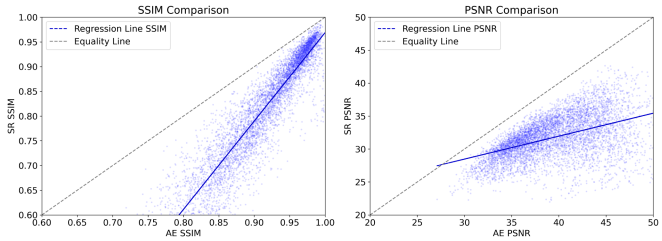

Fig. 5. AE reconstruction quality on HR compared with the performance of the whole SR process on every SEN2NAIP test image.

both tasks (encoding and decoding) on both input data types (synthetic HR in training, S2 interpolated LR in inference).

The AE performs the upsampling and detail retrieval process, which can be considered a bottleneck for overall reconstruction quality. Therefore, it is of interest to analyze how the quality of AE reconstruction impacts SR performance. Fig. 5 shows that the SR metrics never outperform the quality of the AE reconstruction, confirming the importance of a properly trained AE. In addition, the regression lines imply that there is a clear correlation between the two reconstruction qualities, showing that for a given image, the SR metrics can only ever be as good as the AE reconstruction in ideal circumstances.

B. SR Results

For validation purposes, we provide both visual and quantitative metrics. The reconstruction capability of the model based on the synthetic test dataset of SEN2NAIP is demonstrated, as well as the generalization to real-world S2 data. We also compare an LDM model and a pixel-space DM with our results to show the improvements over the baseline performance.

1) Synthetic Dataset Results: Before comparing the results of different SR models, we show the results of the proposed LDSR-S2 on synthetic data. Fig. 6 shows the model output

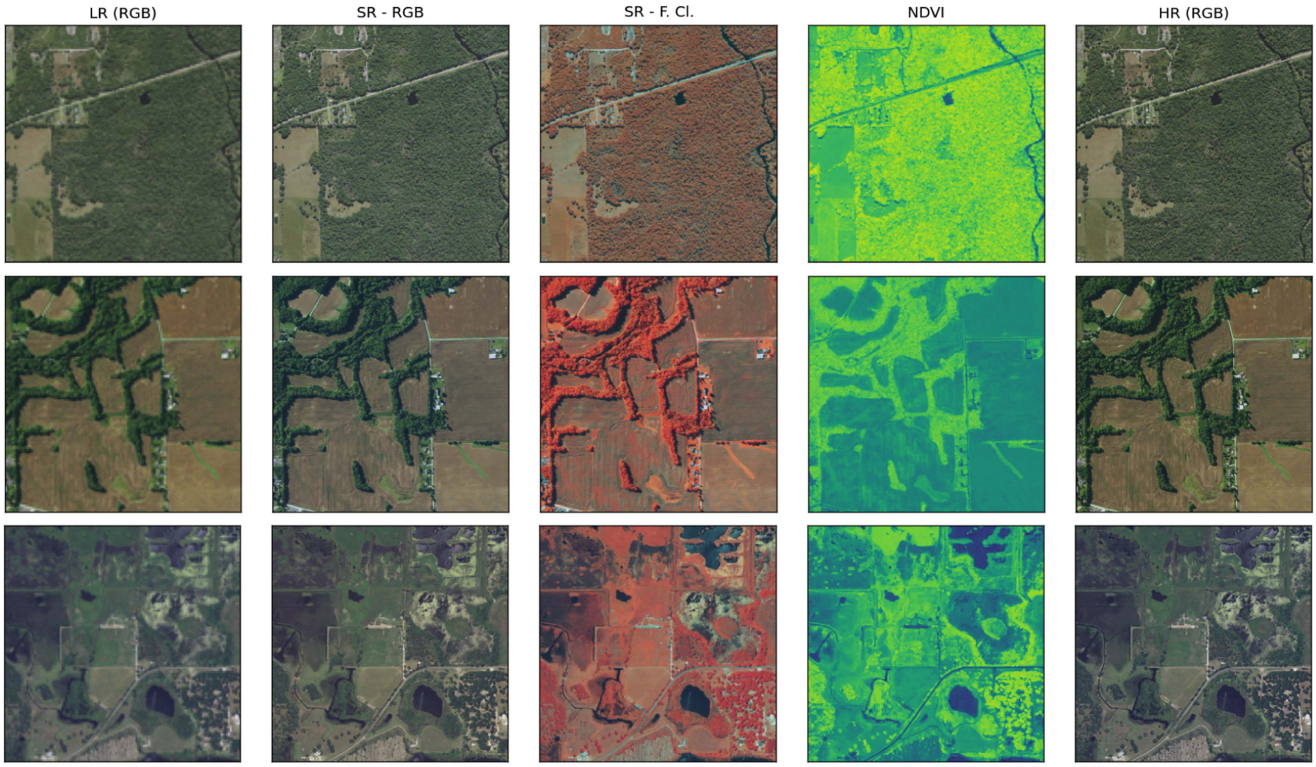

Fig. 6. LDSR-S2 model LR to HR reconstruction quality on test set of SEN2NAIP, additionally showing a false color visualization of the SR product including the NIR band and an NDVI visualization.

quality on the test set of the SEN2NAIP dataset. The quality of the reconstruction is excellent, both in textured areas and in high-frequency regions. The spectral signature is visually similar, and the false-color image and NDVI index show a clear correlation with vegetated regions. Compared to the LR version, the image quality is clearly increased and similar to the HR version. This highlights that the model has learned the distribution of the SEN2NAIP dataset very accurately, showing exceptional SR quality. In order to provide a comparison, our recently proposed *opensr-test* package [58] is used to perform visual and quantitative analysis. It allows the implementation and comparison of different state-of-the-art SR models on various datasets. The LR images of all the validation sets are the S2 and S2-like versions, since we intend to run the model only on S2. The HR versions are sourced from the NAIP program, VEN μS, and SPOT-6. Using different HR datasets and then averaging the results leads to a fairer and broader comparison, since the model does not overfit to a particular HR type, producing, e.g., NAIP-like SR images with NAIP-like characteristics only. The validation set contains worldwide imagery with a large diversity of land covers, vegetation, and climate zones.

