

Cross-Biome Evaluation of Deep Learning Architectures for Semantic Segmentation of Deforestation in the Amazon and Caatinga

Pedro Andrade¹ and Anderson Tenório Sergio¹

CESAR School, Recife, Brazil

 pedrofdeandrade@gmail.com

ats@cesar.school

Abstract. The accurate mapping of deforested areas is critical for environmental monitoring and policymaking, particularly in ecologically diverse countries. Although deep learning models can achieve high in-domain accuracy, their ability to generalize across distinct biomes remains a significant challenge. This study evaluates the cross-biome generalization of Prithvi-EO-2.0-300M-TL, a geospatial foundation model, against a U-Net with a ResNet34 encoder for static semantic segmentation of accumulated deforested and anthropized areas. Using multispectral Sentinel-2 Level-2A imagery and official PRODES/INPE reference masks, both models are evaluated across selected Amazon and Caatinga study areas under in-domain, source-only cross-biome, and multi-biome training configurations. The protocol uses spatial block splits, an all-positive baseline, predicted positive fraction diagnostics, qualitative prediction/error maps, and computational cost measurements. Results show strong in-domain performance in both biomes, with Prithvi providing only modest IoU gains over U-Net. However, source-only Amazon-to-Caatinga transfer collapses both models toward all-positive predictions, yielding IoU values close to the Caatinga all-positive baseline. In the reverse direction, Prithvi again over-predicts near the Amazon all-positive baseline, while U-Net degrades differently through underprediction, achieving higher IoU but missing a substantial proportion of positive pixels. Multi-biome training mitigates collapse and preserves performance close to biome-specific training. Computationally, Prithvi is approximately 13.26 times larger, about 9.3 times slower per seed on average, and substantially more memory-intensive than U-Net. These findings indicate that geospatial foundation models should not be assumed to provide robust cross-biome transfer without target-biome supervision, explicit validation, and careful cost-benefit analysis.

Keywords: Semantic segmentation · Deforestation monitoring · Remote sensing · Sentinel-2 · PRODES · Prithvi-EO · U-Net · Cross-biome evaluation · Amazon · Caatinga

1 Introduction

Deforestation monitoring is central to environmental governance, climate policy, biodiversity protection, and land-use planning. Remote sensing enables the observation of large and often inaccessible regions, while deep learning has improved pixel-level mapping of forest disturbance and land-cover patterns. However, a model that performs well within one ecological domain can fail under spectral, spatial, seasonal, and land-use shifts when applied elsewhere. This issue is particularly relevant in Brazil, where operational monitoring must span ecosystems with distinct vegetation structure, soil exposure, hydrology, seasonality, and anthropization histories.

This work investigates whether segmentation models trained for accumulated deforestation in one biome generalize to another under a source-only supervised fine-tuning protocol. We focus on selected study areas from two contrasting Brazilian biomes: the Amazon, dominated by dense tropical rainforest, and the Caatinga, a semi-arid biome with sparse and seasonally variable vegetation. The target variable is not recent annual deforestation. Following the PRODES class mapping used here, the positive class represents accumulated deforested or anthropized areas up to 2023, including the historical pre-2000 deforestation class d2000 when present. Valid native vegetation classes are negative, while hydrology, nodata, and classes later than the Sentinel-2 reference year are ignored.

We compare a convolutional U-Net baseline with a ResNet34 encoder to Prithvi-EO-2.0-300M-TL, a geospatial foundation model [1]. Foundation models have been proposed as a path toward transferable Earth observation systems, yet recent remote-sensing studies also show that heterogeneous environments and sensor/domain mismatches can require explicit adaptation strategies [2]. This paper evaluates that issue in a compact but controlled cross-biome experiment.

The contributions are fourfold. First, we define a spatially separated protocol for in-domain, source-only cross-biome, and multi-biome training. Second, we compare a lightweight convolutional architecture with a much larger foundation model under the same preprocessing, split, training schedule, and evaluation metrics. Third, we include an all-positive baseline and predicted positive fraction diagnostics to identify degenerate behavior. Fourth, we add qualitative prediction/error maps and computational cost measurements to support interpretation beyond aggregate IoU.

2 Related Work

2.1 Deep Learning for Forest Monitoring and Deforestation Mapping

Deep learning is widely used for semantic segmentation in remote sensing because encoder-decoder architectures can learn spatially structured representations directly from image patches. U-Net remains influential because skip connections combine high-level semantic features with spatial detail [3]. The ResNet family of encoders is also widely used because residual connections support deep feature extraction [4].

In forest-related remote sensing, U-Net-style models remain competitive. For example, Wagner et al. [5] used a U-Net regression model with Planet NICFI imagery and LiDAR-derived canopy height models to map Amazon forest height at high spatial resolution. Although canopy-height regression differs from binary deforestation segmentation, such work demonstrates that locally trained encoder-decoder models can be effective in large-scale Amazon monitoring tasks.

Forest disturbance and land-cover mapping face class imbalance, boundary ambiguity, spatial autocorrelation, and heterogeneous landscape composition. Tian et al. [2] discuss that environmental variation and disturbance-pattern heterogeneity complicate model transfer across regions. These concerns motivate the cross-biome design of the present study.

