

Noise-Aware Data Augmentation for Robust Road Detection in Small Satellite Imagery

Nina Krašovec^{1,2}, Aleš Marsetič^{1,3}

¹ ZRC SAZU, Novi trg 2, 1000 Ljubljana, Slovenia - (nina.krasovec, ales.marsetic)@zrc-sazu.si

² Faculty of Civil and Geodetic Engineering, University of Ljubljana, 1000 Ljubljana, Slovenia

³ SPACE-SI, Aškerčeva 12, 1000 Ljubljana, Slovenia

Keywords: Road Detection, Small Satellite, Data Augmentation, Noise, Blur, Transfer Learning.

Abstract

Road extraction from small satellite imagery is challenging because raw images often suffer from low signal-to-noise ratio (SNR), high radiometric variability, and reduced sharpness. In this work, we investigate whether noise and blur data augmentation during pretraining can improve robustness in such conditions. We use a two-stage transfer-learning framework in which a U-Net with a ResNet-50 encoder is pretrained on PlanetScope RGB imagery and fine-tuned on NEMO-HD imagery. During pretraining, we evaluate Gaussian, ISO-like, and Perlin noise, as well as Gaussian and motion blur, each at three severity levels. On the internal held-out test split, augmentation effects were modest, with the best strict IoU improving from 26.7% for the geometric-only baseline to 27.2%. However, evaluation on external full-scene NEMO-HD images showed clearer benefits. Augmentation-based models consistently improved road detection in raw imagery, mainly by increasing completeness and recall, while there was little or no systematic benefit in stacked imagery. No clear trend was observed across augmentation severity levels, indicating that performance depended more on scene conditions than on perturbation strength. The results show that augmentation is most useful for single-acquisition small satellite imagery, where it improves robustness to lower image quality.

1. Introduction

Satellite imagery from small satellites often exhibits increased noise levels, higher radiometric variability, and lower spatial resolution compared to large missions. These factors impact image quality and make downstream tasks such as object detection and semantic segmentation more challenging. Although such imagery is valuable due to its affordability and sometimes high temporal frequency, extracting roads reliably remains difficult, particularly under lower SNR conditions, occlusions, and varying illumination.

The small satellite dataset used in this work reflects many of the challenges discussed above. NEMO-HD, the first Slovenian microsatellite for Earth observation, usually supports image stacking, which improves SNR and overall image quality. However, operational constraints can limit the number of available acquisitions, particularly when rapid maneuvers are needed to follow specific targets such as rivers. In such cases, only single acquisitions are available. Compared with stacked products, these raw images are more strongly affected by low SNR and radiometric variability, making road extraction particularly challenging.

Motivated by this, we study whether noise-aware augmentation during pretraining can improve robustness to some of the typical challenges of small satellite imagery. Models are pretrained on PlanetScope imagery and fine-tuned on NEMO-HD data, while controlled noise and blur augmentations are used during pretraining to better reflect target-domain conditions. By comparing results on raw single-acquisition images and stacked images, we assess augmentation effects under different image-quality regimes. While the experiments focus on NEMO-HD, the underlying idea is applicable more broadly to small satellite systems affected by low SNR, radiometric variability, and reduced sharpness.

1.1 Related Work

Recent work on road extraction from remote sensing imagery has been driven primarily by advances in deep learning, including improved backbone architectures, multiscale feature fusion, and topology-aware representations of road networks (Chen et al., 2022, Zhang et al., 2024, Lu and Weng, 2025). However, the literature also shows that performance is constrained not only by model design but by the quality and consistency of the source imagery. This issue is especially relevant for small satellite and CubeSat platforms, which are attractive because of their comparatively low cost and high revisit capability, but often require additional preprocessing to achieve geometric and radiometric consistency for quantitative analysis (Leach et al., 2019, Mhangara et al., 2020, Dias et al., 2024). For example, comparative analysis of nSight-1 against Landsat-8 reported lower visual quality and interpretability for the nanosatellite imagery (Mhangara et al., 2020), while studies on PlanetScope and other optical CubeSat products have emphasized persistent radiometric inconsistencies and the need for explicit normalization or harmonization before downstream use (Leach et al., 2019, Dias et al., 2024). These findings are directly relevant for road extraction, where narrow linear structures are particularly sensitive to blur, low SNR, and radiometric instability.

A related line of work addresses robustness to acquisition-related image-quality challenges during training. In remote sensing, reduced SNR and image-quality variations are known to affect mapping accuracy (Xian et al., 2019). This has motivated approaches that simulate acquisition conditions or inject distortions during training in order to improve generalization under realistic operating conditions. Nalepa et al. (2021), for example, studied robustness in hyperspectral image segmentation by simulating atmospheric effects and injecting Gaussian, impulsive, and Poisson noise (Nalepa et al., 2021). Within road extraction itself, augmentation pipelines have also been used to improve generalization. JointNet employed random-noise aug-

mentation including Gaussian, salt-and-pepper, Poisson, and speckle noise (Zhang and Wang, 2019), while RDPGNet used random rotation, flipping, Gaussian blur, masking, and contrast adjustment during training (Liu et al., 2024). However, these studies typically report such augmentations as parts of the overall training pipeline rather than isolating their individual contribution to performance. In addition, the above works were not conducted on small satellite road extraction: Nalepa et al. (2021) considered hyperspectral image segmentation, whereas JointNet and RDPGNet were evaluated on standard road-extraction benchmarks rather than low SNR small satellite imagery (Nalepa et al., 2021, Zhang and Wang, 2019, Liu et al., 2024). Although Perlin/simplex noise has also been proposed as a more structured augmentation mechanism, the available evidence comes from medical imaging rather than remote sensing road extraction (Bae et al., 2018).