For visual comparison, we show a diverse selection of SR models that have been trained for different purposes in order to give a general overview of the products of the RS-SR model (Fig. 7).

Based on the synthetic SEN2NAIP data, which has been spectrally and spatially adapted to match S2 characteristics, the output of the LDSR-S2 is compared with several models (Fig. 7): AllenAI's Satlas SRGAN model [59], EvoLand SR [19], SR4RS [17], and a Swin Transformer adapted for SR [60]. Furthermore, we compared the results to a pixel-space diffusion model and the baseline SR LDM [21], without the adaptations made for LDSR-S2. It should be noted that EvoLand performs a factor of 2$\times$ SR, which has been then interpolated to the 2.5-m resolution to match the other models, which limits the usefulness of the comparison of the spectral metrics. Although some of these models have other SR scale factors, use only RGB bands, or are trained for slightly different applications, it is useful to have visual comparisons of a wide range of models that were trained for specific RS purposes.

The Satlas model exhibited apparent deficiencies that could not be conclusively attributed to generalization errors or specific implementation issues, since it is used here in the single image SR configuration, while the original results are published for multiimage SR [57]. The EvoLand model, as intended specifically, maintained spectral fidelity to input imagery remarkably well, although it prioritizes edge sharpening over the inclusion of *generated* high-frequency SR details. The SR4RS model excelled in urban environments, capturing intricate details effectively; however, it struggled with spectral adherence and maintaining integrity in naturally textured areas. The latent diffusion-based pipeline was not originally trained on RS data and does not incorporate the changes made to adopt the RS requirements. It shows severe limitations in adapting to the unique spectral and radiometric range, as well as the geometric features of RS. Our model, LDSR-S2, distinctly succeeded in introducing detailed textures and structures that closely mirror those in the HR target images, with minimal hallucinations and

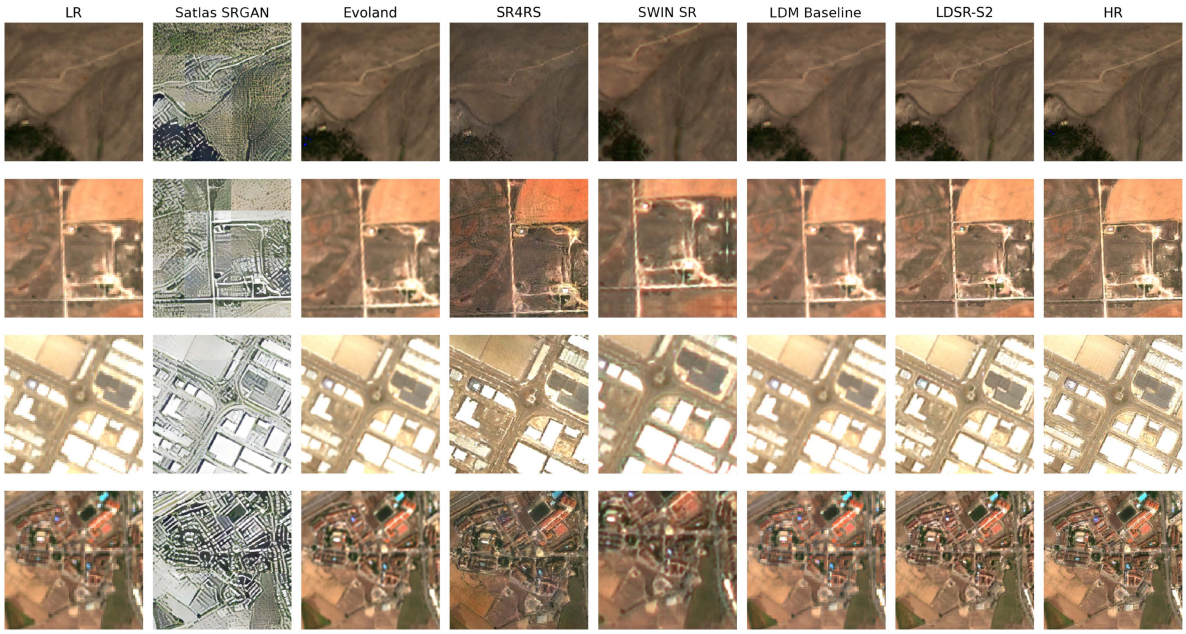

Fig. 7. Comparison of SR performance for several models based on the synthetic *opensr-test* dataset: true color composite (RGB bands) of original Sentinel-2 LR image (left), produced SR images (center), and harmonized HR reference (right).

excellent spectral accuracy. The direct comparison demonstrates that the adaptations and training strategy was successful in adapting LDMs to RS requirements. For the models implemented in the *opensr-test* package, we calculate the performance metrics based on the synthetic dataset SEN2NAIP using the normalized distance metric [58]. The spectral consistency is the only metric that is not normalized, instead being calculated with the spectral angle distance. These models are specifically selected because they are globally applicable and perform SR with a scale factor of $4\times$. To provide a fair comparison, only the RGB bands are taken into account. In addition, we add the LDM baseline model as well as a pixel-space DM to show the improvement of LDSR-S2 over the standard LDM model. The computed quality metrics measure the following properties.

