A Comprehensive Benchmark for Optical Remote Sensing Image Super-Resolution

Cesar Aybar, David Montero, Simon Donike, Freddie Kalaitzis, Luis Gómez-Chova

***Abstract*—In recent years, there has been a growing interest in using image super-resolution (SR) techniques in remote sensing. These techniques aim to reconstruct high-resolution (HR) imagery from low-resolution (LR) sources. Despite the development of sophisticated SR methodologies, determining what constitutes 'good' SR is still a matter of debate. Present-day literature often presents SR models through a strong computer vision perspective, heavily relying on synthetic datasets. Moreover, commonly used metrics often prioritize attributes that do not necessarily correspond to improvements in spatial resolution. To address this challenge, we present *OpenSR-test*, a comprehensive benchmark designed exclusively for evaluating SR of remote sensing images. Our framework incorporates specific quality metrics and curated cross-sensor datasets, each spanning various scale factors with consistent metadata. Utilizing *OpenSR-test*, we evaluate state-of-the-art SR algorithms from a remote sensing perspective. The *OpenSR-test* framework and datasets are publicly available at https://esaopensr.github.io/opensr-test/.**

***Index Terms*—super-resolution, benchmarking, deep learning, datasets, Sentinel-2, NAIP, SPOT.**

I. Introduction

Optical remote sensing is a valuable source of information that predominantly uses satellite sensors to gather data of the Earth's surface. The resolution of these sensors is determined by their ability to distinguish changes over time (i.e. temporal resolution), different wavelengths of the electromagnetic spectrum (i.e. spectral resolution), and spatially separated objects (i.e. spatial resolution). Depending on the application, remote sensing users may require higher resolution in one or more of these dimensions. For instance, in crop monitoring, a high temporal resolution may be preferred to track crop phenology and harvesting, while in the case of building delineation, a very high spatial resolution may be preferred to capture the fine details of individual structures. However, remote sensing sensors do not always provide the required resolution due to technical and economic constraints.

In scenarios where spatial high-resolution images are not available, super-resolution (SR) algorithms emerge as a prominent solution [1]. SR is inherently an inverse problem, as it reconstructs the original high-resolution (HR) image from its degraded low-resolution (LR) counterpart. Moreover, it is characterized as an ill-posed problem, given the one-to-many relation between LR and possible HR representations. Over recent years, substantial research efforts have been dedicated to the super-resolution of natural images, particularly by employing deep-learning methods [1]. Current research has been directed towards developing more efficient architectures that integrate local and global features [2], obtaining more realistic degradation kernels [3], producing SR images over different scales [4], and proposing ingenious loss functions [5].

Despite significant advancements, a crucial question remains in the remote sensing community: Can SR methods designed for natural images be effectively applied to satellite imagery? These SR methods use artificially degraded images through basic bicubic or bilinear interpolation as input. This approach circumvents the need for 'real' LR images and results in LR-HR pairs that are spatially and spectrally consistent [3]. However, this approach may overlook the unique characteristics of the satellite sensors, potentially introducing a domain bias between synthetically generated and real LR observations. To mitigate this issue, recent advancements in SR methods have included simultaneous image fusion and SR processes [6], though these have shown limitations in preserving reflectance coherence, as recently noted in [5]. Alternatively, some studies have explored implementing harmonized cross-sensor SR approaches. Nevertheless, the preparation of training datasets for these approaches faces significant limitations due to the required preprocessing, including spatial alignment, removal of cloud-contaminated pixels, spatial coregistration, and radiometric adaptation [7], which ultimately constrains the dataset size and, consequently, the scalability of the model.

Another concern is the need for quality metrics explicitly designed for measuring the ground-resolved distance (GRD), a key indicator of spatial image resolution. While previous studies have often relied on pixel-wise metrics, like PSNR, and perceptual metrics, such as SSIM and LPIPS [8], these metrics have limitations. On the one hand, pixel-wise metrics can be too sensitive to spatial translations and local reflectance disparities unrelated to the super-resolution problem [9]. On the other hand, perceptual metrics are not specifically intended for super-resolution. They assess the general image quality based on human perception standards. Therefore, these metrics may not be the most accurate solution for domain-specific contexts like remote sensing, which requires specialized expertise [10].