2. Satellite imagery and reference data

Slovenia's first microsatellite, NEMO-HD, was launched in 2020 and is operated by the Slovenian Centre of Excellence for Space Sciences and Technologies SPACE-SI. The satellite provides four spectral channels (R, G, B, NIR) at a spatial resolution of 5.6 m and a panchromatic channel at a spatial resolution of 2.8 m. It is aimed at monitoring smart cities, rivers and the sea, agriculture, droughts, floods, and invasive plants (Rodič et al., 2022).

NEMO-HD uses a full-frame sensor. The bands in single-shot images are radiometrically corrected and co-registered. If multiple images are taken consecutively, images are stacked and the bands are co-registered, which improves SNR (Figure 1). SNR is defined as the ratio of the mean value of the signal and the mean deviation of the pixel values. A stack depth of 12 on the multispectral channels gives SNR above 70 while a single image gives a SNR around 20. All NEMO-HD images in this study were geometrically corrected using the processing chain called STORM (Pehani et al., 2016) and projected to UTM.

NEMO-HD raw imagery exhibits some quality-related effects that are relevant for road extraction. These include reduced sharpness caused by motion and sometimes defocus, visible vignetting in two bands of the raw images, radiometric variability, and occasional poor band-to-band co-registration. In addition, single acquisitions have substantially lower SNR than stacked imagery. These effects vary across scenes and can reduce the visibility and delineation of narrow road structures, making both training and evaluation more challenging.

PlanetScope imagery at 3 m spatial resolution was used for pre-training due to its worldwide availability, improved radiometric stability, and spatial resolution relatively close to that of NEMO-HD. Because PlanetScope has higher spatial resolution, road structures appear clearer and are easier for the network to learn from. In addition, the amount of available NEMO-HD training data is limited, making PlanetScope a suitable source for learning robust initial representations. Pretraining on this dataset therefore provides robust initial weights and improves subsequent fine-tuning on NEMO-HD images.

Reference road data were derived from OpenStreetMap (OSM) and manually refined based on their visibility in the PlanetScope and NEMO-HD imagery. Roads missing from OSM were not added manually. For each road type, widths were measured at several locations to define the rasterization width. The resulting vector road data were rasterized using the corresponding road widths, matched to the spatial resolution of the imagery, and projected to the UTM coordinate system. Both the satellite images and the reference masks were stored in GeoTIFF format.

3. Experimental setup

We adopt a two-stage training procedure for road detection in small satellite imagery. First, models are pretrained on PlanetScope RGB imagery using 50-20-30 data split for training, validation and testing, respectively. All models trained on PlanetScope imagery are then fine-tuned on NEMO-HD RGB imagery using 50-20-30 data split. In total, 14 PlanetScope images

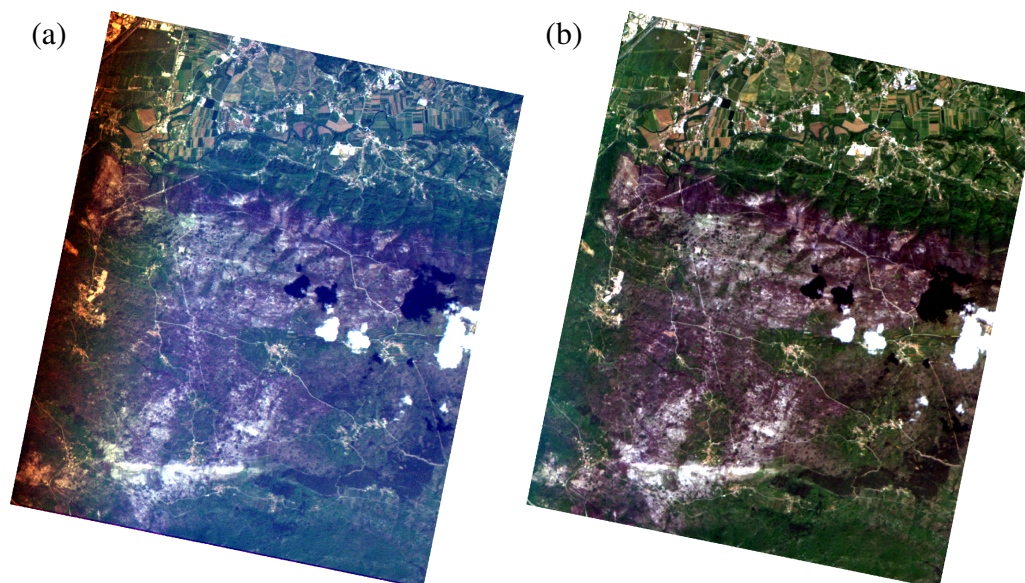


Figure 1. Representative NEMO-HD imagery over the Kras study area: (a) raw single-acquisition image and (b) the corresponding stacked product. The stacked image shows improved radiometric consistency, resulting in a more homogeneous appearance and better visual separability of landscape features.

were divided into tiles of size 256×256 , yielding 1501, 613, and 913 tiles for training, validation, and testing, respectively. Training tiles containing no road pixels were removed before augmentation and model training. In addition, 13 NEMO-HD images were divided into tiles of the same size, resulting in 361, 158, and 231 tiles for the training, validation, and test sets, respectively.