1) *Reflectance Consistency:* Measures the mean absolute error between the reflectance values of the original LR image and the downsampled SR image. This metric assesses how well the SR process preserves the absolute reflectance values, defining its usefulness for downstream EO applications.
2) *Spectral Consistency:* The spectral angle distance serves as a measure to assess the capability of the SR model to retain the spectral characteristics of the LR image [61]. This metric is essential for demonstrating spectral fidelity and adherence to the conditioning image.
3) *Spatial Consistency:* Assesses the spatial alignment and structural integrity by finding matching points in the SR and LR images identified by the LightGlue algorithm and quantifying the error in the affine matrices [62].
4) *Synthesis:* Quantifies the gain in spatial resolution by measuring the high frequencies of the SR image with respect to the LR input image (upsampled using bilinear interpolation to match the SR image spatial resolution). A higher value indicates more introduced spatial detail, regardless of whether these introduced high frequencies are accurate or not.
5) *Omission:* Measures the degree to which the high-frequency features present in the HR image are missing in the SR image, i.e., the SR image is similar in features to the LR input image sampled.
6) *Improvement:* Represents the overall enhancement in high-frequency information in the SR image compared to the HR reference image. Higher improvement scores indicate a more effective SR.
7) *Hallucination:* Evaluates the presence of high-frequency features in the SR image that do not correspond to the real features in the HR image. High hallucination values indicate the generation of unrealistic details.

Table III shows the quality metrics for the models described. As evident from reflectance, spectral, and spatial consistency metrics, LDSR-S2 creates an SR product that closely adheres to the input image characteristics that must be preserved. Those metrics are especially important for maintaining relevant information for downstream EO tasks. The synthesis metrics are higher for the other models, which indicates that these models introduce more high-frequency information, but as evident by the hallucination metric the other models also introduce more information which does not correlate with the actual HR imagery. The omission metric is also lower (i.e., the SR is not similar to the LR input) for the GAN models, but the combination of a high-hallucination metric suggests that the frequencies introduced are not necessarily accurate. The overall improvement, as represented by the improvement metric, shows that the LDSR-2 output has the smallest distance to the HR image, introducing a moderate but accurate amount of high-frequency information, resulting in closer resemblance to the ground truth. The metrics highlight that successfully trained RS SR models perform better than the baseline computer vision models, including

TABLE III
COMPARISON ON THE *OPENSR-TEST* DATASETS AND SR MODELS

	LDM baseline	Pixel-Space Diffusion	LDSR-S2 model	Satlas SRGAN	SR4RS model	SWIN Transformer
Reflectance ↓	0.029	0.089	**0.003**	0.049	0.027	0.005
Spectral ↓	9.68	1.43	**1.26**	12.11	3.40	1.56
Spatial ↓	0.07	0.50	**0.01**	0.27	0.92	0.02
Synthesis ↑	**0.027**	0.013	0.007	0.022	0.013	0.005
Hallucination ↓	0.60	0.73	**0.34**	0.79	0.66	0.39
Omission ↓	0.31	0.13	0.46	**0.11**	0.20	0.59
Improvement ↑	0.09	0.14	**0.21**	0.10	0.14	0.11
PSNR ↑	30.82	36.15	**37.23**	10.78	33.55	34.20
SSIM ↑	0.63	0.83	**0.89**	0.23	0.67	0.84

Comparison on the Opensr-Test Datasets and SR Models, best results highlighted in bold.

a Swin Transformer trained on the same data. Furthermore, the comparison with the baseline model shows the improvement performance of the LDSR-S2 model, even for only the RGB bands. In general, the metrics show that the LDSR-S2 model is more careful with the introduction of details, erring on the side of caution in order not to introduce hallucinations. The results might lack some detail, but are spatially, spectrally, and detailwise closer to the input imagery and HR space, producing a result fit for further EO analysis. LDSR-S2 outperforms both the baseline LDM and the *pixel-space* DM, clearly indicating that the proposed adaptations achieve their intended improvements. Moreover, performing denoising within a compressed latent space introduces no measurable drawbacks compared to pixel-space diffusion. This suggests that compression primarily affects uniform areas, thus preserving the ability to reintroduce high-frequency details into the reconstructed images. Finally, the reduced computational cost of LDMs makes large-scale SR inference on EO datasets both practical and efficient.

2) Generalization to S2 Imagery: The application scenario requires that the model trained on synthetic SEN2NAIP data be applied to S2 imagery (Fig. 8). Due to the general difficulty of extrapolating beyond the training set of diffusion models in combination with the assumptions made while creating the training dataset, we see some degraded reconstruction quality in the S2 SR products compared to the synthetic dataset. The SEN2NAIP dataset incorporates specific agricultural practices, urban development patterns, and surface reflectance characteristics predominant in the US context. One of the primary challenges is the model's ability to handle the diverse spectral and spatial characteristics inherent in global S2 imagery. These characteristics include different types of vegetation, soil, water bodies, and especially built environments, each presenting unique spatial signatures and patterns. The model is able to adapt and reproduce the spectral signature, but especially unseen spatial features pose a challenge at inference time [63].

As outlined in Section III, the WorldStrat dataset is not suitable for training in this context, but we can use its original S2-LR images to compute worldwide validation metrics. We use the entire dataset to compute and average the *opensr-test* metrics (Table IV) and differentiate between geographical context and land cover. The absolute metrics are not necessarily reflective of the performance due to the mismatch between the LR and HR versions, but the relative performance enables conclusions about the generalization performance to different regions and image contents.