2.2 Geospatial Foundation Models

Foundation models for vision and Earth observation aim to learn broad representations that can be adapted to downstream tasks. Recent work on forest

disturbance mapping has shown that foundation-model-based approaches can improve performance when adapted with task-specific architectural components, local-feature branches, and transfer strategies [2]. Nevertheless, this literature also indicates that generic pretraining alone is not necessarily sufficient under domain, sensor, or landscape mismatch, especially when forest structure, disturbance patterns, and image characteristics vary across regions.

The present study examines this issue using Prithvi-EO-2.0-300M-TL, a geospatial foundation model [1], under a source-only supervised fine-tuning protocol. The model checkpoint was accessed through the IBM-NASA geospatial model repository [6]. We compare it with a U-Net baseline using the same preprocessing pipeline, spatial splitting strategy, training schedule, and evaluation metrics in cross-biome experiments between Amazon and Caatinga study areas.

2.3 Cross-Domain Generalization in Remote Sensing

Remote-sensing models are affected by domain shifts caused by acquisition date, illumination, atmosphere, seasonality, vegetation structure, soil exposure, and land-use composition. Random within-scene splits can overestimate robustness when neighboring chips share local texture, phenology, or acquisition conditions. Spatially separated validation and cross-region tests therefore provide a stronger indication of generalization.

Large-area mapping studies also emphasize consistent reference construction, multi-source data integration, and careful treatment of ecological heterogeneity [7]. The present work follows this direction by evaluating Amazon-to-Caatinga and Caatinga-to-Amazon transfer, not only within-biome performance.

3 Methodology

3.1 Study Areas

The experiment uses one selected study area in each biome. The Amazon area is defined by the bounding box (-63.00, -9.00, -62.60, -8.60). The Caatinga area is defined by the bounding box (-40.70, -9.70, -40.30, -9.30), with the Sentinel-2 MGRS tile constrained to 24LUQ. These areas should be interpreted as selected study regions, not as exhaustive samples of the entire biomes. This distinction is important for the external validity of the results.

3.2 Data Sources and Preprocessing

The input imagery consists of Sentinel-2 Level-2A multispectral observations from 2023 [8], obtained using the Microsoft Planetary Computer STAC catalog [9]. The selected Amazon scene was S2A_MSIL2A_20230704T142721_R053_T20LNR_20230704T231904, with 0.0% cloud cover. The selected Caatinga scene was S2A_MSIL2A_20230915T130251_R095_T24LUQ_20240908T123321, also with 0.0% cloud cover. Six bands were used: B02, B03, B04, B8A, B11, and B12, corresponding to blue, green, red, narrow NIR, SWIR-1, and SWIR-2. Each band was clipped to the study-area bounding box and aligned to the reference grid before stacking.

The same Sentinel-2 L2A six-band stack was used for both U-Net and Prithvi-EO-2.0 to ensure a controlled comparison between architectures under identical input conditions. Although Prithvi-EO-2.0 was originally pre-trained on NASA HLS V2 data at 30 m resolution, using HLS only for Prithvi would introduce an additional data-source confound, making it difficult to determine whether performance differences were caused by the model architecture or by differences in the input product. Therefore, we prioritized experimental comparability and the preservation of 10–20 m spatial information relevant to deforestation patterns. This product and resolution mismatch is acknowledged as an intentional methodological trade-off and is partially addressed through the supervised fine-tuning stage.

Reference masks were derived from official PRODES/INPE GeoTIFF products downloaded from the TerraBrasilis data portal [10]. Specifically, we used the files `prodes_amazonia_legal_2025_v20260408.tif` and `prodes_caatinga_2024_v20260407.tif`. Each PRODES raster was clipped to the corresponding study-area bounding box and reprojected to match the Sentinel-2 Level-2A grid using nearest-neighbor resampling. Chips of 224×224 pixels were generated from the aligned Sentinel-2 stack and corresponding PRODES-derived mask. Chips were extracted on a non-overlapping grid (stride = 224 px, i.e., equal to the chip size, so there is no spatial overlap between chips), and only chips with at least 50% valid pixels were retained (a valid pixel is any pixel whose reference mask is not `ignore_index = 255`). Table 1 summarizes the study-area and chip-generation statistics.

Table 1. Study-area and chip-generation summary

Biome	Sentinel-2 scene/date	PRODES source	Valid chips	Mean valid fraction	Mean positive fraction
Amazon	2023-07-04, tile 20LNR	<code>prodes_amazonia_legal_2025_v20260408.tif</code>	360	0.981	0.164
Caatinga	2023-09-15, tile 24LUQ	<code>prodes_caatinga_2024_v20260407.tif</code>	357	0.964	0.534

3.3 PRODES Reference Mask Construction and Temporal Filtering

The reference mask has three possible values: 1 for accumulated deforested/anthropized areas up to 2023, 0 for valid native vegetation classes, and 255 for ignored pixels. The positive class includes d2000, annual deforestation classes up to d2023, and residue classes up to r2023. The d2000 class is treated as positive because it represents accumulated pre-2000 deforestation in the PRODES legend used in this experiment. Future classes relative to the Sentinel-2 reference year, namely d2024, d2025, r2024, and r2025, are not treated as native vegetation; instead, they are assigned to `ignore_index = 255`. These classes are excluded for temporal consistency: because the Sentinel-2 scene is from 2023, areas labelled by PRODES as deforested or residual only in 2024–2025 may not correspond to the land-cover state visible in the input image. Treating them as positive examples would therefore introduce a potential label–image temporal mismatch. Hydrology and nodata pixels are also ignored. Vegetation classes 100 and 101 are treated as the valid negative class.