We use U-Net encoder-decoder architecture (Ronneberger et al., 2015) with ResNet-50 encoder initialized from ImageNet weights. Training uses BCE-Dice loss which is better suited to imbalanced datasets than BCE loss alone. The Adam optimizer is used with a learning rate of $5e-4$ and weight decay of $1e-3$. An adaptive learning-rate scheduler reduces the learning rate when validation performance stagnates. We train each model for up to 50 epochs with batch size of 16 and select the best checkpoint by intersection over union (IoU). The final decision threshold is fixed at 0.4 for binarization.

3.1 Augmentation Design

The augmentation study is organized into three experiment groups. *Baseline* establishes reference performance under standard training conditions. *Noise* evaluates robustness to intensity perturbations and *blur* assesses sensitivity to image image-quality changes caused by loss of sharpness and motion effects. All transforms within noise and blur groups are tested across low (*L*), medium (*M*), and high (*H*) severity levels (Table 1).

The augmentation strategy was designed to reflect the main image-quality variations observed in the dataset, particularly radiometric variability and loss of sharpness. To improve robustness and generalization during pretraining, we employ three groups of data augmentations that capture common image-quality variations present in small satellite optical imagery. The first group consists of geometric transformations, specifically flips and rotations, which increase spatial invariance and effectively enlarge the training set. The second group introduces radiometric perturbations through Gaussian noise, synthetic ISO-like noise, and Perlin noise. These transformations model different forms of intensity variation, ranging from spatially uncorrelated fluctuations to spatially coherent low frequency patterns. The third group comprises Gaussian blur and motion blur, which simulate reduced sharpness and motion-related image distortions. Together, these augmentations enable us to evaluate whether robustness to radiometric and sharpness variability learned during pretraining improves downstream performance. In all pretraining configurations, geometric augmentations are applied by default, with noise- or blur-based transformations introduced additionally according to the respective experimental

setting. During fine-tuning, we restrict augmentation to geometric transformations only, in order to preserve realistic image appearance and avoid introducing artificial distortions into the target dataset.

Gaussian noise is a stochastic perturbation in which each pixel is independently modified by a value drawn from a normal distribution. In the context of satellite imagery, it provides a simple model for random sensor-related fluctuations and other spatially uncorrelated disturbances. Synthetic ISO-like noise is used to mimic grain-like sensor artifacts by introducing random intensity fluctuations with a noisy appearance. Perlin noise, in contrast, produces smooth and spatially correlated random patterns in which neighboring values vary gradually rather than independently. This makes it useful for simulating low-frequency intensity variations and texture-like disturbances. Following prior work in the medical imaging domain (Bae et al., 2018), we use the computationally efficient simplex variant implemented in the Python package *noise* (v. 1.2.2); for brevity, we refer to this variant as *Perlin noise* throughout the paper.

Perlin noise generates a pseudo-random surface over 2D coordinates, so nearby pixels receive similar perturbations rather than independent random values. In our implementation, this field is sampled as $n(u, v)$ and we use *scale* and *intensity* as two additional parameters. *Scale* controls spatial granularity where larger values give broader, lower-frequency structures and smaller values give finer texture, while *intensity* controls perturbation magnitude and is applied after noise generation. The augmented image is computed as $I_{\text{aug}} = \text{clip}(I + \alpha n, 0, 1)$, with $\alpha = \text{intensity}$. For all the remaining augmentations, we used Python library Albumentations (v. 2.0.8) (Buslaev et al., 2020) with their standard implementations.

Gaussian blur is a smoothing operation that suppresses high-frequency image content by convolving the image with a Gaussian kernel. It is used to simulate reduced sharpness caused by limited optical resolution or slight defocus. Motion blur represents directional spatial distortions that smears intensities along a dominant axis and therefore reduces local detail. In remote sensing imagery, it serves as a simple approximation of motion-induced artifacts and acquisition-related loss of sharpness.

We first train a baseline model (A01) with geometric transform alone. The N-series (N01 - N09) evaluates robustness to noise perturbations at three severities: Gaussian noise (N01-N03), Perlin noise (N04-N06), and ISO-like noise (N07-N09). The B-series (B01-B06) evaluates blur robustness for Gaussian blur (B01-B03) and motion blur (B04-B06). This design enables us to compare different noise and blur augmentations as well as the impact of the severity levels.

