Enhancing Low-Cost Satellite Imagery through Self-Supervised Denoising Networks

Filippo Balzarini
Politecnico di Milano
filippo.balzarini@mail.polimi.it

Michele Cavicchioli
Politecnico di Milano
michele.cavicchioli@mail.polimi.it

Abstract

Land monitoring applications rely on high-quality, noise-free satellite imagery to ensure accurate analysis. Traditionally, this requires data from high-end satellites, which are costly to deploy and operate. In this work, we investigate whether imagery acquired from hypothetical low-cost satellites, characterized by spatially correlated noise, can still be made suitable for land monitoring tasks through the use of modern denoising techniques. We evaluate three methods of Blind-Spot Networks (BSN) and assess their ability to restore image quality. Our results demonstrate that while denoising effectively reduces noise and enables the use of degraded imagery, it also introduces some loss of fine details. The restored images remain suitable for tasks involving large-scale structures such as agricultural fields and major buildings, but challenges persist in scenarios requiring fine-grained details, such as dense urban environments.

Figure 1: Quality comparison between AP-BSN, NL-N2V and SSID

1. Introduction

High-resolution, noise-free satellite imagery is a cornerstone for a wide range of Earth observation and land monitoring applications, including agricultural assessment, urban planning, and environmental surveillance. The accuracy of these downstream tasks critically depends on the radiometric and spatial fidelity of the captured images. However, the acquisition of such high-quality data traditionally relies on sophisticated, high-end satellites equipped with advanced optical sensors, which are expensive to design, launch, and maintain. This economic constraint limits the spatial and temporal coverage achievable by spaceborne imaging systems, especially for small-scale research missions or developing regions.

Recent advances in computer vision and deep learning have opened new opportunities for improving the usability of imagery captured under suboptimal conditions. In particular, the emergence of self-supervised and blind-spot denoising networks has demonstrated that neural models can effectively suppress structured and unstructured noise without access to clean ground-truth data. Solutions such as *Noise2Void* (N2V) [5], *Non-Local Noise2Void* (NL_N2V) [10], *Asymmetric PD Blind-Spot Network* (AP-BSN) [6] and Spatially Adaptive Self-Supervised Image Denoising (SSID) [8] are based on Blind Spot Networks (BSNs). BSN is a denoising framework that masks some pixels of the image, i.e., *blind spots*, teaching the network to predict their noise-free value, exploiting the spatial redundancy inherent in natural images to learn noise statistics directly from corrupted data, bypassing the need for paired datasets that are unavailable in satellite imaging contexts.

Motivated by these developments, this work explores whether modern deep denoising architectures can compen-

sate for the degraded quality of images captured by hypothetical low-cost satellites. Such platforms, while significantly cheaper to build and operate, are subject to higher levels of spatially correlated noise due to limitations in sensor design, optical quality, and onboard processing. To replicate the conditions of low-cost satellite imaging, we developed a synthetic noise injection framework that models spatially correlated degradations observed in practical sensor systems, enabling controlled and reproducible evaluation of denoising performance. By evaluating the performance of several representative denoising models on synthetically degraded satellite imagery, we aim to quantify the trade-off between noise suppression and detail preservation.

Our findings indicate that deep denoising methods substantially improve the perceptual and quantitative quality of noisy satellite images, enabling their use in land monitoring tasks that depend primarily on large-scale structural features. However, residual artifacts and a modest loss of fine texture detail remain, which can hinder applications requiring precise delineation of small objects or dense urban features. These insights highlight both the promise and the limitations of learning-based denoising as an enabler for low-cost Earth observation, and motivate future research on architecture design and loss functions tailored to the spatial statistics of satellite imagery.

2. Problem Formulation

We define an image y as a function $I : X \to [0,1]^c$, where c is the number of the channels and $X = [1,\dots,H]\times[1,\dots,W]$ is a rectangular domain of height H and width W. We denote by $\mathcal{I}(X)$ the space of all images with domain X. Our observation model is the following:

$$\tilde{I}(x) = I(x) + \eta(x),\ \forall x \in X \tag{1}$$

where $\tilde{I}(x) \in \mathcal{I}(X)$ is the noisy observation, $I(x) \in \mathcal{I}(X)$ is the clean image and $\eta(x) \in \mathcal{I}(X)$ is the additive noise.