This design frames the task as static semantic segmentation of accumulated deforested and anthropized areas, rather than annual change detection. Therefore, the results should be interpreted as the segmentation of accumulated deforestation/anthropization patterns derived from PRODES reference data, not as the detection of newly deforested areas.

3.4 Model Architectures

The baseline model is a U-Net with a ResNet34 encoder pretrained on ImageNet, implemented through `segmentation_models_pytorch` [11]. The network receives six input channels and outputs two valid semantic classes. Because the ResNet34 encoder is pretrained on three-channel ImageNet data, its first convolutional layer was adapted to accept six input channels using the library's `patch_first_conv` mechanism. The pretrained convolutional filters were copied cyclically across the six input-channel positions and rescaled by a factor of $3/6 = 0.5$, preserving the approximate activation magnitude of the original layer. This initialization is positional rather than spectrally informed: the mechanism does not assign filters according to the physical meaning of Sentinel-2 bands. Because all encoder parameters remain trainable during supervised fine-tuning, the expanded input layer serves as an informed initialization rather than a fixed or random mapping. We used `segmentation_models_pytorch` version 0.5.0. Pixels assigned to `ignore_index = 255` are excluded from the loss and metrics.

The comparison model is Prithvi-EO-2.0-300M-TL [1], implemented through TerraTorch [12] as an encoder-decoder segmentation model. The Prithvi backbone uses the six Sentinel-2-like bands listed above and a single temporal frame. Its decoder is configured as a U-Net-like decoder with channels [512, 256, 128, 64], connected to the backbone through the `SelectIndices`, `ReshapeTokensToImage`, and `LearnedInterpolateToPyramidal` neck modules. In this experiment, Prithvi-EO is evaluated under a full supervised fine-tuning protocol, with all model parameters set as trainable.

The two architectures differ substantially in size. The U-Net/ResNet34 model has 24,445,922 trainable parameters, while Prithvi-EO-2.0-300M-TL has 324,204,676 trainable parameters, approximately 13.26 times more. Thus, the comparison is not a parameter-matched contest. It is instead a practical comparison between a relatively lightweight convolutional baseline and a much larger geospatial foundation model. Table 2 reports the model sizes.

Table 2. Model size

Model	Backbone / encoder	Trainable parameters	Relative size
U-Net	ResNet34	24,445,922	1.00x
Prithvi-EO-2.0	Prithvi-EO-2.0-300M-TL	324,204,676	13.26x

3.5 Experimental Protocol

The protocol evaluates six train/test configurations for each model: Amazon-to-Amazon, Caatinga-to-Caatinga, Amazon-to-Caatinga, Caatinga-to-Amazon, Both-to-Amazon, and Both-to-Caatinga. In the multi-biome condition, the training and validation sets are obtained by concatenating the corresponding Amazon and Caatinga spatial splits, while the final test set remains biome-specific.

Spatial block splitting was used to reduce the risk of spatial leakage between training, validation, and test partitions. Chips were extracted on a non-overlapping grid and grouped according to their row and column indices on the chip grid. Each spatial block was defined as a 3×3 arrangement of chips, corresponding to a nominal extent of 672×672 pixels or approximately 6.72×6.72 km on the 10 m working grid. Whole blocks, rather than individual chips, were randomly assigned to training, validation, or test partitions using a seeded random generator, with nominal proportions of 60%, 20%, and 20%, respectively. This strategy reduces the risk that chips with high local spatial autocorrelation are distributed across different partitions. However, no buffer zone was imposed between blocks; therefore, spatial autocorrelation across adjacent blocks may remain. The split was not stratified by positive-class prevalence. Consequently, the prevalence of accumulated deforestation/anthropization among valid test pixels varied across seeds, ranging from approximately 0.330 to 0.549 in the Caatinga test splits.

The protocol was repeated for seeds 42, 123, and 7. Each seed affects spatial-block allocation, model initialization, and training-data order; therefore, variation across seeds reflects both split variability and optimization randomness. Three seeds were used as a compromise between statistical coverage and computational budget: the study evaluates six train/test configurations for two models, resulting in 12 model-configuration conditions. Prithvi-EO had a mean total runtime per seed approximately 9.3 times that of U-Net; thus, each additional seed would add 12 training/evaluation runs at substantial computational cost. Repeated spatially blocked cross-validation, ideally combined with evaluation on geographically independent regions, is an important direction for future work.

Each model is trained for 15 epochs using AdamW with learning rate $1e-4$, weight decay 0.05, batch size 8, and a cosine annealing learning-rate schedule. The loss is cross-entropy with `ignore_index = 255`. Checkpoint selection is based on validation loss from the training domain. In Amazon-to-Caatinga, for example, the model is trained on Amazon training chips, validated on Amazon validation chips, and evaluated only at the end on Caatinga test chips.

For each seed and configuration, the notebook records training time, evaluation time, total runtime, peak GPU memory allocation, total parameters, and trainable parameters. These measurements are used to assess the cost-benefit trade-off between the compact convolutional baseline and the larger foundation model.