TABLE IV
OPENSR-TEST METRICS ON THE *WORLDSTRAT* DATASET

Metric	Africa	Asia	Europe	N. America	Oceania	S. America
Reflectance ↓	0.003	0.005	0.005	0.007	0.004	0.004
Spectral ↓	0.908	1.447	1.557	1.749	1.201	1.307
Synthesis ↑	0.007	0.011	0.011	0.014	0.008	0.008
Hallucination ↓	0.185	0.160	0.163	0.177	0.177	0.183
Omission ↓	0.722	0.757	0.756	0.731	0.727	0.723
Improvement ↑	0.093	0.083	0.080	0.092	0.096	0.094

TABLE V
APPROXIMATE INFERENCE TIMES FOR DIFFERENT DIFFUSION SR METHODS IN HOURS AND MINUTES FOR A SINGLE SENTINEL-2 TILE (110 × 110 KM) ON 4×A100 GPU, USING 100 DDIM SAMPLING STEPS

Method	Inference Time (hh:mm)
Pixel-Space DDPM Sampling	1656:50
Pixel-Space DDIM Sampling	96:35
Latent-Space DDPM Sampling	06:20
Latent-Space DDIM Sampling - LDSR-S2	0:22

Fig. 9 shows the model performance in terms of SSIM and PSNR for different continents and land cover classes [64]. Quite logically, more uniform land cover classes, such as water, grassland, and agriculture are easier to SR and therefore have better values than settlements. The model performs similarly on all continents, indicating that the SR performance is limited more by type of information than location, showing that the model can be generalized to world-wide application scenarios. Fig. 8 demonstrates this, showing a good visual reconstruction for rural and desert areas. Mixed-use residential and industrial areas, as well as densely populated urban areas, show good quality of reconstruction, but good metrics are much harder to achieve in these images with a high-spatial frequency information content.

C. Inference Speed

The entire model contains around 170 million parameters, which is significantly more than other methodologies, such as GANs. The additional requirement of performing a large number of denoising steps to produce one SR result, where the input has to traverse the denoiser once for each step, further increasing computational complexity. Only the switch from DMs to LDMs and using DDIM sampling make processing of continental-scale regions possible. Table V shows the inference times for a RGB-NIR S2 tile.

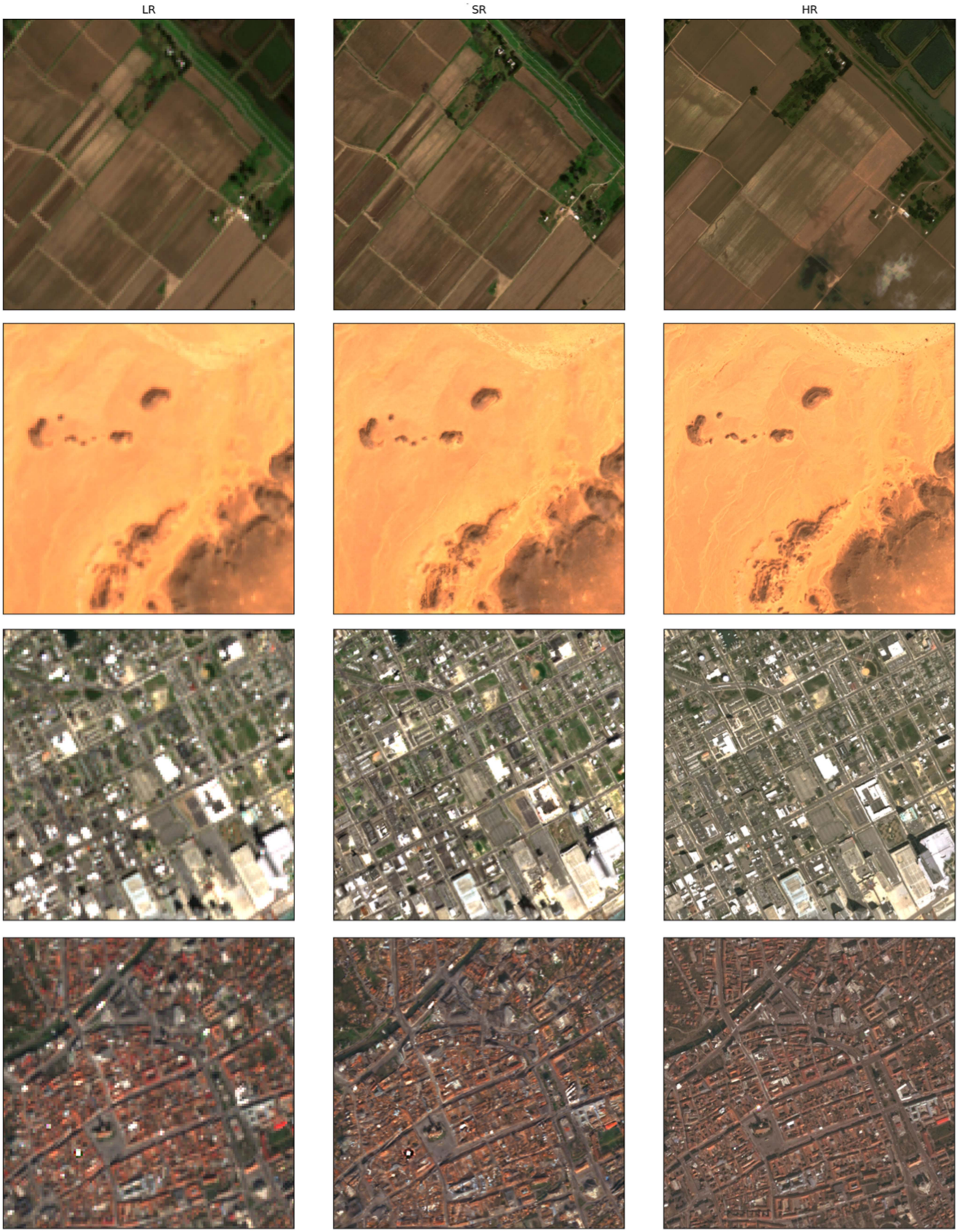

Fig. 8. RGB of SR inference on original S2-LR image compared to SPOT-6 pansharpened HR, from *WorldStrat* validation dataset.