An image denoiser $\mathcal{D}_\theta$, of given parameters $\theta \in \Theta$ is a map $\mathcal{D}_\theta : \mathcal{I}(X) \to \mathcal{I}(X)$ that estimates a denoised image $\hat{I} \in \mathcal{I}(X)$ as $\hat{I} = \mathcal{D}_\theta(\tilde{I})$ of the clean image I.

The considered denoisers $\mathcal{D}_\theta$ are CNNs trained with self-supervised techniques, thus, assuming a training set:

$$T_{tr} = \{\tilde{I}_i\}_{i=1}^N$$

Of noisy images modeled as stated by Equation 1.

The training aims to identify the weights of the network θ such that:

$$\theta = \arg\min_{\theta\in\Theta} \sum_{k=1}^{K} \mathcal{L}(\mathcal{D}_\theta(\tilde{I}_k), \tilde{I}_k)$$

Where $\mathcal{L}$ is a self-supervised loss function computed exclusively on noisy images.

3. Related Work

3.1. Self-Supervised Denoisers

Traditional image denoising methods rely on supervised learning, requiring clean–noisy image pairs for training. However, in many practical scenarios—such as satellite imaging—obtaining perfectly aligned pairs is infeasible due to atmospheric disturbances, sensor noise, and temporal variations. This limitation has driven the development of self-supervised and unsupervised denoising approaches capable of learning directly from corrupted data.

The first work in this direction, Noise2Noise [7], demonstrated that pairs of independently noisy observations are sufficient to train a denoiser without clean targets. Subsequently, Noise2Void [5] eliminated the need for paired samples by predicting each pixel from its surroundings while masking its own value, effectively learning the noise distribution from single noisy images. Other works, such as Neighbor2Neighbor [4], further exploit image structure by constructing paired supervision from sub-images of neighboring pixels. This approach mitigates the impact of spatially correlated noise without relying on the blind-spot mechanism. An alternative masking strategy, adopted by models such as Self2Self, employs dropout-based masking, where pixels are randomly hidden during training. More recent studies have shown that transformer-based architectures achieve significantly higher performance than CNN-based architectures for blind-spot networks. Others, such as AT-BSN [2], demonstrated that integrating custom attention modules can also be highly effective.

Another strategy to overcome the absence of clean–noisy pairs is to construct a noise model that reproduces the characteristics of real noisy images. However, accurately modeling real-world noise is an extremely complex task, and in most practical cases, this approach is avoided.

3.2. Blind-Spot Networks

A Blind-Spot Network (BSN) is an architecture introduced by Noise2Void (N2V) [5] to perform image denoising in a self-supervised manner using only noisy images. The core idea is to mask pixels within the receptive field of the input image, forcing the model to infer the denoised value of a pixel based on its surroundings. This prevents the network from learning a trivial identity mapping and enables training without clean targets.

Noise2Void [5] introduced random masking of central pixels, while Noise2Self [1] generalized this principle as invariant prediction under data transformations. Building on this idea, APBSN [6] introduced an asymmetric pixel-shuffle downsampling (AP-) technique to generate multiple downsampled patches with reduced noise correlation between neighboring pixels. The blind-spot network (BSN) approach is then applied to these downsampled images.

Non-Local N2V (NL-N2V) [10] extended Noise2Void by introducing a non-local masking strategy, replacing masked regions with similar distant patches. This enables effective denoising under spatially correlated noise while maintaining the simplicity of the original BSN architecture. The Spatially Adaptive SSID model [8] further combines a blind-neighborhood network for flat regions with a locally aware network for textured areas, weighted by local flatness, to effectively remove real-world noise without clean–noisy pairs.

Subsequent research focused on leveraging attention mechanisms within state-of-the-art architectures. AT-BSN [2] extended CNN-based BSNs by integrating a self-masked attention mechanism, allowing the network to capture context beyond local neighborhoods while maintaining the blind-spot constraint. It also introduced tunable blind-spot sizes and multi-teacher distillation strategies to enhance robustness and reduce the detail loss commonly observed in CNN-based architectures.