3.6 Evaluation Metrics, Baseline, and Qualitative Outputs

Precision, recall, F1-score, and intersection over union (IoU) are computed over valid pixels only, where a valid pixel is any pixel whose reference-mask value differs from `ignore_index = 255`. Metrics are computed by pooling the predictions and reference masks from all test chips into a single set of valid pixels and deriving a single confusion matrix for each seed and experimental configuration. Thus, the reported per-seed metrics correspond to pixel-level micro-averaging rather than to an average of chip-level metrics. Every valid pixel contributes equally, while chips contribute in proportion to their number of valid pixels.

The main metric is IoU for the positive class, defined as accumulated deforested/anthropized area. It is computed as $\text{IoU} = \text{TP} / (\text{TP} + \text{FP} + \text{FN})$, where TP, FP, and FN denote true-positive, false-positive, and false-negative pixels for the

positive class. Aggregate results are reported as the mean and standard deviation of the per-seed metrics.

The experiment also reports the reference positive fraction, predicted positive fraction, and an all-positive baseline. The reference positive fraction is the proportion of valid test pixels labelled as accumulated deforestation/anthropization in the PRODES-derived mask. The all-positive baseline is a trivial prediction baseline in which every valid pixel is predicted as positive, while the ground-truth labels remain unchanged. Its IoU therefore equals the positive-class prevalence in the corresponding test set. The predicted positive fraction (Pred+) is the proportion of valid pixels classified as positive by the model. This baseline is important because, when positive-class prevalence is high, a model can obtain moderate IoU while failing to distinguish native vegetation from accumulated deforested or anthropized areas.

Qualitative outputs were generated for representative test chips, including Sentinel-2 RGB, false-color composites, PRODES reference masks, model predictions, and pixel-level error maps. In the mask visualizations, black represents native vegetation, white represents accumulated deforestation/anthropization, and yellow represents ignored pixels. In error maps, white denotes true positives, black true negatives, orange false positives, blue false negatives, and yellow ignored pixels.

4 Experiments

4.1 In-Domain Performance

In-domain performance is strong for both architectures. In Amazon-to-Amazon, U-Net reaches an IoU of 0.875, while Prithvi reaches 0.885. In Caatinga-to-Caatinga, U-Net reaches 0.763, while Prithvi reaches 0.773. These results indicate that both models can learn the target segmentation task when target-biome training data are available. Performance is nevertheless lower in the Caatinga study area than in the Amazon study area, suggesting that the semi-arid target domain is more challenging under this setup.

Prithvi produces modest in-domain gains over U-Net: +0.010 IoU in Amazon and +0.010 IoU in Caatinga. These gains are small relative to the difference in model size and computational cost. Figure 1 summarizes the mean IoU across the six train/test configurations.

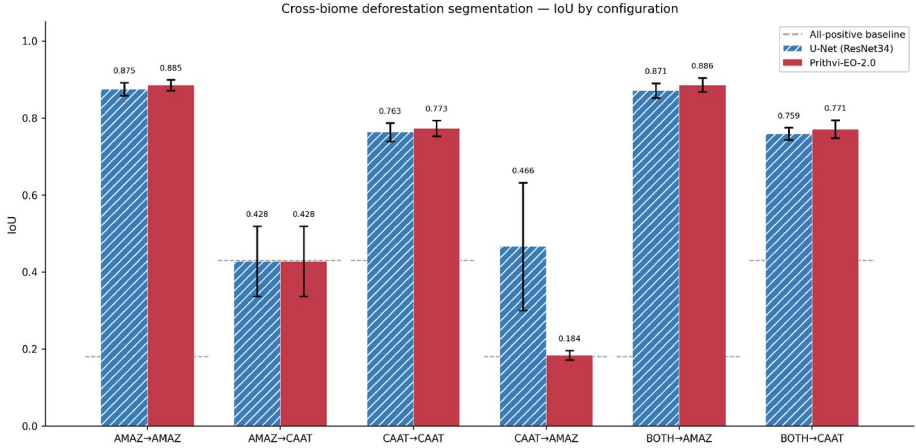


Fig. 1. IoU by train/test configuration. Bars show mean IoU across three seeds; error bars show standard deviation. Dashed segments indicate the all-positive baseline for the corresponding test biome.

4.2 Cross-Biome Transfer

Cross-biome transfer provides evidence of substantial domain shift between the selected study areas. In Amazon-to-Caatinga, U-Net obtains an IoU of 0.428 and Prithvi obtains an IoU of 0.428, both essentially equal to the Caatinga all-positive baseline of 0.430. Proximity to this baseline alone is not sufficient to establish degenerate behavior, because IoU depends jointly on false positives and false negatives.

The direct evidence of near-all-positive prediction behavior is provided by the predicted positive fraction and the associated precision-recall pattern. Pred+ reaches 0.993 for U-Net and 0.995 for Prithvi, indicating that both models classify nearly all valid Caatinga pixels as accumulated deforestation/anthropization. Mean recall is approximately 0.991 for U-Net and 0.993 for Prithvi, whereas mean precision is approximately 0.430 for both models, close to the mean positive-class prevalence of the Caatinga test sets. Together, these results show that the models recover nearly all positive pixels while incorrectly labelling nearly all negative pixels as positive. Therefore, both models collapse toward all-positive predictions, and their near-baseline IoU is a consequence and corroborating indicator of this systematic overprediction, rather than sufficient evidence of it on its own.

