
Optimizing Onboard Earth Observation AI for Robustness against Sensor Degradations

Ivar van der Spoel¹ Lorenzo Papa² Roberto Del Prete^{2,3} Andre Deutz¹ Gabriele Meoni²
Jan N. van Rijn¹

¹LIACS, Leiden University

² Φ -lab, European Space Agency

³Setanta Space

Abstract Onboard artificial intelligence can reduce Earth observation downlink requirements and shorten the time from acquisition to action, but deployed models must remain robust when sensor effects, atmospheric conditions, or onboard preprocessing alter their inputs. This work studies sensor-aware robustness on imagery from ESA’s Φ Sat-2 mission. We construct Easy LandCover (EZLC), a four-class semantic segmentation task comprising 10,000 real Φ Sat-2 patches with labels transferred from ESA WorldCover. We define five controlled degradations: band-dependent sensor noise, optical blur, atmospheric haze, radiometric gain error, and residual band co-registration error. Their formulations and severities are informed by sensor characteristics, empirical properties of the acquisitions, and Earth observation robustness literature. We investigate two model paradigms suited to Φ Sat-2 deployment. First, a Neural Architecture Search framework ‘PyNAS’ is extended with NSGA-II-like multi-objective selection over performance, robustness, and efficiency. Second, the pretrained Φ SatNet encoder is adapted through supervised perturbation-aware fine-tuning or label-free multi-layer self-distillation. Robustness-aware NAS increases mean perturbed mIoU from 0.48 to 0.54 and prediction consistency from 0.78 to 0.89, while the selected model contains fewer than 0.1M parameters. Supervised Φ SatNet encoder adaptation raises mean perturbed mIoU from 0.47 to 0.54 and prediction consistency from 0.81 to 0.91 while preserving a clean mIoU of 0.57. Self-distillation provides smaller but measurable robustness gains without semantic labels during encoder adaptation. It improves representation alignment between clean and perturbed patches across MSE, cosine similarity, CKA, and MMD, yet produces smaller downstream gains than supervised adaptation. These results indicate that representation invariance is a useful diagnostic of robustness, but not a substitute for task-level evaluation.

1 Motivation and Contributions

Earth observation (EO) satellites support applications ranging from land-cover mapping to disaster response. Conventional workflows downlink imagery before analysis, whereas onboard AI can filter observations, create compact products, and support faster decisions Giuffrida et al. (2022). ESA’s Φ Sat-2 mission demonstrates onboard AI under strict resource constraints: its eight-band optical sensor is paired with an Intel Movidius Myriad 2 processor, making model size, memory use, inference latency, and hardware compatibility important deployment constraints Longép   et al. (2023).

Compactness alone is insufficient. Models are commonly developed on selected, preprocessed, and quality-controlled data, whereas deployment imagery may follow a different data distribution. Detector noise, reduced optical sharpness, atmospheric scattering, radiometric conversion errors, or imperfect band alignment can alter the input. The central question is therefore how to evaluate and improve the robustness of compact models operating on real Φ Sat-2 imagery against sensor, environmental, and onboard-processing degradations.

This work makes five connected contributions: (i) a labeled semantic-segmentation task on real Φ Sat-2 acquisitions, (ii) a sensor-aware robustness evaluation framework, (iii) robustness-aware multi-objective NAS for performance, robustness, and efficiency trade-offs, (iv) supervised and label-free adaptation of the compact pretrained Φ SatNet encoder, and (v) analysis of whether improved representation alignment corresponds to improved downstream robustness.

2 Sensor-Aware Benchmark and Robustness Methods

2.1 Easy LandCover and controlled degradations

EZLC is constructed from real Φ Sat-2 patches produced by an external alignment pipeline that pairs Φ Sat-2 acquisitions with corresponding Sentinel-2 observations¹. Using the resulting georeferencing information, we transfer ESA WorldCover labels to the Φ Sat-2 image grid Zanaga et al. (2022). WorldCover’s eleven classes are aggregated into four: background, vegetation, built-up surfaces, and water. Patches whose dominant class exceeds 80% are removed, and 10,000 256×256 patches are selected to target a 25% aggregate pixel share per class. Experiments use an 8,000/2,000 train-validation split with seed 42. Perturbations are applied before band-wise z-score normalization.

We study controlled covariate shifts in which the input changes while the semantic target remains fixed. The five default perturbations and their corresponding strengths are: band-dependent Gaussian sensor noise with SNR factor $s = 0.25$; optical blur with modulation-transfer-function value $M_b = 0.039$; atmospheric haze with transmission $t = 0.50$; multiplicative radiometric gain $\alpha = 1.4$; and residual band misalignment with maximum displacement $\Delta_{\max} = 5.2$ pixels. Their mechanisms and strengths are motivated by Φ Sat-2 sensor characteristics, measurements on real acquisitions, and comparable EO robustness methodology Long  p   et al. (2024); Li et al. (2025). Perturbation-aware supervised training samples one of clean, noise, blur, haze, gain, misalignment uniformly per observation. Only one perturbation is applied at a time, enabling controlled evaluation of individual degradation mechanisms.