Despite these advancements, CNN-based BSNs remain limited in reconstructing fine details, especially in images with complex or structured noise. Transformer-based architectures address this limitation by modeling long-range spatial dependencies and spectral correlations, yielding higher performance at the cost of increased computational requirements. TBSN [9] is a fully transformer-based BSN that employs self-masked attention to enforce blind spots, combining the benefits of global context modeling with self-supervised denoising.

Nevertheless, CNN-based approaches remain attractive for practical applications due to their ease of training, lower memory requirements, and robust performance across diverse noise types.

4. Proposed approach

The *self-supervised approach* was chosen to reflect a potential real-world scenario in which imagery is acquired exclusively from low-quality satellites, where clean reference data are unavailable. Unfortunately, the present problem domain lacks real datasets with the noise characteristics of interest. This limitation arises because the context of application corresponds to low-cost, reduced-quality satellite systems, which are not yet realized in practice and, therefore, do not provide empirical data.

To overcome this challenge, we introduce a controlled *noise injection module*. By designing this synthetic noise process, we approximate the statistical properties expected from realistic satellite acquisitions while maintaining flexibility in noise intensity and correlation structure. This approach enables the training and evaluation of denoising models under conditions that emulate the imaging environment, despite the absence of real-world noisy data.

4.1. Noise Injection Model

To simulate the noise characteristics, we apply a two-stage noise injection process to the clean image $I \in [0,1]^{3\times H\times W}$ (RGB channels normalized to the $[0,1]$ range). The model accounts for both *photon-limited*, modeled with a Poisson noise, and *sensor/electronic*, as a Gaussian noise.

Step 1: Poisson (Photon) Noise. Photon shot noise is modeled as a Poisson process. Given a scale factor $\alpha > 0$, we draw random counts from the Poisson distribution, denoted as $\mathcal{P}$, described in Equation 2

$$I^P(x,y,c) \sim \frac{1}{\alpha}\,\mathcal{P}(\alpha\, I(x,y,c)) \tag{2}$$

where (x,y) denotes spatial coordinates, $c \in \{R,G,B\}$ the color channel, and $I(x,y,c)$ the clean image intensity. This simulates the arrival of finite photons at the sensor.

Step 2: Gaussian (Electronic) Noise. For each spectral band c, we define a noise standard deviation proportional to the mean reflectance value of that band, weighted by the signal-to-noise ratio (SNR) of the band. Let Equation 3 be the average intensity in channel c.

$$\mu_c = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H} I(x,y,c) \tag{3}$$

Then the Gaussian noise variance is described by Equation 4 and derives from the definition of the mean SNR

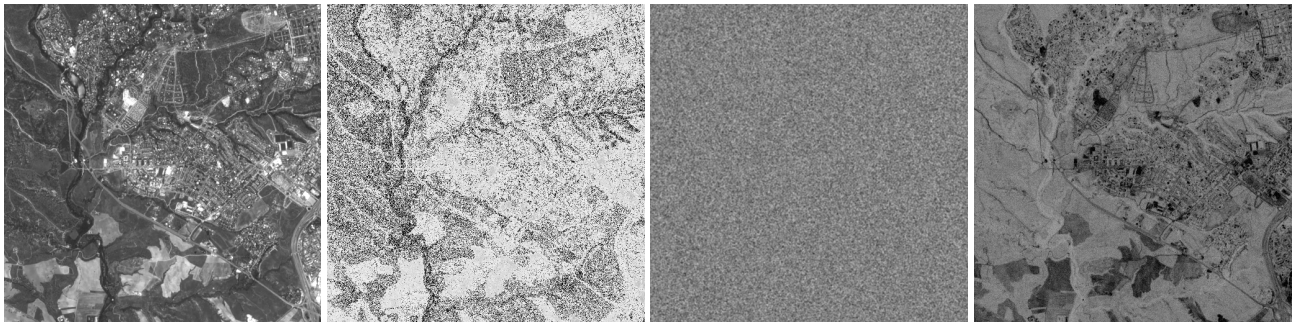
$$\sigma_c^2 = \left(\frac{\mu_c}{\mathrm{SNR}_c}\right)^2 \tag{4}$$

with $\mathrm{SNR}_c \in \{230, 249, 214\}$ for the R, G, B bands respectively as denoted by [3]. A global scaling factor β is introduced to emulate the effect of reduced sensor quality, as would be observed in lower-cost satellite systems; so the effective variance becomes $(\beta\sigma_c)^2$. The Gaussian noise distribution is stated in Equation 5

$$N_c(x,y) \sim \mathcal{N}(0, (\beta\sigma_c)^2). \tag{5}$$

To enforce correlation, the Gaussian noise field is spatially filtered with a Gaussian kernel $K_{\sigma,k}$ of edge size k and variance σ as shown in Equation 6, where $\circledast$ denotes the convolution operation.