The reverse transfer is asymmetric and model-dependent. In Caatinga-to-Amazon, Prithvi also exhibits a near-all-positive prediction pattern: Pred+ = 0.936, mean precision = 0.186, and mean recall = 0.971. Its mean IoU of 0.184 is close to the Amazon all-positive baseline of 0.180.

U-Net displays a different error regime. It achieves a higher mean IoU of 0.466, but with substantial variation across seeds. Its predicted positive fraction is 0.105, below the mean Amazon positive-class prevalence of 0.180. Together with mean recall of 0.500 and mean precision of 0.857, this pattern indicates conservative underprediction: the model identifies relatively few positive pixels, most of which are correct, but misses a substantial proportion of accumulated deforestation/anthropization pixels.

Thus, source-only cross-biome transfer produces model- and direction-dependent error patterns: near-all-positive overprediction for both models in Amazon-to-Caatinga and for Prithvi in Caatinga-to-Amazon, versus underprediction for U-Net in Caatinga-to-Amazon. These patterns are further analysed in Section 4.4. Figure 2 contrasts in-domain and source-only cross-biome IoU for both transfer directions, and Fig. 3 reports the full train/test IoU matrices.

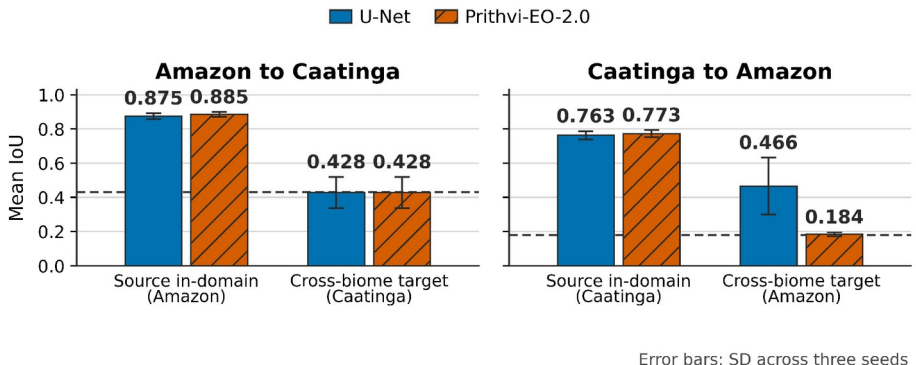


Fig. 2. Cross-biome degradation. The bars compare source-biome in-domain performance with source-only cross-biome performance. U-Net is solid and Prithvi-EO-2.0 is hatched; error bars show the standard deviation across three seeds. Dashed lines indicate the target-specific all-positive IoU baseline.

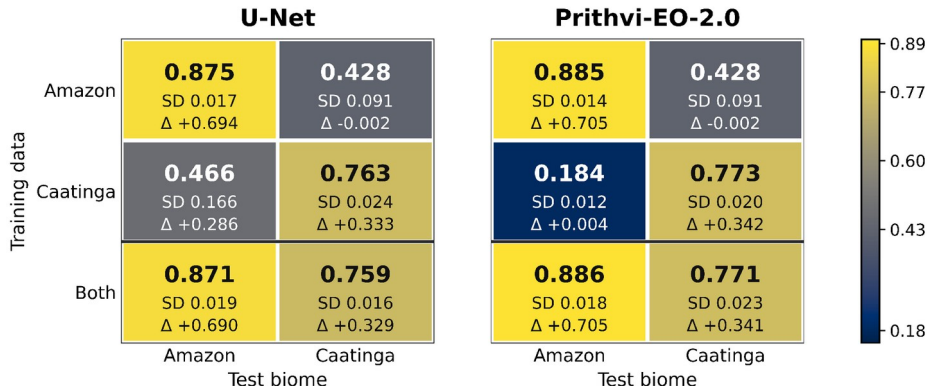


Fig. 3. Train/test IoU matrices for U-Net and Prithvi-EO-2.0. Each cell reports mean IoU, standard deviation across seeds, and the margin relative to the target-specific all-positive baseline.

4.3 Multi-biome Training

Multi-biome training substantially reduces the cross-biome degradation observed under source-only transfer. In Both-to-Amazon, U-Net reaches an IoU of 0.871 and Prithvi reaches 0.886. In Both-to-Caatinga, U-Net reaches 0.759 and Prithvi reaches 0.771. These values remain close to those obtained with target-biome training, suggesting that including samples from both biomes during training mitigates the

extreme prediction failures observed under source-only transfer in the selected study areas.

The differences between multi-biome and target-biome training are small. This suggests that multi-biome training is primarily useful as a robustness strategy when a system must operate across more than one biome, rather than as a strong accuracy enhancer over biome-specific training.

4.4 Collapse Analysis: Predicted Positive Fraction

Predicted positive fraction (Pred+) diagnoses collapse toward all-positive behavior. In-domain and multi-biome configurations have Pred+ close to their reference positive fractions. In Amazon-to-Caatinga, both models predict nearly all valid pixels as positive (Pred+ $\approx$ 0.99), yielding widespread false positives and IoU near the all-positive baseline. In Caatinga-to-Amazon, Prithvi over-predicts (high recall, low precision), whereas U-Net under-predicts (Pred+ = 0.105 versus a 0.180 reference positive fraction; recall = 0.500), with substantial false-negative errors. Thus, source-only transfer causes class shifts that aggregate IoU alone can obscure. These contrasts show why IoU should be interpreted jointly with Pred+, precision, and recall.