This letter introduces a benchmark for comprehensively evaluating super-resolution in optical remote sensing images: *OpenSR-test*. The benchmark comprises three distinct curated cross-sensor datasets with varying scale factors. In addition,

C. Aybar, S. Donike, and L. Gómez-Chova are with the Image Processing Laboratory (IPL), University of Valencia, Valencia, Spain. D. Montero is with the Remote Sensing Centre for Earth System Research (RSC4Earth), Leipzig University, Leipzig, Germany. F. Kalaitzis is with the Department of Computer Science, University of Oxford, UK. Corresponding author: cesar.aybar@uv.es.

This work has been funded by the European Space Agency (ESA, Φ-lab, OpenSR project). C.A. acknowledges support by CONCYTEC, Peru ("Proyectos de Investigación Básica – 2023-01" program, PE501083135-2023-PROCIENCIA). L.G.-C. acknowledges support from the Spanish Ministry of Science and Innovation (project PID2019-109026RB-I00 funded by MCIN/AEI/10.13039/501100011033).

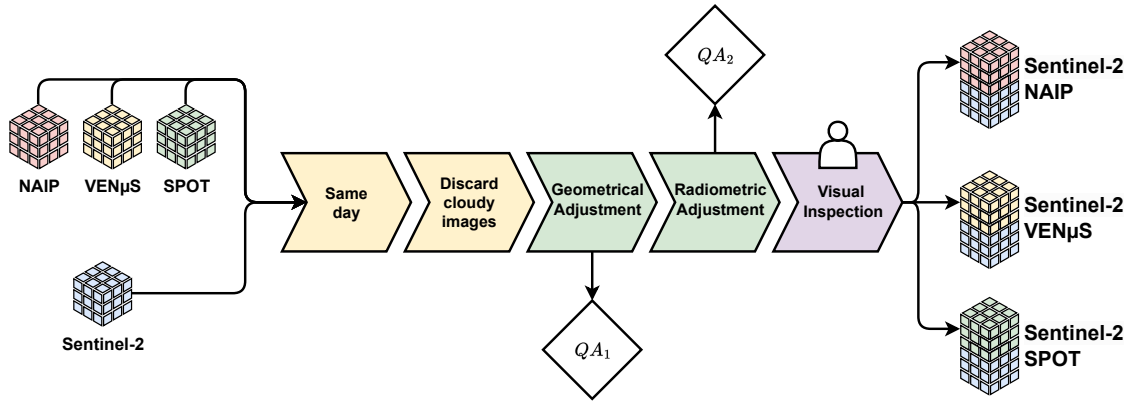

Fig. 1. Workflow for dataset generation: image selection (yellow blocks), harmonization (green blocks), and human inspection (purple block). For each LR-HR pair within each dataset, we report one spatial quality metric (QA_1) and one spectral quality metric (QA_2).

we introduce a set of tailored quality metrics based on a remote sensing super-resolution protocol. To evaluate the proposed approach, we showcase its use by detailing the accuracy assessments of three state-of-the-art SR methods and conducting an in-depth analysis of the resulting metrics. Finally, we conclude the letter by underscoring the importance that robust validation practices can play in shaping the future of SR in the remote sensing field.

II. Cross-sensor Multi-scale Datasets

One of the primary goals of *OpenSR-test* is to prepare cross-sensor multi-scale datasets with the maximum possible consistency between LR and HR image pairs. We selected Sentinel-2 Level-2A (10m bands) as the input LR imagery and three different pre-processed image sources to define the target HR domain: NAIP (2.5m), SPOT (2.5m), and VENμS (5m). The entire process can be divided into three steps, as depicted in Fig. 1: selection of potential LR-HR pairs, harmonization, and visual inspection. For the LR-HR pair selection, we discarded all the LR-HR pairs that were not acquired on the same day to ensure similar atmospheric conditions. LR-HR pairs with cloud cover in the LR image were automatically discarded using a cloud detection algorithm trained on the CloudSEN12 dataset [11].

