
This is the **published version** of the bachelor thesis:

Stoiciu, Adrian. *Detecting deforestation and vegetation-cover change with deep learning on satellite imagery*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334459>

under the terms of the  license.

Detecció de la desforestació i els canvis de la coberta vegetal mitjançant l'anàlisi d'imatges de satèl·lit

Adrian Stoiciu, Grau d'Enginyeria Informàtica, Universitat Autònoma de Barcelona.
Tutor: Mehmet Oguz Mulayim. Curs 2025/2026.

Abstract—Resum—Aquest treball desenvolupa un detector bi-temporal de desforestació anual a l'Amazònia: una Attention Res-UNet amb un codificador ResNet34 preentrenat que rep dos anys consecutius d'imatge de satèl·lit (8 canals, any Y-1 i any Y) i prediu, píxel a píxel, on hi ha hagut pèrdua de bosc nova. S'entrena amb composicions anuals sense núvols del producte HLS de la NASA i amb la banda de pèrdua de Hansen com a etiqueta. Plantejar la tasca com a detecció de canvis, i no com a segmentació d'una sola imatge, és el canvi clau: sobre un conjunt de test honest i sense fuga de dades espacial, el model bi-temporal arriba a un F1 de 0,55 davant del 0,43 del model d'una sola imatge, amb un guany gran en precisió (0,54 a 0,71). La validació contra una referència independent (MapBiomass) mostra que el desacord restant és del mateix ordre que el desacord entre les pròpies referències.

Abstract—This work develops a bi-temporal detector of yearly deforestation in the Amazon: an Attention Res-UNet with an ImageNet-pretrained ResNet34 encoder that takes two consecutive years of satellite imagery (8 channels, year Y-1 and year Y) and predicts, per pixel, where new forest loss has occurred. It is trained on annual cloud-free composites from NASA's HLS product with the Hansen loss-year band as labels. Framing the task as change detection rather than single-image segmentation is the central contribution: on an honest, leakage-free test set the bi-temporal model reaches an F1 of 0.55, up from 0.43 for a single-image model, with a large precision gain (0.54 to 0.71). A cross-check against an independent reference (MapBiomass) shows the residual disagreement is at the level of the disagreement between the references themselves, so much of the error is label noise rather than model error.

Index Terms—deforestation, change detection, deep learning, semantic segmentation, remote sensing, Amazon rainforest, attention U-Net, bi-temporal model.

I. Introduction

THE Amazon rainforest is one of the planet's largest carbon stores and one of its most biodiverse regions, and it is being cleared every year. Monitoring this by hand is impossible: the area spans millions of square kilometres and new clearings appear continuously. Satellites give repeated, wide and consistent coverage, so the natural way to monitor deforestation is to process satellite imagery automatically, and deep learning, especially semantic segmentation, is well suited to turning that imagery into per-pixel maps of forest loss [1]. Two sensor families exist: optical, which I use, records reflected sunlight and is intuitive but blind through cloud; radar sees through cloud at the cost of a harder signal. The Amazon's persistent

cloud is the central difficulty, and how I handle it is what makes an optical-only approach viable.

Why it matters operationally is worth stating plainly. The Brazilian Legal Amazon alone loses on the order of ten thousand square kilometres in a bad year, and the official annual products that quantify it appear months after the year closes. A model that flags new loss from freely available imagery, soon after a year ends and at low cost, would be a useful early signal for the people and agencies who act on deforestation. That is the practical target I kept in mind.

The central idea, and what this report is built around, is that deforestation in a given year is a change, not a state. Forest present one year becomes bare ground the next; from a single image you cannot tell a clearing made this year from one made three years ago, because both look bare. The main system is therefore a bi-temporal model that sees the year before and the year of interest at once [4], [5], [9]. I reached this the hard way: an earlier single-image version stalled and, as I explain honestly in Sections III-B and IV, its results were also inflated by spatial data leakage. Diagnosing and correcting both is what made the bi-temporal result trustworthy, and that honest correction is the part of the project I learned the most from.

The specific objectives are: (i) a clean, reproducible acquisition, training and evaluation pipeline; (ii) an honest evaluation with no spatial leakage; (iii) the bi-temporal reformulation and a measurement of its effect; (iv) a cross-check against an independent reference to separate model error from reference noise; and (v) a 2021 to 2025 evolution study, including a forecast year. The contribution is less a new architecture than a rigorous, honest treatment of a deceptively hard task.

II. State of the art

The literature on deep learning for deforestation guided most of my decisions, and reading it carefully is also what eventually made me distrust my own early results. A systematic review by Jelas et al. [1] concludes that, for this task, the amount and diversity of the data matter more than the exact network, and that U-Net variants dominate. Eliason et al. [2] compare architectures head to head and find a plain U-Net beating heavier models such as DeepLabV3+, with the Attention U-Net the most

TABLE I

Selected related work. Reported scores are not comparable across different tasks and references; the last column says how each relates to this project.