2.2 Robustness-aware Neural Architecture Search

The task-specific route extends PyNAS, a neural architecture search framework for compact models Del Prete et al. (2025). Architectures exceeding 100,000 trainable parameters are rejected. After candidate training, the multi-objective search considers

$$\max; \text{mIoU}_{\text{clean}}(a), \quad \max; \text{Cons}(a), \quad \min; \tau(a), \quad (1)$$

where Cons is mean clean-perturbed pixel-level prediction consistency and τ is cumulative validation-set inference time measured on an H100 GPU. Perturbed mIoU is intentionally not a search objective: a model can remain consistently wrong, so consistency must be optimized together with clean task performance.

Candidates are ranked by non-dominated sorting and crowding distance, following NSGA-II Deb et al. (2002), within the PyNAS evolutionary loop. The search uses a population of 10 candidates for 15 iterations, with three elites, a mutation probability of 0.1, and three random injections per iteration. To limit the computational cost of evaluating many candidate architectures, each candidate is trained for at most 10 epochs during the search. After the search, the nondominated architecture with the highest clean validation mIoU is retrained from random initialization for 50 epochs under the corresponding clean or perturbation-aware regime. This longer training budget is used for the final model evaluation and is consistent with the 50-epoch training budget used for the other models in this study.

¹<https://github.com/clementgilli/phisat2>

2.3 Robust adaptation of Φ SatNet

The task-agnostic route starts from the compact pretrained Φ SatNet encoder, denoted by E_0 ¹. First, the encoder is frozen and a clean EZLC decoder is trained. For supervised robustness adaptation, this clean decoder is then frozen while all encoder stages are fine-tuned with the same perturbation-aware regime, giving a supervised encoder E_{sup} . Any gain must therefore be produced by features that remain compatible with the existing clean task head.

For label-free robustness adaptation, a frozen copy of the pretrained encoder acts as a clean teacher and a trainable copy as the student. At encoder stage l , teacher feature T_l is produced from clean input x , while student features C_l and P_l are produced from x and perturbed $c(x)$, respectively. To account for differences in feature magnitude between encoder stages, each layer loss is normalized by the teacher’s mean squared activation over the self-distillation training set, denoted by e_l . The objective is

$$\mathcal{L}_{\text{SD}} = \frac{1}{4} \sum_{l=1}^4 \frac{\text{MSE}(P_l, T_l)}{e_l} + \lambda \frac{1}{4} \sum_{l=1}^4 \frac{\text{MSE}(C_l, T_l)}{e_l}. \quad (2)$$

We set $\lambda = 2$, giving the clean-anchor term twice the weight of the perturbation term to discourage drift from the pretrained representation. The perturbation term maps degraded inputs toward the clean teacher representation, whereas the clean-anchor term limits representation drift. This self-distilled encoder is denoted by E_{SD} . No EZLC labels or segmentation decoder are used during this stage. The adapted encoder is subsequently frozen and a new decoder is trained only on clean labeled data, so robustness gains must transfer through the encoder representation.

3 Evaluation and Results

We report clean mIoU, mean mIoU over the five default perturbations, relative robustness, and mean prediction consistency. Prediction consistency measures the fraction of pixels whose predicted class remains unchanged between a clean observation and its perturbed counterpart, averaged over the five perturbations. Unlike mIoU, it does not measure correctness with respect to the reference mask and is therefore interpreted together with clean and perturbed task performance.

For Φ SatNet, globally pooled features from four encoder stages are additionally compared using MSE, cosine similarity, linear CKA Kornblith et al. (2019), and MMD Gretton et al. (2012). Self-distillation uses a separate 9,000/1,000 split over all 10,000 images to expose the label-free adaptation stage to more observations while retaining a validation subset for checkpoint selection.

Table 1: Clean and mean perturbed segmentation performance and prediction consistency for the proposed compact approaches and selected larger reference models. R_{rel} is mean perturbed mIoU divided by clean mIoU. Mean consistency is the average clean–perturbed pixel-level prediction consistency across the five perturbations.

Family	Variant	Params. [M]	Clean	Mean pert.	R_{rel}	Mean cons.
NAS	clean-trained	< 0.1	0.60	0.48	0.800	0.78
NAS	robustness-aware	< 0.1	0.57	0.54	0.947	0.89
Φ SatNet	E_0	0.35	0.57	0.47	0.825	0.81
Φ SatNet	E_{sup}	0.35	0.57	0.54	0.930	0.91
Φ SatNet	E_{SD}	0.35	0.55	0.48	0.855	0.84

Across the clean-trained models, haze is the most consistently damaging default perturbation: it produces both large reductions in mIoU and the lowest prediction consistency. Noise particularly affects the clean NAS model, whereas Φ SatNet is more sensitive to radiometric gain. Blur and band misalignment have smaller effects at the evaluated default strengths.