Fig. 9. PSNR and SSIM over *WorldStrat* validation set, grouped by Continent and IPCC land-cover class.

VI. Uncertainty Estimation

SR workflows inherently struggle with the challenge that the problem at hand is ill-posed, which results in a very large range of possible, plausible, and probable reconstructions. The ambiguity arises from the finite nature of the dataset [65], which itself cannot perfectly represent the actual distribution of the data, only an approximation, resulting in a bias toward the training dataset and therefore generalization problems at inference time. Diffusion models have been analyzed using multigenerational statistical measures [66], which we now employ in an LDM framework.

Given the probabilistic framework of diffusion models, our approach seeks to identify and extract the most probable result from the distribution of possibilities. By equipping researchers with the ability to evaluate the reliability and precision of super-resolved features and objects of interest, we establish a foundation for trustworthy SR. Users can discern and quantify how well objects and features they prioritize are represented in the enhanced imagery, thus aligning the model's output with practical reliability and applicability in real-world scenarios. SR products can only be of benefit if we can convince end-users that these products are valuable and useful.

We denote uncertainty estimation as a function CI that takes as input an LR image x and outputs a confidence interval $\mathrm{CI}(x)$. The $\mathrm{CI}(x)$ is provided for each of the super-resolved pixels $\hat{y}$ and is defined as $\mathrm{CI}(x) = [\hat{y} - \epsilon, \hat{y} + \epsilon]$. The ϵ is a parameter that controls the width of the confidence interval. To provide the interval with statistical soundness, we assume that the distribution of the super-resolved pixels is Gaussian. We then use the standard deviation of the distribution to estimate ϵ as $\epsilon = \sigma \cdot \alpha$, where σ is the standard deviation of the distribution, and α is a hyperparameter that controls the width of the confidence interval. By default, we set $\alpha = 1$. The uncertainty map is therefore the pixelwise confidence interval over n generations of the LR input patch.

Although our method is relatively simple, it demonstrates significant efficacy. We observe that the generated intervals exhibit a strong correlation with actual hallucinations (Fig. 10). Examples include textured areas, which require wider explorations of the probability distribution, or the edges of complex shapes. This correlation underscores the practical utility of our method in enhancing the reliability of SR tasks. Although this method is powerful, creating the different samples increases the computational requirements 10- to 20-fold compared to standard inference, depending on the number of samples generated.

Fig. 10 shows the confidence intervals for three example images. Mainly high-frequency areas, such as the road in the middle image or field delineations in the lower image show a larger confidence interval, resulting in brighter colors. In these areas, the model has to create a unique solution on where exactly to place the road, line, or texture, resulting in more variability in the outputs. Also, some textured areas like the forest in the lower image can be classified as hallucinations, where information loss needs to be filled by inventing these textures. Although these textures look highly realistic, they do not necessarily exactly reproduce the original texture. An interesting observation can also be made in the upper image, where the faint line diagonally across the field is almost completely lost in the LR representation and is not recreated by the SR model even in multiple SR creations, leading to an omission. Fig. 11 shows an example of the uncertainty product that has been calculated on real S2 input, where there is no HR ground truth available. The uniform water surface has a low diversity and therefore high confidence, while the bridge and especially the building require more generative capability to be super-resolved, which leads to more diversity in the sampling and therefore higher uncertainty scores.

VII. Discussion

A key achievement of this study is the successful demonstration of utilizing diffusion models for multispectral EO SR tasks via latent-space denoising. This method significantly lowers the computational cost compared to pixel-space diffusion, thus facilitating practical applications in real-world EO scenarios. Furthermore, the proposed architecture and training techniques effectively adapt the fundamental LDM to a reliable, explainable, and globally applicable workflow to enhance the spatial resolution of Sentinel-2 images.

We show that LDSR-S2 achieves state-of-the-art performance in terms of visual quality and quantitative metrics. When applied to the SEN2NAIP test dataset, the model excels in reconstructing high-frequency details and maintaining spectral consistency. When applied to Sentinel-2 images, the LDSR-S2 model shows a