The harmonization phase comprises two key tasks: improving spatial collocation and radiometric matching. For spatial alignment assessment, ground control points between each LR and HR image pair are automatically identified using the LightGlue and DISK algorithms [12], [13]. Following the approach outlined in SEN2VENμS [7], we exclude any erroneous correspondence of points with a displacement greater than 30 meters (3 pixels). The remaining points are used for calculating the mean square collocation error (quality flag QA_1 in Fig. 1). After collocating the LR-HR images, any image pair with a $QA_1 > 0.75$ is removed from the dataset. For radiometric harmonization, we apply histogram matching individually for each band and then compute the average spectral angle distance (SAD) between the LR-HR images (quality flag QA_2 in Fig. 1). LR-HR pairs with a $QA_2 > 5$ are not considered.

Finally, to ensure the optimal quality of the dataset, we carefully inspect the remaining LR-HR pairs through a manual review process. Any LR-HR pairs containing saturated or defective pixels, or other noticeable inconsistencies detected through visual inspection are excluded. The final datasets are accessible on HuggingFace at https://huggingface.co/datasets/isp-uv-es/opensr-test. Furthermore, for each LR-HR pair in the datasets, we report the spatial (QA_1) and spectral (QA_2) errors. The mean values of these quality metrics are presented in Table I for the three datasets (the lower the better).

TABLE I
Datasets available in *OpenSR-test*. The HR and LR pairs consist of 4 bands: red, green, blue, and near infrared. The width of the Gaussian degradation kernel is estimated by comparing the LR and downsampled HR images.

HR source imagery	Number of scenes	HR patch size (pix.)	Scaling factor	Gauss. kernel width (m)	Quality flags	
					$QA_1 \downarrow$	$QA_2 \downarrow$
NAIP [14]	30	1024	×4	5.50	0.33	1.30
SPOT 6/7 [6]	12	512	×4	5.75	0.44	1.21
VENμS [7]	59	512	×2	2.10	0.26	1.57

III. A protocol for remote sensing image SR evaluation

The second main goal of this work is to quantify the quality of SR images. Quantifying the quality of synthesized images is a well-known problem in remote sensing. Twenty-five years ago, Wald et al. [15] introduced a set of properties that an ideal pansharpening algorithm must satisfy to ensure its practical utility. Since then, numerous adaptations have been introduced, leveraging this protocol to tackle a comprehensive range of image enhancement issues, even transcending those related to multispectral data [16], [17]. However, there is still no formal quality assessment protocol for super-resolution in remote sensing. Wald's protocol specifies two core quality properties that algorithms must satisfy: consistency and synthesis. Consistency entails that the LR image can be recovered by degrading the reconstructed (SR) image, although the degradation strategy is usually unknown. Synthesis, on the other hand, states that a reconstructed (SR) image must preserve the characteristics of the original LR image at an