Across seeds 42, 123, and 7, U-Net obtains Caatinga-to-Amazon IoU values of 0.612, 0.553, and 0.234 (Fig. 4); seed 7 has Pred+ = 0.057 and recall = 0.251, contributing to an IoU SD of 0.166. This run represents a strongly under-predicting regime, and the across-seed variation reflects both split and optimization variability.

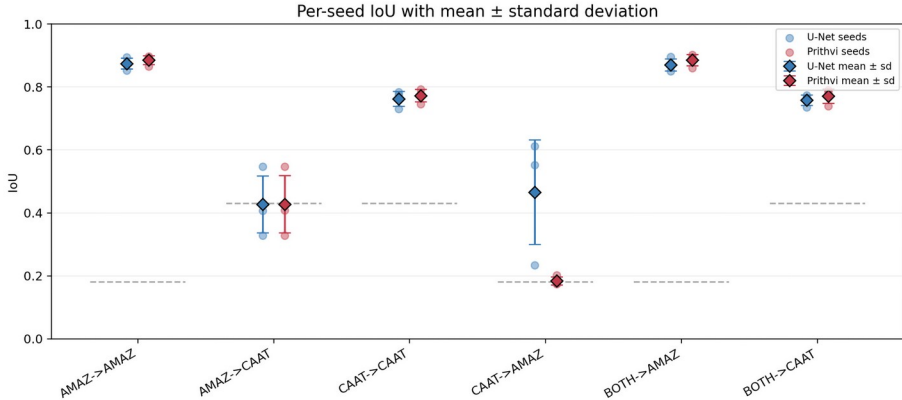


Fig. 4. Per-seed IoU distribution. Dots represent individual seeds; diamonds and error bars represent mean and standard deviation. Dashed segments indicate the all-positive baseline.

Table 3 reports the aggregate quantitative results across the three seeds for all train/test configurations.

Table 3. Aggregate quantitative results across three seeds. Prec., Recall, F1, IoU, and Pred+ are means of per-seed pixel-level micro-averaged metrics for the positive class. IoU SD denotes the standard deviation across the three per-seed IoU values. Base is the mean IoU of the all-positive predictor, which is equivalent to the mean positive-class prevalence among valid test pixels. Delta is the mean IoU minus the mean Base. Both denotes multi-biome training using samples from Amazon and Caatinga. Because metrics are computed separately for each seed before aggregation, F1 is not necessarily equal to the harmonic mean of the displayed mean Precision and Recall values.

Train	Test	Model	Prec.	Recall	F1	IoU	IoU SD	Pred+	Base	Delta
amazon	amazon	U-Net	0.937	0.929	0.933	0.875	0.017	0.178	0.180	+0.694
amazon	amazon	Prithvi	0.934	0.945	0.939	0.885	0.014	0.182	0.180	+0.705
caatinga	caatinga	U-Net	0.882	0.850	0.865	0.763	0.024	0.413	0.430	+0.333
caatinga	caatinga	Prithvi	0.870	0.874	0.872	0.773	0.020	0.433	0.430	+0.342
amazon	caatinga	U-Net	0.430	0.991	0.594	0.428	0.091	0.993	0.430	-0.002
amazon	caatinga	Prithvi	0.430	0.993	0.594	0.428	0.091	0.995	0.430	-0.002
caatinga	amazon	U-Net	0.857	0.500	0.617	0.466	0.166	0.105	0.180	+0.286
caatinga	amazon	Prithvi	0.186	0.971	0.311	0.184	0.012	0.936	0.180	+0.004
both	amazon	U-Net	0.938	0.924	0.931	0.871	0.019	0.177	0.180	+0.690
both	amazon	Prithvi	0.937	0.942	0.939	0.886	0.018	0.181	0.180	+0.705
both	caatinga	U-Net	0.875	0.852	0.863	0.759	0.016	0.415	0.430	+0.329
both	caatinga	Prithvi	0.876	0.865	0.871	0.771	0.023	0.425	0.430	+0.341

4.5 Qualitative Error Analysis

The qualitative outputs confirm the quantitative collapse analysis. In the Amazon-to-Caatinga Prithvi example, the prediction becomes almost entirely positive, while the PRODES reference contains a mixture of native vegetation, accumulated deforestation, and ignored pixels. The resulting error map is dominated by false positives. This visual pattern is consistent with the predicted positive fraction of 0.995 and the IoU being near the all-positive baseline. A representative Amazon-to-Caatinga Prithvi example is shown in Fig. 5.



Fig. 5. Qualitative example of Amazon-to-Caatinga transfer failure for Prithvi-EO. The prediction collapses toward the positive class, producing widespread false positives in the Caatinga target region. In the error map, white = true positive, black = true negative, orange = false positive, blue = false negative, and yellow = ignored.

When Caatinga samples are included during multi-biome training, the same target chip no longer yields a degenerate all-positive mask. In the Both-to-Caatinga Prithvi example, the prediction becomes spatially structured and better aligned with the PRODES reference, although boundary errors and localized false positives and false

negatives remain. This comparison indicates that the Caatinga target is learnable and that the main failure in Amazon-to-Caatinga transfer is associated with source-only domain shift rather than an intrinsic inability of the architecture to represent the target patterns. The corresponding Both-to-Caatinga result for the same chip is shown in Fig. 6.