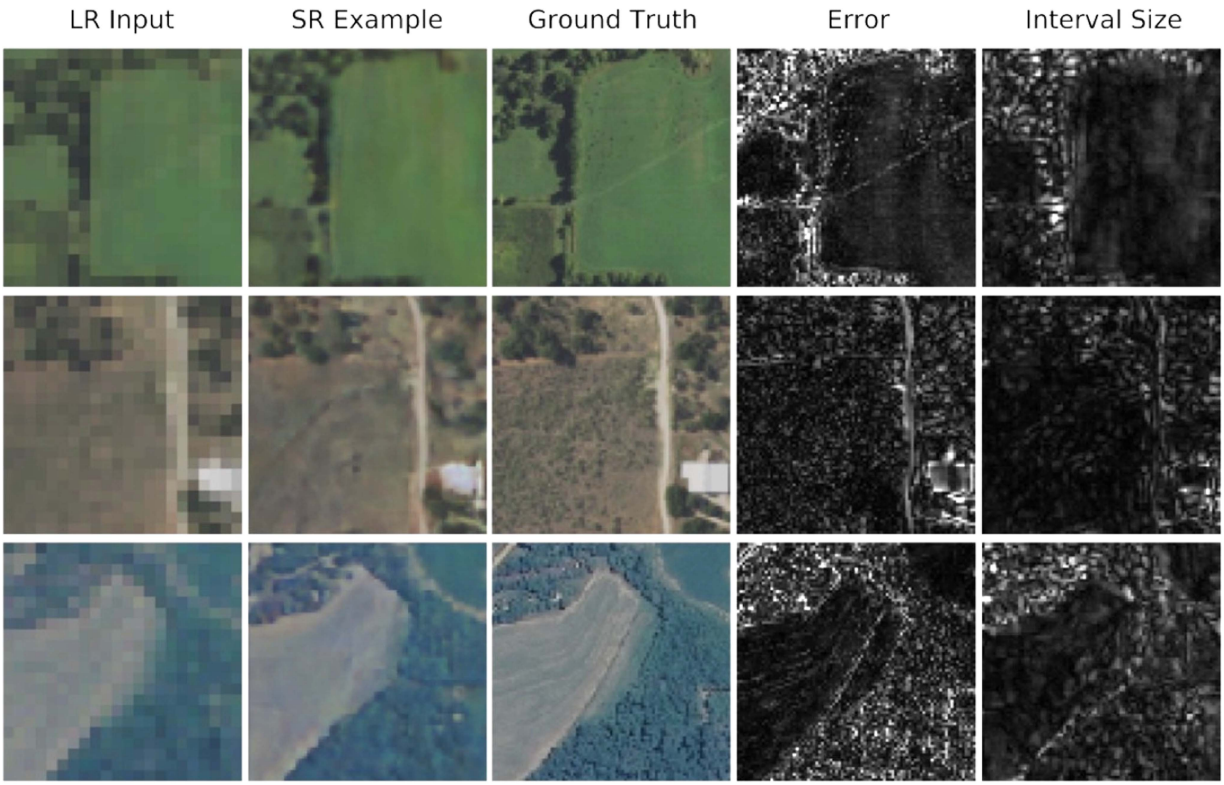

Fig. 10. RGB-bands of the LR, SR, and HR images. Since HR versions of real S2 images are not available, this graphic uses the SEN2NAIP dataset in order to enhance the interpretability (see Fig. 11 for a real S2 example). The interval size image shows the CI explained in Section VI, with lighter pixels showing higher values.

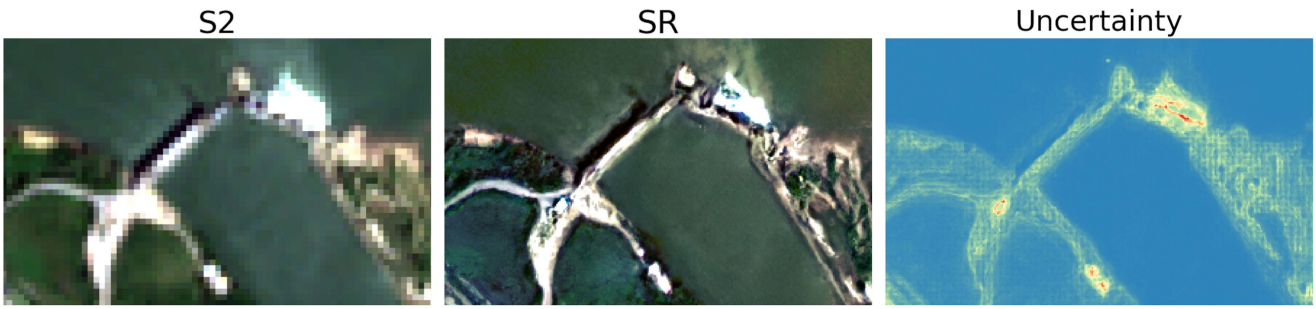

Fig. 11. Example of the uncertainty product delivered with the SR result, based on a real S2 input. The uncertainty values are visualized on a scale from blue (low) to red (high), stretched to the min-max range of the standard deviation.

decrease in performance compared to the synthetic dataset, hinting at generalization issues. However, the model still provides an enhanced product through the injection of high-frequency information and close spectral adherence, making it a valuable tool for applications where HR imagery is otherwise unavailable or too expensive. The probabilistic nature of our diffusion model allows us to provide uncertainty metrics along with the SR product, which expose areas of possible hallucinations to the user in order to increase the trustworthiness of the product.

While the LDSR-S2 model demonstrates considerable promise in enhancing multispectral SR, the generalization of these results to real-world Sentinel-2 imagery is presently constrained by the quality and diversity of the training data used. To advance the robustness and applicability of our model, a crucial next step involves expanding our dataset to include a wider array of real-world scenarios, which will help improve the spectral and spatial harmonization between synthetic and actual Sentinel-2 datasets.