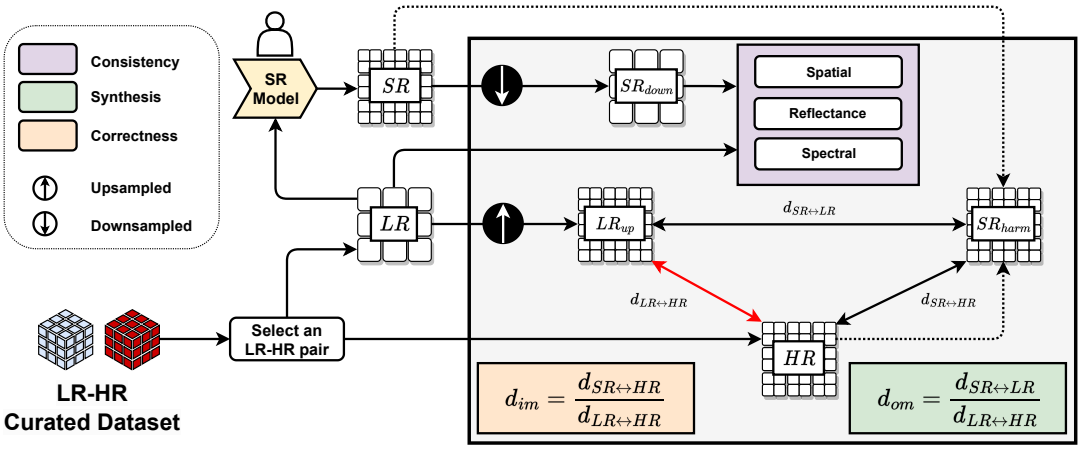

Fig. 2. A high-level summary of our workflow to estimate the consistency, synthesis, and correctness metrics. We propose three metrics to evaluate consistency and three more to assess synthesis and correctness. The dotted lines represent the optional harmonization step conducted before the triple distance process. The red line ($d_{LR\leftrightarrow HR}$) is used to normalize the distances $d_{SR\leftrightarrow LR}$ and $d_{SR\leftrightarrow HR}$.

increased spatial resolution. Inspired by Wald's protocol, we propose a set of properties that can serve as guidelines for designing super-resolution methods, metrics, and datasets, as presented in the previous section.

SR images derived from LR images must satisfy the following three properties:

Consistency property. Any SR image, when degraded (downsampled) to the original LR spatial resolution, SR_{down}, must maintain consistent reflectance values and spatial alignment with its LR counterpart. Testing this property can be challenging because the degradation model to generate LR (i.e., SR_{down}) from SR is typically unknown. However, assuming that the high-frequency characteristics between SR and HR should be similar, a degradation model learned from HR to LR might be employed to degrade SR.

Synthesis Property. Any SR image must improve the effective spatial resolution. In addition, the SR model must preserve the low-frequency details from the original LR image. If this condition is met, the metric of synthesis can be quantified by measuring the distance between the SR image and a reference image above the Nyquist frequency. The reference image can be obtained by simply upsampling the LR image (see next section). The larger the distance between the SR and the LR reference image, the more effective the gain in spatial resolution and detail representation.

Correctness property. Any SR model must avoid hallucinations and omissions. Hallucinations are related to generating high-frequency features that do not align with the real ones present in the HR image. On the other hand, omissions denote the missing high-frequency features that the SR image failed to capture. One practical approach to evaluate the correctness property is to measure the distances between the SR image to the HR and LR images, as we present in the next section.

IV. SR Quality Metrics

The third objective of the *OpenSR-test* framework is to introduce specialized quality metrics that fulfill the described protocol for remote sensing image SR. The proposed quality metrics for SR can be computed at various aggregation levels: individual pixels, patches, or entire images; and are categorized into three distinct groups attending to the described *Consistency*, *Synthesis*, and *Correctness* properties (Fig. 2).

In the realm of *consistency*, our framework introduces three distinct metrics, which are delineated in purple in Fig. 2: *reflectance*, *spectral*, and *spatial* consistency. The core objective here is to evaluate the fidelity with which the SR image retains the intrinsic properties of the LR image. To facilitate this assessment, we downsample the SR image to the exact spatial resolution as the LR image, producing SR_{down}. We generate SR_{down} using bilinear interpolation by default, applying an anti-aliasing filter and then resampling. Firstly, the *reflectance* metric is quantified using the mean absolute error. This metric focuses on determining the extent to which the super-resolution process affects the reflectance values. Secondly, the *spectral* metric aims to ascertain the SR model's capability in preserving the LR image's spectral signatures, an essential factor in maintaining the spectral quality of the image. To compute this metric, we employ the spectral angle distance. Finally, the *spatial* consistency metric is derived by calculating the mean absolute error between the matching points identified by LightGlue in both SR and LR images. This measure is crucial for assessing the spatial alignment and structural integrity of the SR image relative to the LR image.