Fig. 6. Qualitative example of Both-to-Caatinga segmentation for Prithvi-EO on the same Caatinga target chip. Including Caatinga samples during training mitigates the all-positive collapse observed in Amazon-to-Caatinga transfer and produces a spatially structured prediction, although boundary and localized classification errors remain. In the error map, white = true positive, black = true negative, orange = false positive, blue = false negative, and yellow = ignored.

4.6 Computational Cost

The computational measurements strengthen the cost-benefit interpretation. Averaged across configurations, Prithvi requires approximately 738.7 seconds per seed, compared with 79.6 seconds for U-Net. This is about 9.3 times slower. Prithvi also uses a mean peak GPU memory of approximately 6.46 GB, compared with 1.14 GB for U-Net. Thus, the foundation model is not only 13.26 times larger in trainable parameters, but also substantially more expensive to fine-tune and evaluate under this protocol. Figure 7 summarizes the computational cost by architecture.

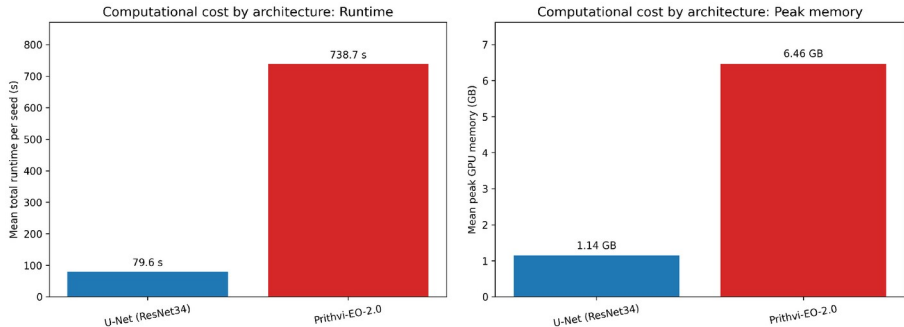


Fig. 7. Computational cost by architecture. Values are averages across train/test configurations and seeds.

5 Discussion

5.1 Domain Shift between Biomes

The main empirical finding is strong evidence of domain shift between the selected Amazon and Caatinga areas. Both models perform well in-domain but fail or degrade

sharply when transferred across biomes under source-only fine-tuning. The Amazon-to-Caatinga direction is particularly revealing because both models produce IoU values at the Caatinga all-positive baseline. High Amazon performance therefore does not imply Caatinga generalization, even with the same sensor, target class, chip size, and training schedule.

This result is plausible from a remote-sensing perspective. Dense tropical forest and semi-arid vegetation differ in canopy closure, soil exposure, spectral contrast, phenology, hydrology, and anthropized landscape texture. A decision boundary learned in one spectral-spatial context may not remain valid in another. The visual inspection of the Caatinga scene further shows strong internal heterogeneity, including urban areas, irrigated agriculture, exposed soil, and riverine structures.

5.2 Limits of Prithvi-EO under Source-Only Cross-Biome Transfer

Prithvi-EO performs slightly better than U-Net in the two in-domain conditions and in multi-biome training. Nevertheless, the larger model does not provide consistent source-only cross-biome robustness. In Amazon-to-Caatinga, both models collapse toward all-positive predictions. In Caatinga-to-Amazon, Prithvi performs near the all-positive baseline and substantially below U-Net. This result should not be interpreted as a universal limitation of all foundation models, but it does show that Prithvi-EO pretraining is not a substitute for target-domain validation, target-biome supervision, or explicit adaptation strategies.

The cost-benefit trade-off is important. Prithvi has 324.2 million trainable parameters, about 13.26 times more than U-Net, but its in-domain IoU gains are 0.010 in each biome. It is also about 9.3 times slower per seed on average and requires far more GPU memory. For operational monitoring, such a trade-off requires careful justification, especially when a lighter convolutional baseline achieves comparable or superior robustness in some transfer settings.

5.3 Multi-biome Training as a Viable Strategy

Multi-biome training is the most effective strategy tested in this work for mitigating cross-biome collapse. It prevents the collapse observed under source-only cross-biome transfer and maintains high performance in both test biomes. However, it does not substantially outperform target-biome training. This suggests that labeled target-biome examples are important under this protocol, and that a monitoring system intended to operate across biomes should include representative samples from each ecological domain.

5.4 Implications for Operational Monitoring

For operational environmental monitoring, the results suggest caution when deploying a model trained only in the Amazon to the Caatinga or vice versa. A system may appear accurate in one biome and still produce near-trivial predictions elsewhere. The all-positive baseline is therefore not just a diagnostic convenience; it is a useful operational safeguard. If a model barely exceeds this baseline, the resulting maps may be unreliable for decision-making even when absolute IoU appears moderate.

Operational systems should report predicted class prevalence in addition to IoU, precision, recall, and F1-score. Predicted positive fraction is easy to compute and can reveal domain-shift failures that aggregate metrics alone may obscure. Qualitative error

maps also help identify whether performance degradation arises from boundary uncertainty, localized confusion, or global class collapse.