$$\hat{N}_c = N_c \circledast K_{\sigma,k}. \tag{6}$$



(a) Clean (b) Poisson component (c) Gaussian component (d) Noisy image

Figure 2: Noise injection procedure

Step 3: Noise Combination. The final noisy image is obtained as the additive combination shown in Equation 7

$$\tilde{I}(x, y, c) = I^P(x, y, c) + \hat{N}_c(x, y) \tag{7}$$

The above-described formulation captures both the signal-dependent photon statistics and the band-dependent electronic noise properties of Sentinel-2 imagery, while allowing tunable control over the noise correlation and intensity. Figure 2 shows examples of the different components building the final noisy image starting from the clean sample.

5. Experiments

5.1. Dataset and Metrics

To construct the dataset for our denoising experiments, we gathered high-resolution satellite images on the bands B04 (red), B03 (green), and B02 (blue) from the Planetary Computer API provided by Microsoft [11]. Table 1 summarizes the exact data that were collected.

The raw images were first inspected, cropped into fixed-size patches of 512×512 pixels, and cleaned. To increase variability and ensure coverage of different regions within each image, we applied an offset-based cropping strategy, shifting the crop window by 128 pixels across the horizontal and vertical axes. This procedure yielded a diverse set of partially overlapping patches while preserving local spatial context. The final dataset contained a total of 3979 images, which we successively split into 80% for training, 10% for validation, and 10% for testing.

Location	Latitude [°]	Longitude [°]	Date
Rome	41.9 - 42.0	12.45 - 12.5	June 2024
Paris	48.9 - 49.0	2.4 - 2.45	June 2023
Madrid	40.4 - 40.5	3.7 - 3.75	April 2023
Tokyo	35.7 - 35.8	139.7 - 139.75	June 2023
London	51.5 - 51.6	-0.2 - -0.25	June 2023
Milan	45.5 - 46.5	9.2 - 9.25	June 2023

Table 1: Geographical bounding boxes and acquisition dates of the raw satellite images.

During the training process, we provided our network with only noisy images and minimize the self-supervised loss of the model. The noisy images are generated following Section 4.1, the parameters are shown in Table 2.

We evaluated the performance of the studied models by Structural Similarity Index Measure (SSIM) [12] and Peak Signal to Noise Ratio (PSNR), those are briefly described in Section 5.1.1 and Section 5.1.2.

5.1.1 Structural Similarity

To better align objective image quality metrics with human perception, [12] introduced the Structural Similarity Index (SSIM). The method is based on the observation that human vision is highly adapted to extract structural information from natural scenes, and that perceived distortions are strongly linked to changes in this structural content. SSIM evaluates the similarity between two image patches using three complementary components: *luminance*, *contrast*, and *structure*.

5.1.2 Peak Signal-to-Noise Ratio

The Peak Signal-to-Noise Ratio (PSNR) is one of the most widely used fidelity measures in image processing. It is derived from the Mean Squared Error (MSE) between a reference image and its distorted version, expressing the ratio between the maximum possible signal power and the power of the distortion noise. Due to its simplicity and clear physical interpretation, PSNR has long served as a standard baseline for evaluating image quality, although it often correlates poorly with human visual perception.

5.2. Evaluated Models

In this study, three models were evaluated: AP-BSN [6], NL-N2V [10], and Spatially Adaptive Self Supervised Image Denoising [8]. All models were trained, validated, and tested using the exact same dataset split to ensure a consistent experimental setup and enable a fair comparison of their performance.

5.3. Results

This section presents the results obtained across the trained models.