Before the synthesis and correctness estimation, the OpenSR-test framework performs two sequential steps:

1) The LR image is upsampled to match the HR spatial resolution, creating LR_{up}. Users can choose different methods, but bilinear interpolation is the default.
2) The SR image undergoes a harmonization process (SR_{harm}) to handle systematic errors identified during consistency inspection. The process includes correcting the reflectance values using histogram matching and performing spatial alignment using ground control points. These control points are obtained by applying LightGlue in the SR-HR pair.

With the three images standardized (i.e. three points in a

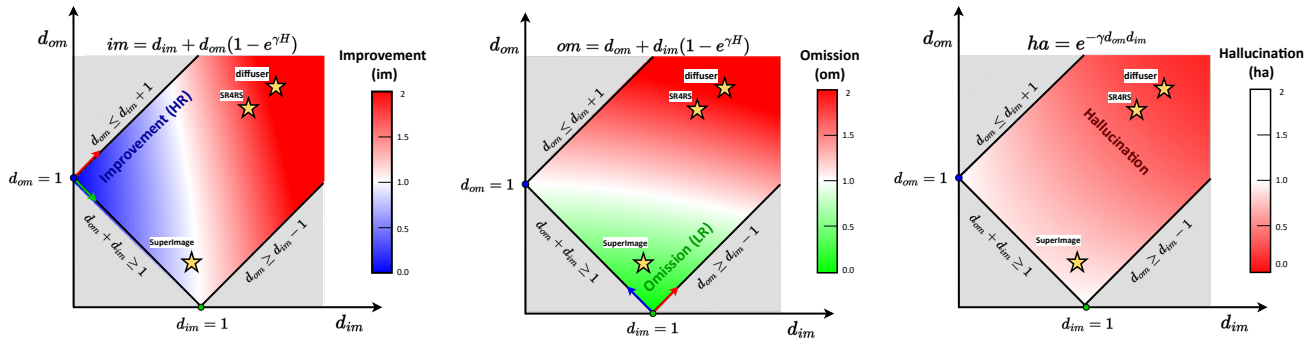

Fig. 3. Visual representation of the three fundamental components within the correctness metric evaluation. The $L1$ norm defines the metric space. The first diagram in the figure delineates the improvement space (im), the second delineates the omission space (om), and the third depicts the hallucination space (ha). In all metrics, a value closer to zero indicates better accomplishment. Each diagram displays three stars representing the SR models' correctness evaluation. Refer to the results section for further information on the SR model setup. The gray areas depict regions where the Euclidean triangle inequality is not satisfied, i.e., where there are no possible solutions.

high-dimensional space defining a triangle), we proceed to distance estimation, as depicted in Fig. 2. First, we calculate the distance between SR_{harm} and LR_{up}, and call it $d_{SR \leftrightarrow LR}$. This distance provides insights into the divergence of the SR image from the naive upsampled LR image (*omission*). Next, we calculate the distance between SR and HR, referred to as $d_{SR \leftrightarrow HR}$. This particular distance evaluates the extent to which the SR image aligns with the content of the reference HR image (*improvement*). Finally, we calculate the distance between LR_{up} and HR, referred to as $d_{LR \leftrightarrow HR}$. This distance indicates the amount of high-frequency information present in the HR image but absent in the LR image.

The default setting for distance calculations in the *OpenSR-test* is the L1 norm. However, users have the option to choose from eight different distance metrics: L2 norm, SAD, Percentage Bias, Inverted Peak Signal-to-Noise Ratio, Kullback-Leibler divergence, and LPIPS. We normalize $d_{SR \leftrightarrow LR}$ and $d_{SR \leftrightarrow HR}$ by the reference distance $d_{LR \leftrightarrow HR}$. This normalization results in relative distances, which we define as omission (d_{om}) and improvement (d_{im}) dimensions, respectively. The d_{om} dimension measures the total gain in high-frequency information and is used to quantify the synthesis property.