Exp. IDs	Augmentation Group	Low (L)	Medium (M)	High (H)
N01-N03	Gaussian noise	noise std. range [0.01, 0.03]	noise std. range [0.03, 0.07]	noise std. range [0.07, 0.12]
N04-N06	Perlin noise	scale 140.0 intensity 0.05	scale 100.0 intensity 0.10	scale 70.0 intensity 0.20
N07-N09	ISO noise	color shift [0.005, 0.02] intensity [0.05, 0.20]	color shift [0.01, 0.05] intensity [0.10, 0.40]	color shift [0.03, 0.08] intensity [0.30, 0.70]
B01-B03	Gaussian blur	blur limit [3,3] sigma limit [0.6,0.8]	blur limit [7,7] sigma limit [1.0,1.4]	blur limit [11,11] sigma limit [1.8,2.4]
B04-B06	Motion blur	blur limit [3,3] direction range [-0.2,0.2]	blur limit [7,7] direction range [-0.5,0.5]	blur limit [11,11] direction range [-0.8,0.8]

Table 1. Overview of augmentation settings: experiment IDs mapped to noise and blur augmentations, with low/medium/high severity parameter ranges.

3.2 Evaluation

We report strict F1-score and IoU along with 2-pixel tolerance metrics to account for minor spatial offsets and differences in road width between the reference and the predictions. Results are evaluated on additional full NEMO-HD images rather than test tiles from the 13-image set, where conditions differ substantially. This enables a more controlled comparison and improves interpretability. We use raw satellite imagery from Kras and Lenart (Slovenia), Toronto (Canada), and Madrid (Spain). Stacked imagery is available for Kras, Toronto, and Madrid, while only single image is available for Lenart. The RGB images have a spatial resolution of 5.6 meters. Raw images exhibit some vignetting along with higher radiometric variability and lower SNR, whereas this effect is removed in the stacked images.

The Kras image covers approximately 118 km². Its northern part is flat and rural with several villages, while the remaining area is mostly forested and hilly. Roads primarily connect villages or pass through forests, so vegetation occlusion is present. The Toronto image covers approximately 118 km² of the city outskirts. The terrain is predominantly flat, with open and wide roads that are clearly visible from a top-down view. Most roads are concentrated in the west, where the raw image also shows some vignetting. The Madrid image covers approximately 130 km² outside the urban area. It is mostly rural with villages. The image is slightly defocused, resulting in reduced detail due to smearing. The Lenart image covers approximately 139 km² of agricultural and forested, hilly terrain with small villages. Roads are partially occluded by vegetation.

4. Results and discussion

On the internal held-out test data, performance differences across augmentation settings were modest but revealed several consistent trends (Table 2). The best result was obtained by B01 (low-severity Gaussian blur), with IoU of 27.2% and F1-score of 42.8%, compared with 26.7% IoU and 42.1% F1-score for the geometric-only baseline. Thus, the best augmented configuration improved performance only slightly, by about 0.5% IoU and 0.7% F1-score.

Exp. ID	F1 (%)	IoU (%)	F1 _{2px} (%)	IoU _{2px} (%)
A01	42.1	26.7	73.2	57.7
N01 _L	41.1	25.9	71.6	55.7
N02 _M	41.4	26.1	72.4	56.7
N03 _H	40.8	25.7	71.7	55.9
N04 _L	42.2	26.7	72.6	57.0
N05 _M	42.0	26.6	72.8	57.2
N06 _H	42.7	27.1	72.9	57.3
N07 _L	42.3	26.8	73.2	57.7
N08 _M	41.7	26.3	72.5	56.9
N09 _H	42.6	27.0	72.9	57.4
B01 _L	42.8	27.2	73.3	57.8
B02 _M	41.0	25.8	72.3	56.7
B03 _H	42.3	26.9	73.3	57.9
B04 _L	41.1	25.9	72.5	56.8
B05 _M	42.5	27.0	73.2	57.8
B06 _H	41.8	26.4	73.0	57.4

Table 2. Summary of internal held-out test results for the geometric-only baseline (A01), noise-based (N01–N09), and blur-based (B01–B06) augmentation settings. Metrics are reported as strict F1 and IoU and the corresponding 2-pixel tolerance.

Augmentation did not improve performance uniformly. Only 7 of the 15 augmented variants outperformed the baseline IoU, showing that augmentation improved performance only for some augmentation types and severity levels, rather than across all tested settings. On average, blur- and noise-based augmentation performed almost identically (26.52% and 26.47% mean IoU, respectively). Gaussian noise consistently underperformed the baseline, whereas Perlin and ISO-like noise produced the strongest results.

The reference road masks have several limitations that should be considered when interpreting the results. The rasterized road widths do not always match the apparent road widths in the imagery, some roads are missing, and some labels exhibit small spatial offsets. These inaccuracies arise partly from the incompleteness and positional uncertainty of OSM data and partly from the difficulty of orthorectifying lower-quality small satellite imagery. For this reason, in addition to strict metrics, we report 2-pixel tolerance metrics, which reduce the impact of minor spatial misalignments and small discrepancies in road width. This threshold was selected based on a sensitivity analysis showing only small additional gains for larger tolerances.