Table 3 reports the mean values for both the SSIM and PSNR metrics. The NL-N2V model was evaluated using multiple patch sizes to investigate how spatial context influences performance. We observed that increasing the patch size generally improves SSIM, indicating better structural preservation, but tends to slightly decrease PSNR.

Overall, the results indicate that SpatiallyAdaptiveSSID consistently outperforms the other models in terms of structural similarity, which, in this context, represents the most relevant evaluation metric. Nevertheless, SpatiallyAdaptiveSSID is also significantly more computationally demanding. When trained on a GTX 1080 Ti equipped with 12 GB of RAM, it requires approximately seven days of computation, compared to 8–10 hours for the other models. This substantial increase in training time can be attributed to the three-stage training procedure adopted by SpatiallyAdaptiveSSID.

Among the evaluated models, APBSN achieved the highest PSNR, even though qualitative assessments indicate that it loses the most fine details. The additional smoothing observed in APBSN is mainly due to its architectural design, which involves downsampling operations that inherently discard some spatial information. The model was trained by minimizing an L1 loss, as previous experiments showed that using an L2 loss led to noticeably stronger blurring in the denoised images. This behavior can be explained by the fact that, under an L2 objective, large errors contribute quadratically more to the total loss compared to L1. Consequently, the network learns to avoid large deviations, producing smoother and less detailed results in general. The high PSNR achieved by APBSN can thus be explained by its lower average pixel-wise error: since PSNR is directly derived from the mean squared error (MSE), a model that generates smoother predictions with smaller deviations, like APBSN, achieves a higher PSNR, even if it sacrifices fine detail. In contrast, other models that attempt to preserve more texture may incur larger local errors, resulting in slightly lower PSNR values.

Figure 4 illustrates the PSNR values across the entire test set, while Figure 3 presents the corresponding SSIM values. These figures confirm that even when evaluated on individual samples, SpatiallyAdaptiveSSID is the model that better fits the challenge.

Parameter	Value
photon_scale	1000
noise_boost	10.0
kernel_edge	5
kernel_sigma	0.55

Table 2: Noise injection Parameters

Model	SSIM ↑	PSNR ↑	Patch Size
APBSN	0.795	**30.71**	480x480
NLN2V	0.787	28.56	240x240
NLN2V	0.789	28.45	512x512
SSID	**0.799**	29.92	512x512

Table 3: Results for Structural Similarity and Peak Signal-to-Noise Ratio for the trained models

Figure 5 shows how the SSIM changes with the PSNR value, which states the robustness and consistency of the results of SpatiallyAdaptiveSSID compared to the other models analyzed.

5.3.1 Qualitative results

In Figure 6 in the appendix, we provide four examples of denoised images alongside their noisy counterparts, to illustrate where each model performs best and worst. It is quite clear that none of the tested models can produce images with a satisfactory level of detail — all of them suffer from significant information loss. Moreover, the results get worse when there is not a sufficient contrast in the original image. For example, in the second comparison image, the lack of contrast between structures makes the result appear as a single homogeneous component rather than distinct elements.

6. Conclusion

In this report, we evaluated several denoising models on real noisy optical satellite images. The results indicate that although all tested models effectively reduce noise, none can preserve fine structural details, resulting in a significant loss of information. Blind-spot networks demonstrated particular limitations in this manner, as their intrinsic masking mechanism inherently suppresses small-scale textures and fine features, which are critical for accurate image reconstruction.

These findings suggest that blind-spot architectures are better suited for tasks emphasizing large-scale or high-contrast structures, but are less appropriate for domains characterized by low contrast and fine-grained details, such as optical satellite imagery. Future research should focus

on developing denoising strategies capable of preserving detailed information while maintaining robustness to real-world noise. Promising directions include hybrid architectures and training approaches that integrate blind-spot learning with explicit detail-preservation mechanisms.