Regarding *correctness*, the primary objective is to classify and quantify how much of the high-frequency representation is an improvement (SR close to HR), omissions (SR close to LR), or hallucinations (SR far from both LR and HR). The *OpenSR-test* framework reports the percentage of pixels classified within these categories. Using the d_{om} and d_{im} we estimate the omission (om), improvement (im), and hallucination (ha) scores using the following equations (Fig. 3):

$$H = d_{im} + d_{om} - 1 \tag{1}$$

$$om = d_{om} + d_{im}\left(1 - e^{-\gamma_{om} H}\right) \tag{2}$$

$$im = d_{im} + d_{om}\left(1 - e^{-\gamma_{im} H}\right) \tag{3}$$

$$ha = e^{-\gamma_{ha} d_{im} d_{om}} \tag{4}$$

The parameters γ_{om}, γ_{im}, and γ_{ha} govern the respective domains of the omission, improvement, and hallucination scores. We set γ to 0.8 for both the om and im scores, and to 0.4 for the ha score. These values were determined empirically based on the consensus of three remote sensing experts. The scores for om, im, and ha are used to categorize data units as hallucinations, omissions, or improvements based on the score that registers the minimum value. Refer to Fig. 3 for a visual representation of these dimensions.

V. Results

A. Experimental Setup

We compare three pre-trained SR models using different datasets and architectures to showcase the effectiveness of our benchmark. One of the models, *SR4RS* [18], employed Sentinel-2 images, while the other two models, diffuser and SuperImage, were trained using natural images. The evaluation metrics reported in our comparison only considered the red, green, and blue bands to ensure consistency across all SR model configurations.

Despite the extensive literature on remote sensing super-resolution, our experiment did not include other SR models because only *SR4RS* provides open access to both the weight parameters and the code of their trained models. Pre-trained models available in the SUPER-RES [5] and WorldStrat [6] projects were omitted as they focus on multi-image super-resolution, whereas *OpenSR-test* is designed explicitly for single-image super-resolution.

B. Overall Performance Comparison

The comparative performance analysis of the three super-resolution models, presented in Table II, offers a detailed assessment across the *OpenSR-test* datasets, including NAIP, SPOT, and VENμS.

From a *consistency* perspective, the *SuperImage* and *SR4RS* instances display an acceptable balance in preserving spatial, reflectance, and spectral characteristics across all evaluated datasets. This indicates its suitability for remote-sensing downstream tasks. On the other hand, the *diffusers* model exhibits significant alterations in reflectance and spectral values, along with spatial displacements greater than 1 pixel at low resolution (10 meters), which may suggest a degree of local geometric distortion.

TABLE II
BENCHMARK COMPARISON OF SR MODELS. DOWNWARD ARROWS (↓) DENOTE METRICS IN WHICH LOWER VALUES ARE PREFERABLE, AND UPWARD ARROWS (↑) INDICATE METRICS IN WHICH HIGHER VALUES REFLECT BETTER PERFORMANCE.

		Consistency			**Synthesis**	**Correctness**		
		reflectance ↓	**spectral ↓**	**spatial ↓**	**high-frequency ↑**	**ha ↓**	**om ↓**	**im ↑**
NAIP	**SuperImage**	**0.008**	7.286	**0.131**	0.003	**0.117**	**0.784**	**0.098**
	SR4RS	0.016	**3.471**	1.156	0.010	0.869	0.077	0.054
	diffusers	0.463	12.437	2.88	**0.013**	0.905	0.055	0.040
SPOT	**SuperImage**	**0.009**	3.512	**0.062**	0.006	**0.160**	**0.794**	**0.046**
	SR4RS	0.039	**3.232**	1.151	**0.023**	0.834	0.115	0.051
	diffusers	0.417	11.730	0.817	0.014	0.686	0.251	0.063
VENμS	**SuperImage**	0.009	8.687	**0.099**	0.003	**0.403**	**0.380**	**0.217**
	SR4RS	0.014	**3.394**	1.122	**0.012**	0.971	0.017	0.012
	diffusers	0.467	13.303	0.806	0.009	0.933	0.043	0.024

From a *synthesis* perspective, *SR4RS* and *diffusers* models exhibit close scores. This outcome aligns with expectations, as generative models like *SR4RS* and *diffusers* tend to introduce more high-frequency energy into images compared to discriminative models such as *SuperImage*. Notably, *SuperImage* introduces approximately two times less high-frequency information, indicating a more conservative approach.