Metrics with 2-pixel tolerance yielded substantially higher absolute scores for all models, reflecting the sensitivity of strict evaluation to small spatial offsets and width mismatches. For example, the baseline IoU increased from 26.7% under strict evaluation to 57.7% under the tolerance evaluation. However, these metrics did not change the main conclusion that the geometric-only baseline remained highly competitive.

To assess generalization beyond the internal data split, all trained models were additionally evaluated on seven full NEMO-HD images: Kras (raw, stack), Toronto (raw, stack), Madrid (raw, stack), and Lenart (raw). These images were not used during model development. For Kras and Toronto, the evaluation was carried out on additional full-scene areas from the same acquisition stripes as the data used in the internal split, but outside the training, validation, and test regions. In contrast, Madrid and Lenart represent fully unseen scenes that were not used at any stage of training or model selection. Table 3 summarizes the best-performing model for each image under both strict and 2-pixel tolerance evaluation, together with the corresponding road density in the reference data.

Image	Rd. dens. (km/km ²)	Strict		2-pixel tolerance	
		Best	IoU (%)	Best	IoU (%)
Kras raw	1.255	N08 _M	16.3	N08 _M	49.0
Kras stack	1.255	A01	22.0	N03 _H	59.3
Toronto raw	2.248	B04 _L	27.0	N08 _M	73.0
Toronto stack	2.248	N09 _H	33.1	N04 _L	80.4
Madrid raw	0.883	N03 _H	12.4	N03 _H	24.5
Madrid stack	0.883	A01	16.4	A01	37.4
Lenart raw	1.592	B04 _L	15.3	B04 _L	36.2

Table 3. Best-performing models on the external full-scene set under strict and 2-pixel tolerance evaluation. Road density is given in km/km², and A01 denotes the geometric-only baseline.

The full-scene evaluation shows a strong scene dependence (Figure 2). Under strict evaluation, the best IoU ranges from 12.4% to 33.1%, while under 2-pixel tolerance evaluation it ranges from 24.5% to 80.4% for Madrid (raw) and for Toronto (stack), respectively. Toronto yields the strongest results in both raw and stacked imagery and also has the highest road density (2.248 km/km²), whereas Madrid is the most challenging scene and has the lowest road density (0.883 km/km²). Part of the

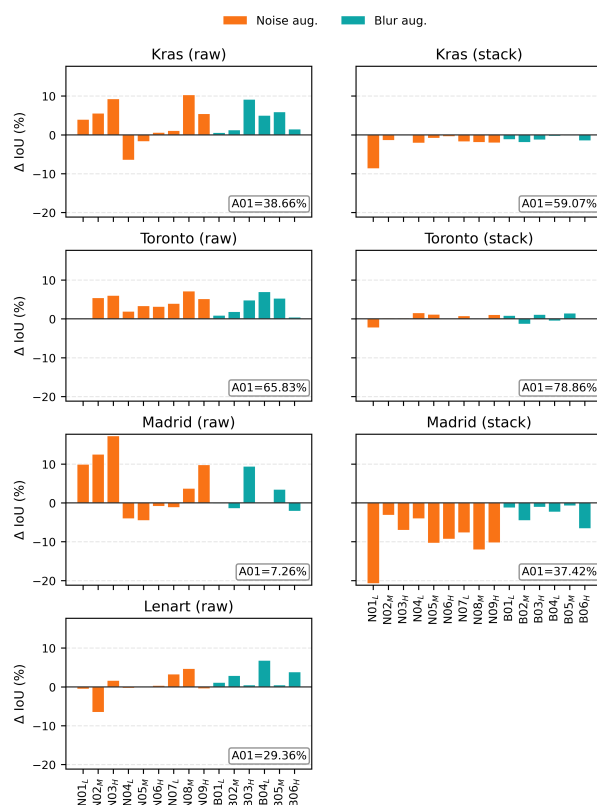


Figure 2. Relative change in 2-pixel tolerance IoU compared with the baseline (A01).

stronger performance on Kras and Toronto may be attributed to the fact that these evaluation scenes originate from the same acquisition stripes as the internal split, which could reduce the domain shift relative to Madrid and Lenart. Nevertheless, road density and partial acquisition overlap do not explain all differences. For example, Lenart has a higher road density than Kras, yet its best raw-image performance is lower under both strict and tolerance evaluation (strict IoU 15.3% vs. 16.3%; 2-pixel IoU 36.2% vs. 49.0%), indicating that scene characteristics and image quality also play an important role.

Using the best model for each version of the scene (raw and stack), strict IoU improves by 5.7% for Kras, by 6.1% for Toronto, and by 4.0% for Madrid. The corresponding 2-pixel tolerance IoU improves by 10.3%, by 7.4%, and by 12.9%, respectively. These gains are consistent with improved quality of stacked imagery.