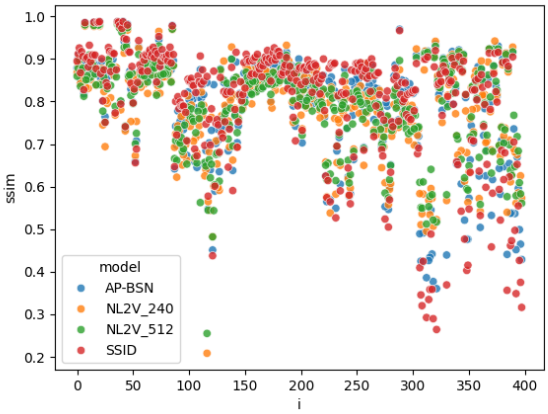


Figure 3: SSIM metric results per single image

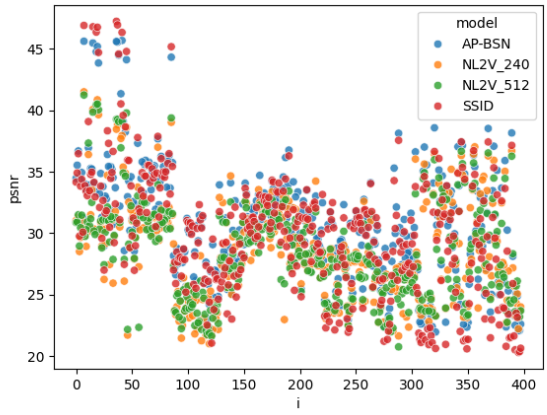


Figure 4: PSNR metric results per single image

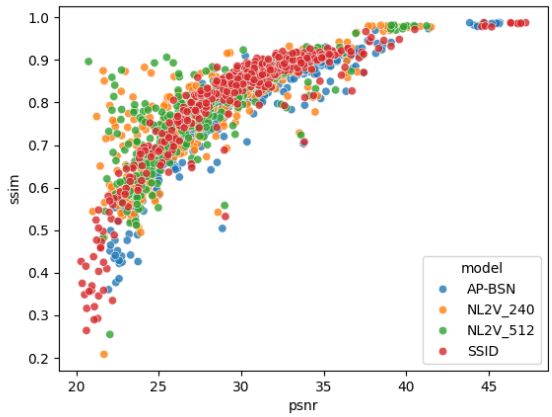


Figure 5: SSIM (y-axis) and PSNR (x-axis) for different models.

References

[1] J. Batson and L. Royer. Noise2Self: Blind denoising by self-supervision. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 524–533. PMLR, 09–15 Jun 2019. 2

[2] S. Chen, J. Zhang, Z. Yu, and T. Huang. Exploring efficient asymmetric blind-spots for self-supervised denoising in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2814–2823, June 2024. 2, 3

[3] F. Gascon, C. Bouzinac, O. Thépaut, M. Jung, B. Francesconi, J. Louis, V. Lonjou, B. Lafrance, S. Massera, A. Gaudel-Vacaresse, F. Languille, B. Alhammoud, F. Viallefont, B. Pflug, J. Bieniarz, S. Clerc, L. Pessiot, T. Trémas, E. Cadau, R. De Bonis, C. Isola, P. Martimort, and V. Fernandez. Copernicus sentinel-2a calibration and products validation status. *Remote Sensing*, 9(6), 2017. 3

[4] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14781–14790, June 2021. 2

[5] A. Krull, T. Buchholz, and F. Jug. Noise2void - learning denoising from single noisy images. *CoRR*, abs/1811.10980, 2018. 1, 2

[6] W. Lee, S. Son, and K. M. Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5

[7] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2noise: Learning image restoration without clean data. In *International Conference*

on Machine Learning (ICML), volume 80, pages 2971–2980, March 2018. 2

[8] J. Li, Z. Zhang, X. Liu, C. Feng, X. Wang, L. Lei, and W. Zuo. Spatially adaptive self-supervised learning for real-world image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9914–9924, June 2023. 1, 3, 5

[9] J. Li, Z. Zhang, and W. Zuo. Rethinking transformer-based blind-spot network for self-supervised image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 3

[10] D. Martin, E. Peretti, and G. Boracchi. Non-local n2v: Improving n2v networks for spatially correlated noise. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 2175–2180, 2025. 1, 3, 5

[11] M. O. Source, M. McFarland, R. Emanuele, D. Morris, and T. Augspurger. microsoft/planetarycomputer: October 2022, Oct. 2022. 4

[12] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

A. Supplementary Material

A.1. Qualitative results

Figure 6 presents four examples comparing denoised images with their corresponding noisy inputs. These visual comparisons highlight the strengths and weaknesses of each model discussed in Section 5.3.1, illustrating the loss of fine details and contrast that limits their performance on real noisy optical satellite images.