Lastly, in terms of *correctness*, the generative models (*SR4RS* and *diffuser*) demonstrate a higher incidence of hallucinations across all datasets (Fig. 3). These models tend to introduce artifacts that, while visually appealing, do not accurately correlate with the HR data. Conversely, the *SuperImage* model tends more towards the omission space (Fig. 3), implying that the added high-frequencies are similar to the LR_{up} image. While this approach results in fewer hallucinations, it may also lead to the exclusion of finer details, which is the primary motivation for utilizing an SR algorithm. A trade-off between hallucinations and omissions highlights the importance of a balanced approach in remote sensing super-resolution algorithms.

Finally, a particular characteristic can be noted in the VENμS dataset, where all algorithms tend to show higher hallucination values (Table II). Visual analysis indicates that the VENμS imagery appears smoother when compared to the SR results of the three SR networks. This observation is supported by comparing the degradation kernel widths (Table I); specifically, the kernel width of VENμS is less than half that of NAIP and SPOT, i.e. the effective spatial resolution of VENμS images appears to be lower than 5 m. Since the correctness metrics are based on comparing distances, if the reference HR is inadequate, the results will always lean towards hallucination.

VI. CONCLUSIONS

This letter presents the *OpenSR-test*, a novel and comprehensive benchmark tailored for SR of optical remote sensing images. Designed around three foundational properties, the *OpenSR-test* provides a more precise interpretation of SR performance than traditional metrics, and provides three curated cross-sensor datasets covering various scales. The *OpenSR-test* framework can be seamlessly extended to tackle other image synthesis problems such as cloud removal, image dehazing, or SAR-optical fusion. We hope our work will encourage further research in the field of SR performance assessment.

REFERENCES

[1] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Science Reviews*, p. 104110, 2022.

[2] R. Cheng, H. Wang, and P. Luo, "Remote sensing image super-resolution using multi-scale convolutional sparse coding network," *Plos one*, vol. 17, no. 10, p. e0276648, 2022.

[3] 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 Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 155–170, 2022.

[4] H. Wu, N. Ni, and L. Zhang, "Learning dynamic scale awareness and global implicit functions for continuous-scale super-resolution of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

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

[6] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The WorldStrat dataset–with application to super-resolution," *arXiv preprint arXiv:2207.06418*, 2022.

[7] J. Michel, J. Vinasco-Salinas, J. Inglada, and O. Hagolle, "SEN2VENμS, a dataset for the training of Sentinel-2 super-resolution algorithms," *Data*, vol. 7, no. 7, p. 96, 2022.

[8] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[9] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[10] H. B. Li, C. Prabhakar *et al.*, "A domain-specific perceptual metric via contrastive self-supervised representation: Applications on natural and medical images," *arXiv preprint arXiv:2212.01577*, 2022.

[11] C. Aybar, L. Ysuhuaylas, J. Loja, K. Gonzales *et al.*, "CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Scientific Data*, vol. 9, no. 1, p. 782, 2022.

[12] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," *arXiv preprint arXiv:2306.13643*, 2023.

[13] M. Tyszkiewicz, P. Fua, and E. Trulls, "DISK: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 254–14 265, 2020.

[14] USDA-FSA-APFO Geospatial Enterprise Operations, "NAIP: National Agriculture Imagery Program," 2023.

[15] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric engineering and remote sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[16] W. G. C. Bandara and V. M. Patel, "Hypertransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1767–1777.

[17] S. Chen, R. Zhang, H. Su, J. Tian, and J. Xia, "SAR and multispectral image fusion using generalized IHS transform based on à trous wavelet and EMD decompositions," *IEEE Sensors Journal*, vol. 10, no. 3, pp. 737–745, 2010.

[18] R. Cresson, "SR4RS: A Tool for Super Resolution of Remote Sensing Images," *Journal of Open Research Software*, vol. 10, no. 1, 2022.