On all raw scenes, the best augmentation-based model improves recall relative to the geometric-only baseline, indicating that more road pixels are detected and fewer are missed (Figure 3). This trend is consistent with the more challenging nature of raw imagery, where lower SNR and higher radiometric variability leave greater room for improvement through augmentation. The gain in recall, however, comes at the cost of reduced precision, showing that the additional true positives are accompanied by additional false positives. This trade-off is relatively mild in Kras and Toronto, where recall improves substantially while precision decreases only modestly. In Madrid and Lenart, the trade-off is less favorable: recall still increases, but precision drops more noticeably. This is consistent with the higher difficulty of these scenes, as Madrid is visibly defocused and Lenart exhibits poorer band-to-band co-registration, both of which can make the model more prone to spurious detections.

By contrast, stacked imagery shows little or no clear benefit from augmentation. In Kras, a few models perform better than the baseline with respect to recall while precision remains below the baseline performance for all models. In Madrid the baseline remains superior overall despite some augmented models achieving slightly higher recall. Only Toronto shows a small positive effect of augmentation in the stacked image. Overall, the precision–recall analysis suggests that augmentation is most useful when the target imagery is of lower quality, where it helps recover additional road pixels, but that this benefit weakens once the imagery becomes cleaner and less noisy.

No clear or substantial trend was observed across the low, medium, and high-severity augmentation settings. Performance differences within individual augmentation groups were generally modest and less pronounced than the differences observed between scenes. This indicates that augmentation effectiveness was determined mainly by scene-specific factors, such as image quality and acquisition conditions, rather than by the exact perturbation severity. At the same time, no single augmentation strategy dominates across all scenes. Blur-based, noise-based, and even the geometric-only baseline each achieve the best result for at least one image.

Visual inspection of the raw-scene predictions supports the quantitative results and helps explain the precision–recall trade-off (Figure 4). In Kras and Toronto, the best augmentation-based models produce visibly more complete and better connected road networks than the geometric-only baseline. In Kras, augmentation improves the detection of difficult intersections and some partially occluded road segments that are missed entirely by the baseline model, although many segments remain fragmented or undetected. In Toronto, augmentation also improves road continuity and completeness, particularly in the

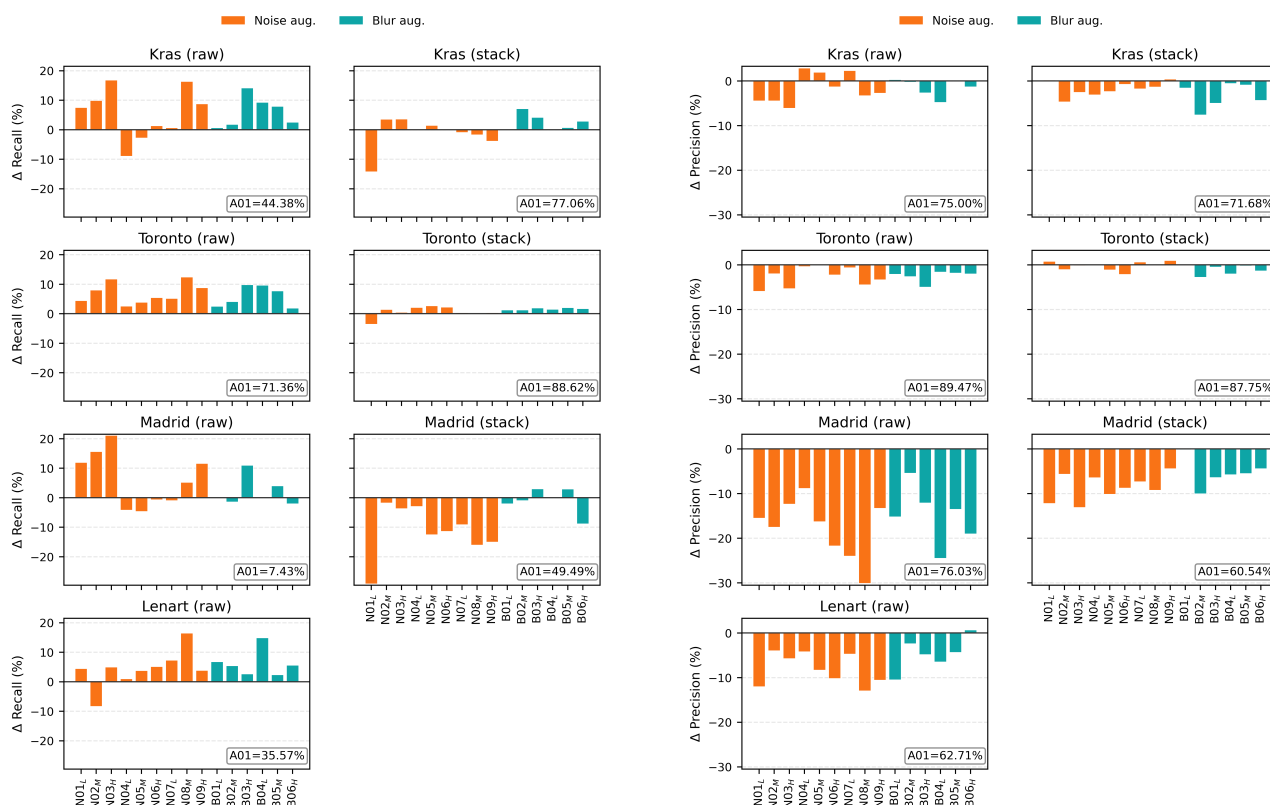


Figure 3. Relative change in 2-pixel tolerance recall (left) and precision (right) compared with the baseline (A01).

western part of the scene where vignetting is present and road density is highest (Figure 5). In both scenes, some apparent false positives correspond to roads that are missing from the reference data, indicating that part of the measured error is due to annotation incompleteness rather than model failure alone. At the same time, some false positives remain, for example along agricultural field boundaries in Kras.