In addition, while our uncertainty metrics provide valuable insight, further refinement and validation are necessary to ensure their accuracy and usability across different applications. The only conclusive proof lies within a measurable performance increase in downstream applications, which is out of the scope of this article. Moreover, integrating the other S2 spectral bands beyond RGB-NIR could provide more comprehensive solutions for a wider range of EO applications. The LDSR-S2 model could additionally benefit from a multiimage approach. Our future efforts will focus on overcoming these data constraints,

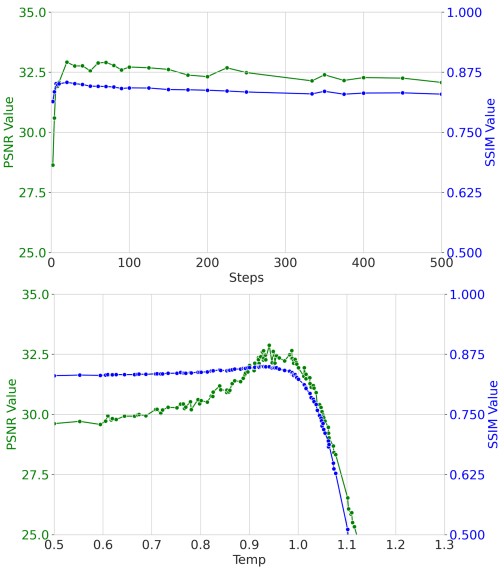

Fig. 12. Inference hyperparameter ablation study of number of inference steps (top) and sampling temperature (bottom).

which currently serve as the primary bottleneck, to unlock the full potential of SR technologies in global EO tasks.

VIII. Conclusion

The proposed LDSR-S2 model presents a significant advancement in the field of RS SR. Our adaptations enable the usage of diffusion models in EO tasks, forcing closer adherence to the input imagery, expanding to more spectral bands, and improving computational efficiency compared to previous RS SR diffusion models. The resulting product outperforms state-of-the-art models. By taking advantage of the generative power of LDMs, we can provide high-quality, HR imagery from low-resolution inputs which incorporate the NIR band to enable RS-typical downstream applications. The probabilistic nature of the model enables the provision of pixelwise uncertainty maps, which can help users to determine the usability of the product to their application. Our approach balances computational efficiency with output quality, making it a practical solution for large-scale EO applications. The method's significance lies in its potential to augment EO practices by providing high-spatial, high-temporal frequency image data at a fraction of the usual cost, making advanced RS capabilities accessible to a wider range of users and applications.

VIII. Appendix A
Sampling Parameters Ablation

Fig. 12 plots the number of inference steps and the sampling temperature against the SSIM and PSNR performance metrics, allowing the identification of optimal parameter values. About 100 inference steps offer a good balance between inference time and performance metrics. For the temperature parameter, we can see that both metric graphs peak around 0.95, indicating that a slightly scaled-down amount of noise compared to the training noise leads to a more truthful reconstruction.

VIII. Resources

Models, weights, dataset, reproducible examples, and the testing procedure both as *pip* packages and their source code are provided at the following.

1) Model: https://github.com/ESAOpenSR/opensr-model
2) Data: https://huggingface.co/datasets/isp-uv-es/SEN2NAIP
3) OpenSR-Utils: https://github.com/ESAOpenSR/opensr-utils
4) OpenSR-Test: https://github.com/ESAOpenSR/opensr-test

Acknowledgment

The Sentinel-2 Level 2 A and NAIP data were provided by the European Space Agency (ESA) and USDA, respectively. The authors would like to thank Nicolas Longépé for his collaboration and support in the framework of the ESA Φ-lab OpenSR project: *Explainable AI: Application to Trustworthy Super-Resolution (OpenSR)*.

References

[1] P. Kansakar and F. Hossain, "A review of applications of satellite Earth observation data for global societal benefit and stewardship of planet Earth," *Space Policy*, vol. 36, pp. 46–54, 2016.
[2] K. Anderson, B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl, "Earth observation in service of the 2030 agenda for sustainable development," *Geo-Spatial Inf. Sci.*, vol. 20, no. 2, pp. 77–96, 2017.
[3] F. Gascon et al., "Copernicus sentinel-2a calibration and products validation status," *Remote Sens.*, vol. 9, no. 6, 2017, Art. no. 584.
[4] J. A. Benediktsson, J. Chanussot, and W. M. Moon, "Very high-resolution remote sensing: Challenges and opportunities [point of view]," *Proc. IEEE*, vol. 100, no. 6, pp. 1907–1910, Jun. 2012.
[5] D. A. Quattrochi and M. F. Goodchild, *Scale in Remote Sensing and GIS*. Boca Raton, FL, USA: CRC press, 1997.
[6] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 129–152, 2023.
[7] J. Kong et al., "Super resolution of historic Landsat imagery using a dual generative adversarial network (GAN) model with CubeSat constellation imagery for spatially enhanced long-term vegetation monitoring," *ISPRS J. Photogrammetry Remote Sens.*, vol. 200, pp. 1–23, 2023.
[8] P. M. Atkinson and P. J. Curran, "Choosing an appropriate spatial resolution for remote sensing investigations," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 12, pp. 1345–1351, 1997.
[9] S. H. Tabesh, M. Babadi Ataabadi, and D. Chen, *Advancements in Deep Learning-Based Super-Resolution for Remote Sensing: A. Comprehensive Review and Future Directions*. Cham, Switzerland: Springer, 2024, pp. 51–91.
[10] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.
[11] C. Li et al., "Casformer: Cascaded transformers for fusion-aware computational hyperspectral imaging," *Inf. Fusion*, vol. 108, 2024, Art. no. 102408. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524001866
[12] J. Hu, T. Huang, and L. Deng, "Fusformer: A transformer-based fusion approach for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021,
[13] J. He, Q. Yuan, J. Li, Y. Xiao, X. Liu, and Y. Zou, "Dster: A dense spectral transformer for remote sensing spectral super-resolution," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, 2022, Art. no. 102773. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030324342200099X
[14] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," 2017, *arXiv:1703.09470*.
[15] X. He et al., "Pan-Mamba: Effective pan-sharpening with state space model," *Information Fusion*, vol. 115, Art. no. 102779.