5.5 Limitations

This study has several limitations. First, each biome is represented by one selected study area; therefore, the results should be interpreted as evidence from selected Amazon and Caatinga regions rather than definitive biome-wide behavior. Second, each area uses a single Sentinel-2 observation from 2023. This is especially relevant for the Caatinga, where seasonal variation can substantially change spectral appearance. Third, the spatial block split mitigates local leakage but does not replace external validation on fully independent regions.

Fourth, the task is accumulated deforestation segmentation, not annual change detection. Fifth, although qualitative prediction and error maps are included, they cover a limited number of representative chips and should be expanded in future work. Sixth, runtime and GPU memory are measured in a Kaggle Tesla T4 environment; broader computational benchmarking across hardware and implementation settings would be needed for general deployment conclusions. Seventh, because this study prioritizes a controlled comparison using Sentinel-2 L2A inputs for both architectures, Prithvi-EO-2.0 was not evaluated under its HLS-based pre-training input regime.

6 Conclusion and Future Work

This paper evaluated cross-biome semantic segmentation of accumulated deforested/anthropized areas in selected Amazon and Caatinga study areas. Both U-Net and Prithvi-EO-2.0-300M-TL achieved strong in-domain performance, whereas source-only cross-biome transfer showed unstable behavior and, in some configurations, collapsed toward all-positive predictions. Prithvi-EO-2.0 provided modest in-domain improvements over U-Net but did not consistently improve cross-biome robustness despite its substantially larger computational cost. Multi-biome training mitigated the observed collapse in the selected study areas, suggesting that exposure to samples from both domains can improve robustness under the evaluated protocol.

Future work should evaluate geographically independent areas, seasons, and years within each biome to disentangle biome-related differences from scene-specific, geographic, and seasonal variation. Where computationally feasible, repeated spatially blocked cross-validation with block-level uncertainty estimation should complement the current three-seed holdout protocol. Additional segmentation architectures, such as DeepLabV3+ and SegFormer, should be compared under the same preprocessing and evaluation protocol. Alternative Prithvi-EO adaptation strategies, including frozen-backbone training, partial fine-tuning, layer-wise learning rates, and few-shot target-biome adaptation, should also be evaluated; the latter constitutes a distinct setting because it assumes access to labelled target-biome samples.

Acknowledgments. The authors acknowledge the use of generative artificial intelligence tools during the preparation of this manuscript. These tools were used to support language revision, structural organization, and editorial refinement. The authors take full responsibility for the scientific content, methodological decisions, data analysis, interpretation of results, originality, plagiarism verification, and final version of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Szwarcman, D., Roy, S., Fraccaro, P., et al.: Prithvi-EO-2.0: A Versatile Multitemporal Foundation Model for Earth Observation Applications. *IEEE Trans. Geosci. Remote Sens.* 64, 4400120 (2026). <https://doi.org/10.1109/TGRS.2025.3642610>
2. Tian, Y., Zhao, F., Meng, R., Sun, R., Zhang, Y., Shen, Y., et al.: A vision foundation model-based method for large-scale forest disturbance mapping using time series Sentinel-1 SAR data. *Remote Sens. Environ.* 325, 114775 (2025). <https://doi.org/10.1016/j.rse.2025.114775>
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778. IEEE (2016). <https://doi.org/10.1109/CVPR.2016.90>
5. Wagner, F.H., Dalagnol, R., Carter, G., Hirye, M.C.M., Gill, S., Takoungoum, L.B.S., et al.: Wall-to-wall Amazon forest height mapping with Planet NICFI, Aerial LiDAR, and a U-Net regression model. *Remote Sens. Ecol. Conserv.* (2025). <https://doi.org/10.1002/rse2.70041>
6. IBM-NASA Geospatial: Prithvi-EO-2.0-300M-TL. Hugging Face. <https://huggingface.co/ibm-nasa-geospatial/Prithvi-EO-2.0-300M-TL> (Accessed: 24 Jun 2026).
7. Souza, C.M., Jr., Ferreira, B.G., Brandão, I.M., Rios, S., Aguilar-Brand, J., Schirmbeck, J., et al.: Enhanced Amazon Wetland Map with Multi-Source Remote Sensing Data. *Remote Sens.* 17, 3644 (2025). <https://doi.org/10.3390/rs17213644>
8. ESA, Copernicus: Sentinel-2 Level-2A multispectral imagery. Copernicus Data Space Ecosystem. <https://dataspace.copernicus.eu/explore-data/data-collections/sentinel-data/sentinel-2> (Accessed: 24 Jun 2026).
9. Microsoft: Planetary Computer. <https://planetarycomputer.microsoft.com/> (Accessed: 24 Jun 2026).
10. INPE: PRODES—Brazilian Deforestation Monitoring Programme. TerraBrasilis. PRODES Amazonia Legal 2025 v20260408; PRODES Caatinga 2024 v20260407. <https://terrabrasilis.dpi.inpe.br/> (Accessed: 24 Jun 2026).
11. Iakubovskii, P.: Segmentation Models PyTorch. GitHub (2019). https://github.com/qubvel/segmentation_models.pytorch (Accessed: 24 Jun 2026).
12. Gomes, C., Blumenstiel, B., de Sousa Almeida, J.L., et al.: TerraTorch: The Geospatial Foundation Models Toolkit. *arXiv:2503.20563* (2025). <https://doi.org/10.48550/arXiv.2503.20563>