The effect of augmentation is less pronounced in the more challenging scenes. In Lenart, augmentation resolves some occluded roads, but many expected road segments remain undetected and overall connectivity is still limited. In Madrid, the baseline extracts only a small number of road segments, and although augmentation clearly improves coverage, the result remains poor overall. This is consistent with the severe defocus in the Madrid image, which appears to have a strong negative effect on road extraction and is only partly mitigated by the tested augmentations. Similarly, the limited gains in Lenart suggest that the observed image-quality challenges in that scene are not fully addressed by the current augmentation design. Overall, the probability maps retain more road-like structure than is present in the final binary predictions, suggesting that part of the remaining error may be related to the fixed binarization threshold of 0.4. Hence, an improved threshold selection, calibration, or post-processing could further improve the balance between completeness and false detections.

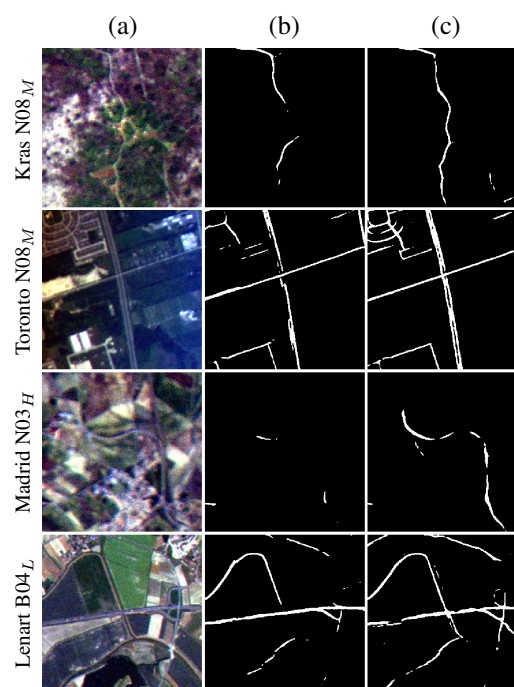


Figure 4. Qualitative comparison of raw-image predictions for representative patches: (a) the input RGB patch, (b) the prediction of the baseline model (A01), and (c) the prediction of the best augmentation-based model selected separately for each scene. The examples illustrate some observed effects of augmentation, including improved road continuity and completeness, moderate gains, and limited but visible improvement in the more challenging scenes.

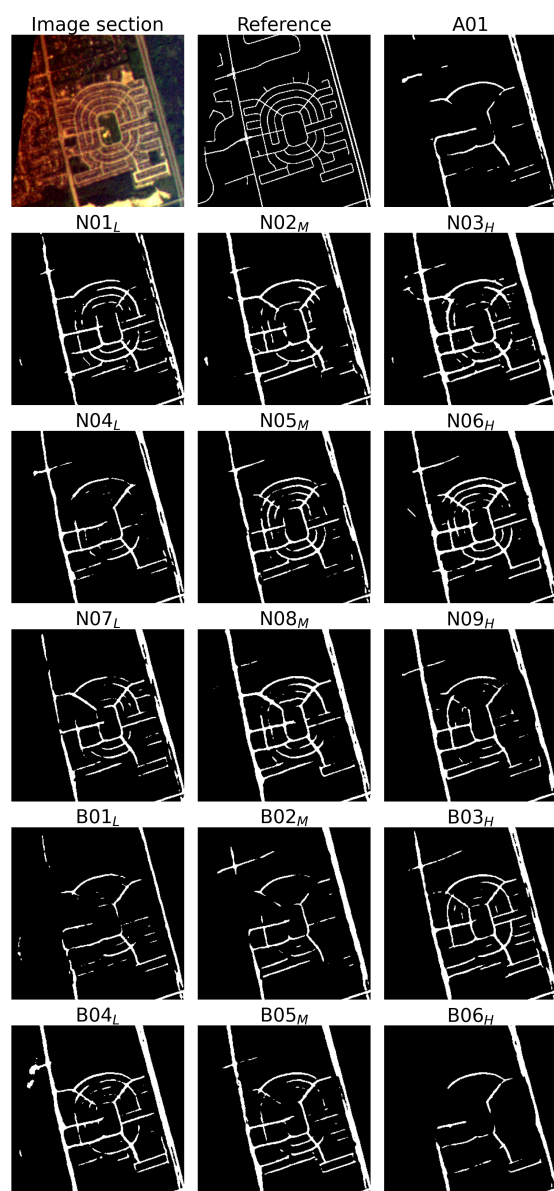


Figure 5. Example from the Toronto raw image showing the input RGB patch, the reference road mask, and the baseline prediction in the first row, followed by predictions from all augmentation-based models. The subsequent rows correspond to Gaussian noise, Perlin noise, ISO noise, Gaussian blur, and motion blur, respectively. Within each row, the three columns represent low, medium, and high augmentation severity.