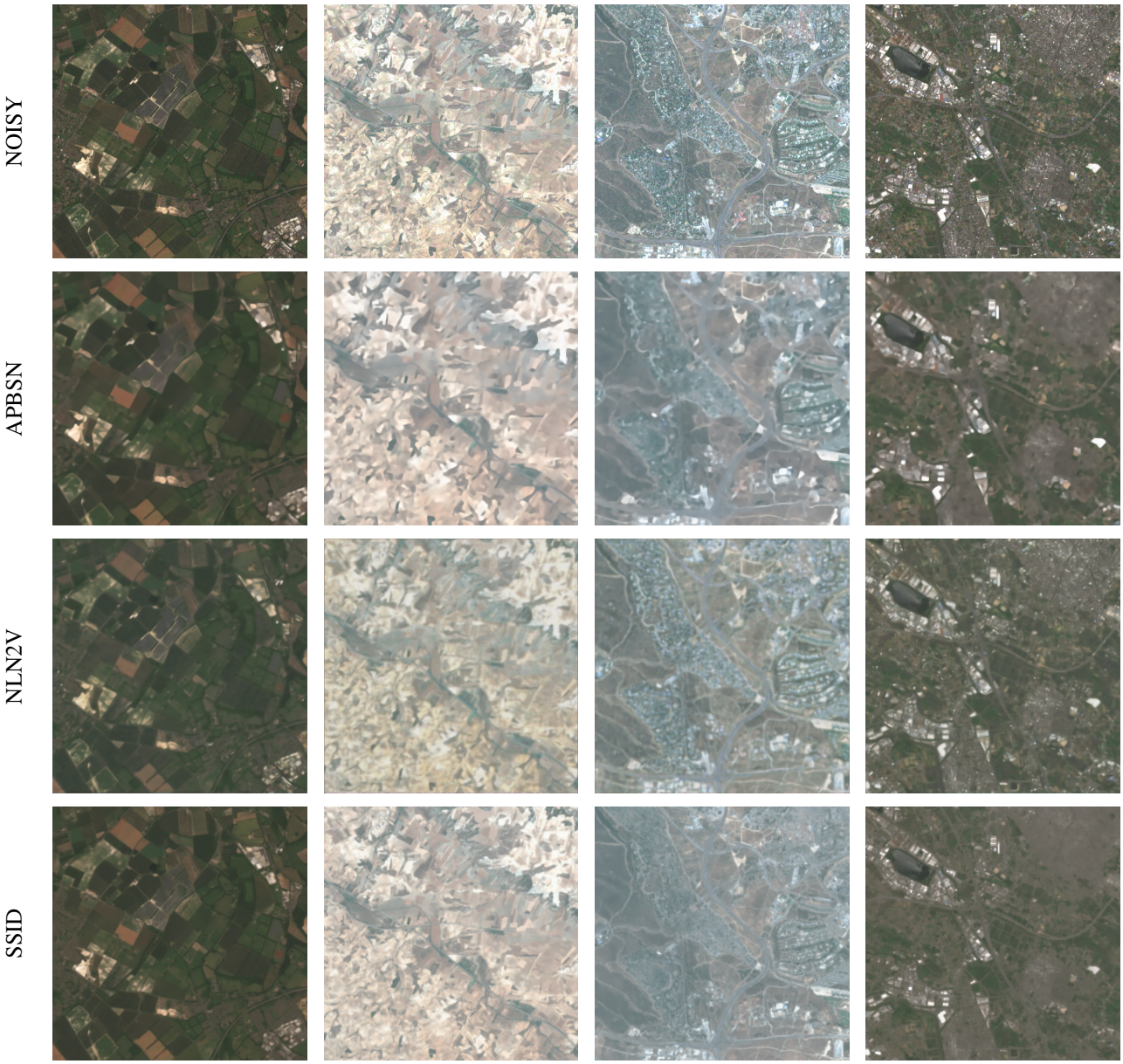


Figure 6: Results comparion between models