[16] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019.

[17] R. Cresson, "SR4RS: A tool for super resolution of remote sensing images," *J. Open Res. Softw.*, vol. 10, no. 1, 2022.

[18] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16772–16782.

[19] EvoLand, "Sentinel 2 SR," GitHub, 2024, Accessed: Jun. 21, 2024. [Online]. Available: https://github.com/Evoland-Land-Monitoring-Evolution/sentinel2_superresolution

[20] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[22] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 8–10.

[23] R. Tanno et al., "Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI," *NeuroImage*, vol. 225, 2021, Art. no. 117366.

[24] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learning-based single image super-resolution," *PeerJ Comput Sci.*, vol. 62, 2021, Art. no. e621.

[25] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 305–319, 2018.

[26] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Sci. Rev.*, vol. 232, 2022, Art. no. 104110.

[27] P. L.-C. R. Fernandez-Beltran and F. Pla, "Single-frame super-resolution in remote sensing: A practical overview," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, 2017.

[28] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[29] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[30] C. Ayala, C. Aranda, and M. Galar, "Guidelines to compare semantic segmentation maps at different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4702416.

[31] A. Okabayashi, N. Audebert, S. Donike, and C. Pelletier, "Cross-sensor super-resolution of irregularly sampled Sentinel-2 time series," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, pp. 502–511.

[32] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, "Diffusion model with detail complement for super-resolution of remote sensing," *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4834.

[33] L. Han et al., "Enhancing remote sensing image super-resolution with efficient hybrid conditional diffusion model," *Remote Sens.*, vol. 15, no. 13, 2023, Art. no. 3452.

[34] X. Lu et al., "Effective variance attention-enhanced diffusion model for crop field aerial image super resolution," *ISPRS J. Photogrammetry Remote Sens.*, vol. 218, pp. 50–68, 2024.

[35] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[37] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.

[38] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative AI: A survey," 2023, *arXiv:2303.07909*.

[39] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, 2023.

[40] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[41] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution," 2022.

[42] R. Dong, L. Mou, L. Zhang, H. Fu, and X. X. Zhu, "Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 191, pp. 155–170, 2022.

[43] Z. Qiu, H. Shen, L. Yue, and G. Zheng, "Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 199, pp. 226–241, 2023.

[44] A. Kuznetsova et al., "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, 2020.

[45] C. Aybar, D. Montero, J. Contreras, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "SEN2NAIP: A large-scale dataset for Sentinel-2 Image Super-Resolution," *Sci. Data*, vol. 11, Dec. 2024, Art. no. 1389.

[46] C. Aybar, D. Montero, J. Contreras, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "SEN2NAIP: A large-scale dataset for Sentinel-2 Image Super-Resolution," 2024, Accessed: Jun. 21, 2024. [Online]. Available: https://huggingface.co/datasets/isp-uv-es/SEN2NAIP

[47] V. T. Sambandham, K. Kirchheim, F. Ortmeier, and S. Mukhopadhaya, "Deep learning-based harmonization and super-resolution of Landsat-8 and Sentinel-2 images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 212, pp. 274–288, 2024.

[48] D. P. Kingma et al., "An introduction to variational autoencoders," *Foundations Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[49] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[50] L. Kantorovich, "On the translocation of masses," *Doklady Akademii Nauk SSSR*, vol. 37, no. 7/8, pp. 199–201, 1942.

[51] C. Villani, *The Wasserstein Distances*. Berlin, Heidelberg, Germany: Springer, 2009, ch. 6, pp. 93–111, doi: 10.1007/978-3-540-71050-9_6.

[52] S. Kolouri, C. E. Martin, and G. K. Rohde, "Sliced-Wasserstein autoencoder: An embarrassingly simple generative model," 2018, *arXiv:1804.01947*.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] C. Luo, "Understanding diffusion models: A unified perspective," 2022, *arXiv:2208.11970*.

[56] C. Saharia et al., "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH Conf. Proc.*, 2022, pp. 1–10.

[57] M. T. Razzak, G. Mateo-García, G. Lecuyer, L. Gómez-Chova, Y. Gal, and F. Kalaitzis, "Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 1–13, 2023.

[58] C. Aybar, D. Montero, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "A comprehensive benchmark for optical remote sensing image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5003105.

[59] P. Wolters, F. Bastani, and A. Kembhavi, "Zooming out on zooming in: Advancing super-resolution for remote sensing," 2023, *arXiv:2311.18082*.

[60] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[61] F. Kruse et al., "The spectral image processing system (sips) – interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 145–163, May 1993, doi: 10.1016/0034-4257(93)90013-N.

[62] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023.

[63] A. Malczewska and M. Wielgosz, "How does super-resolution for satellite imagery affect different types of land cover? Sentinel-2 case," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 340–363, 2024.

[64] Food Agriculture Org. United Nations, Rome, Italy, "World census of agriculture," 1999, land-use categories recognized; FAO (1986, 1995a); FAO/UNEP (1999). [Online]. Available: http://www.fao.org

[65] D. Draper, "Assessment and propagation of model uncertainty," *J. Roy. Stat. Soc.: Ser. B. (Methodological)*, vol. 57, no. 1, pp. 45–70, 1995.

[66] E. Horwitz and Y. Hoshen, "Conffusion: Confidence intervals for diffusion models," 2022, *arXiv:2211.09795*.

[67] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, U.K.: Addison-Wesley, 1999.