5. Conclusion

This study examined whether noise and blur data augmentation during pretraining can improve road detection in small satellite imagery, with particular focus on the comparison of raw single-acquisition images and stacked images. The results show that augmentation has only a limited effect on the internal held-out test split, where differences between models remained small and the geometric-only baseline was already highly competitive. However, the full-scene evaluation revealed that augmentation is most beneficial in raw imagery, where lower SNR, vignetting, radiometric variability, and reduced sharpness make road extraction very challenging. In

these conditions, the best augmentation-based models generally improved recall, road continuity, and overall completeness, although often at the cost of a moderate increase in false positives. Stacked imagery showed little additional benefit from augmentation. No single augmentation group or severity level consistently dominated across all scenes. Scene characteristics and image quality had a stronger influence on performance than the severity level of the augmentation.

Overall, the findings indicate that augmentation can improve robustness to the image-quality challenges typical of small satellite imagery, especially when only raw acquisitions are available. More accurate and complete reference data would also help to provide better model evaluation and result interpretation. Future work should focus not only on augmentation design, but also on improved threshold selection strategy, post-processing, and architectures that better preserve road completeness and connectivity. The presented framework shows that targeted augmentation during pretraining is a practical way to improve road extraction under challenging small satellite imaging conditions and may be relevant for other small satellite systems facing similar radiometric and geometric constraints.

Acknowledgements

The project is financially supported by the Slovenian Research Agency under the codes ID J2-4488, P6-0079, and MR-58022.

References

- Bae, H.-J., Kim, C.-W., Kim, N., Park, B., Kim, N., Seo, J. B., Lee, S. M., 2018. A Perlin noise-based augmentation strategy for deep learning with small data samples of HRCT images. *Scientific reports*, 8(1), 17687.
- Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A. A., 2020. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2). <https://www.mdpi.com/2078-2489/11/2/125>.
- Chen, Z., Deng, L., Luo, Y., Li, D., Junior, J. M., Gonçalves, W. N., Nurunnabi, A. A. M., Li, J., Wang, C., Li, D., 2022. Road extraction in remote sensing data: A survey. *International journal of applied earth observation and geoinformation*, 112, 102833.
- Dias, R. L. S., Amorim, R. S. S., da Silva, D. D., Fernandes-Filho, E. I., Veloso, G. V., Macedo, R. H. F., 2024. Relative radiometric normalization for the PlanetScope nanosatellite constellation based on Sentinel-2 images. *Remote Sensing*, 16(21), 4047.
- Leach, N., Coops, N. C., Obrknezev, N., 2019. Normalization method for multi-sensor high spatial and temporal resolution satellite imagery with radiometric inconsistencies. *Computers and Electronics in Agriculture*, 164, 104893.
- Liu, G., Shan, Z., Meng, Y., Akbar, T. A., Ye, S., 2024. RD-PGNet: A road extraction network with dual-view information perception based on GCN. *Journal of King Saud University-Computer and Information Sciences*, 36(3), 102009.
- Lu, X., Weng, Q., 2025. Deep learning-based road extraction from remote sensing imagery: Progress, problems, and perspectives. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228, 122–140.

Mhangara, P., Mapurisa, W., Mudau, N., 2020. Image interpretability of nSight-1 nanosatellite imagery for remote sensing applications. *Aerospace*, 7(2), 19.

Nalepa, J., Myller, M., Cwiek, M., Zak, L., Lakota, T., Tulecziejew, L., Kawulok, M., 2021. Towards on-board hyperspectral satellite image segmentation: Understanding robustness of deep learning through simulating acquisition conditions. *Remote sensing*, 13(8), 1532.

Pehani, P., Čotar, K., Marsetič, A., Zaletelj, J., Oštir, K., 2016. Automatic geometric processing for very high resolution optical satellite data based on vector roads and orthophotos. *Remote Sensing*, 8(4), 343.

Rodič, U., Urbas, A., Fröhlich, H., Lamut, M., Bošnjak, M., Smerdu, A., Marsetič, A., Kokalj, Ž., Johnston-Lemke, B., Roth, N. et al., 2022. NEMO-HD In-Orbit Results for Agile Scanning of Cities, River Basins and 3D Objects by Satellite Video and Multispectral Cameras. *4S Symposium 2022*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Xian, G., Shi, H., Anderson, C., Wu, Z., 2019. Assessment of the impacts of image signal-to-noise ratios in impervious surface mapping. *Remote Sensing*, 11(22), 2603.

Zhang, L., Yuan, S., Dong, R., Zheng, J., Gan, B., Fang, D., Liu, Y., Fu, H., 2024. Swcare: Switchable learning and connectivity-aware refinement method for multi-city and diverse-scenario road mapping using remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 127, 103665.

Zhang, Z., Wang, Y., 2019. JointNet: A common neural network for road and building extraction. *Remote Sensing*, 11(6), 696.