Robustness-aware NAS increases mean perturbed mIoU from 0.48 to 0.54 and mean prediction consistency from 0.78 to 0.89, while clean mIoU changes from 0.60 to 0.57. The improvement therefore reflects both better task performance under degradation and more stable predictions. Consistency cannot, however, be interpreted independently of accuracy: one candidate on the NAS Pareto set reaches 0.96 consistency with only 0.17 clean mIoU. The selected candidate instead combines 0.57 clean mIoU with 0.89 consistency, illustrating why the search jointly considers clean performance, prediction stability, and efficiency.

For Φ SatNet, supervised encoder adaptation increases mean perturbed mIoU from 0.47 to 0.54 and mean consistency from 0.81 to 0.91 while preserving clean mIoU at 0.57. Label-free self-distillation gives a smaller improvement: E_{SD} reaches 0.48 mean perturbed mIoU and 0.84 consistency, while clean mIoU decreases to 0.55. Supervised adaptation therefore produces the stronger robustness improvement for the fixed EZLC task, whereas self-distillation demonstrates that some prediction stability can be introduced into the encoder without semantic labels during robustness adaptation.

Representation alignment does not reliably rank downstream robustness. Relative to E_0 , E_{SD} improves MSE, cosine similarity, and CKA for all five perturbations and MMD for four, yet its downstream improvement is small. Conversely, E_{sup} produces substantially larger downstream gains despite mixed changes in the representation metrics. Representation alignment is therefore useful as a diagnostic of encoder behavior, but not as a substitute for task performance or prediction consistency.

4 Discussion and Conclusion

Three conclusions follow. First, robustness cannot be characterized by clean accuracy or prediction consistency alone. Clean and perturbed mIoU measure whether predictions remain correct, whereas prediction consistency measures whether they remain stable. The NAS Pareto set demonstrates the distinction directly: very high consistency can coexist with poor segmentation performance. Robust evaluation should therefore consider clean performance, perturbed performance, and prediction consistency together.

Second, supervised perturbation-aware training is the strongest strategy for the studied fixed task. A robustness-aware NAS model below 0.1M parameters reaches the same rounded mean perturbed mIoU (0.54) as the supervised Φ SatNet encoder, while making efficiency explicit during search.

Third, self-distillation demonstrates that some robustness can be introduced into a reusable encoder without semantic labels during robustness adaptation, but the smaller gain and loss of clean performance show the cost of enforcing global feature alignment without task information. Future representation objectives should better preserve pretrained features.

The conclusions are bounded by several limitations. EZLC represents one four-class downstream task, whereas Φ SatNet is intended for multiple downstream tasks. The controlled perturbations approximate plausible sensor, environmental, and processing effects and are evaluated separately rather than in combination. Finally, parameter count and H100 inference time do not establish on-board suitability: latency, memory use, numerical precision, energy consumption, and compatibility must ultimately be measured on the Intel Movidius Myriad 2 processor.

Overall, sensor-aware perturbations make robustness an explicit design and adaptation objective rather than an assumption inferred from clean validation accuracy. For a known task with labels, perturbation-aware supervision provides the strongest gains, whereas under strict capacity constraints, multi-objective NAS discovers competitive robust models at very small scale. When labels are unavailable during encoder adaptation, multi-layer self-distillation provides a weaker but reusable label-free route.

References

- Deb, K., Pratap, A., Agarwal, S., and Meyerivian, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.
- Del Prete, R., Thind, P. K., Mazzeo, A., Whitley, M., Papa, L., Longép  , N., and Meoni, G. (2025). Optimizing deep learning models for on-orbit deployment through neural architecture search. *Scientific Reports*, 15(1):37783.
- Giuffrida, G., Fanucci, L., Meoni, G., Bati  , M., Buckley, L., Dunne, A., van Dijk, C., Esposito, M., Hefele, J., Vercruyssen, N., Furano, G., Pastena, M., and Aschbacher, J. (2022). The Φ -Sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Sch  lkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Li, X., Tao, Y., Zhang, S., Liu, S., Xiong, Z., Luo, C., Liu, L., Pechenizkiy, M., Zhu, X., and Huang, T. (2025). REOBench: Benchmarking robustness of earth observation foundation models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Long  p  , N., Melega, N., Marchese, V., Paskeviciute, A., Babkina, I., Petrelli, I., Casaburi, M., Peressutti, D., and Oman Kadunc, N. (2023). Φ sat-2 mission overview for the #orbitalai challenge. Technical note, European Space Agency. Issue 1, Revision 0.
- Long  p  , N., Petrelli, I., Oman Kadunc, N., Peressutti, D., Del Prete, R., Casaburi, M., Babkina, I., Vercruyssen, N., Callejo Luis, E., Mor  n Elorza,   ., Marchese, V., Paskeviciute Kidron, A., and Melega, N. (2024). Simulation of multispectral and hyperspectral EO products for onboard machine learning application. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:17651–17665.
- Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., and Arino, O. (2022). ESA WorldCover 10 m 2021 v200.