Work	Task	Reported	Relation to this work
Jelas [1]	review	data > arch.	data-first strategy
Eliason [2]	forest/non-forest seg.	U-Net best	method precedent
John-Zhang [3]	deforestation seg.	attn. helps	supports attention
Sound. [4]	static forest map	IoU > 0.9	easier, balanced task
Sartor [5]	change det. (SAR+opt.)	tiling	non-overlap tiling
arXiv [9]	bi-temporal vs PRODES	F1 ≈ 0.80	realistic target
This work	annual-loss change det.	F1 0.55	honest, leakage-free

efficient; they name their narrow geographic coverage as their main limitation. John and Zhang [3] apply an attention-based U-Net specifically to deforestation, which directly supports my use of attention gates. Lee et al. [11] classify deforested landscapes, and ForestNet [12] classifies the drivers of loss; together they show the breadth of the field.

On the change-detection framing central here, Sartor et al. [5] compare consecutive years and fuse optical with SAR using non-overlapping tiling (the discipline I later adopted), and Soundararajan et al. [4] use a multi-temporal model; their reported IoU above 0.9 is on a static forest/non-forest benchmark, an easier and better-posed task than annual-loss attribution and not directly comparable. A 2025 Amazon study that frames the task as bi-temporal change detection against the PRODES reference reaches an F1 of about 0.80 [9], the strongest evidence that the bi-temporal formulation is the right direction and a realistic target. Table I places these works side by side; the point is that the numbers are not comparable unless the task is the same.

Two choices set this project apart and are, I believe, its main contributions as a TFG. First, the evaluation is made rigorously leakage-free with a geographically blocked split, after I discovered that a naive split had been inflating my numbers. Second, the residual error is interpreted against a second, independent reference rather than treating a single noisy reference as ground truth. High benchmark scores are not a fair yardstick: those benchmarks label whole scenes as forest or non-forest at one date, a balanced two-class problem, whereas annual-loss attribution has a rare positive class, a change target, and a reference that is itself uncertain about the year.

III. Data and methodology

This section describes the data and the methods, and the order in which the work happened, because the project

went through a deliberate reassessment. The plan in the initial report was to detect deforestation from imagery; the first phases produced a single-image Attention Res-UNet and preliminary metrics. Instead of adding features on top of an approach I had not fully understood, I stopped to audit the foundations, found the two problems below (an ill-posed single-image formulation and spatial data leakage), and the final phase was spent correcting them, enlarging the dataset, and building and validating the bi-temporal model.

A. Imagery, labels and dataset

All imagery is gathered through Google Earth Engine. I use two optical missions, Sentinel-2 (collection HLSS30) and Landsat 8/9 (HLSL30), both from NASA’s Harmonized Landsat Sentinel-2 (HLS) product, which co-registers and cross-calibrates the two so they can be treated as one stream. I take four bands each (red, green, blue and near-infrared) at 30 m. The near-infrared (NIR) band is the most informative: healthy vegetation reflects it strongly while bare ground does not, so forest and loss separate clearly in NIR and in the derived vegetation index far more than in the visible bands (Fig. 1). I use the Normalized Difference Vegetation Index, $NDVI = (NIR - Red)/(NIR + Red)$, which is near one over dense canopy and drops sharply over cleared ground; the bi-temporal model is, in effect, learning to find where NDVI has fallen between two years.

Because the Amazon is cloudy most of the year, a single date is rarely usable. I read the Fmask quality band that ships with HLS, mask any pixel flagged as cloud, adjacent cloud, cloud shadow or cirrus, and then take the per-pixel median of all the remaining clear scenes of a year. The median is robust to the occasional missed cloud, and fusing the two missions roughly doubles the clear looks available, which together yield a mostly cloud-free annual composite and let the project stay optical-only.

The labels come from the Hansen Global Forest Change loss-year band [6], which gives, for every 30 m pixel, the year of clearing. It is the standard global product, but it is important to be clear about what it is: Hansen is itself the output of a model, a bagged decision-tree classifier trained on Landsat time-series metrics, validated on about 1,500 reference points to roughly 87% accuracy, not a field measurement [6]. Independent assessments document real biases, especially in the tropics. A study dedicated to the Amazon finds Global Forest Change overestimates deforestation relative to careful supervised maps (by about 55% at a Brazilian site and roughly twofold at a Peruvian one), largely through mixed pixels and regenerating secondary forest, and stresses that comparing it to another map is model-to-model, not ground-truth, verification [13]. A formal comment in Science had already questioned its accuracy for conflating forest types [14]. This is exactly why I do not treat Hansen as truth and cross-check against an independent reference in Section IV. I use it nonetheless because it is the only free, global, annual product with a per-pixel year.

Why near-infrared matters: forest is bright in NIR and high in NDVI; cleared ground is dark and low

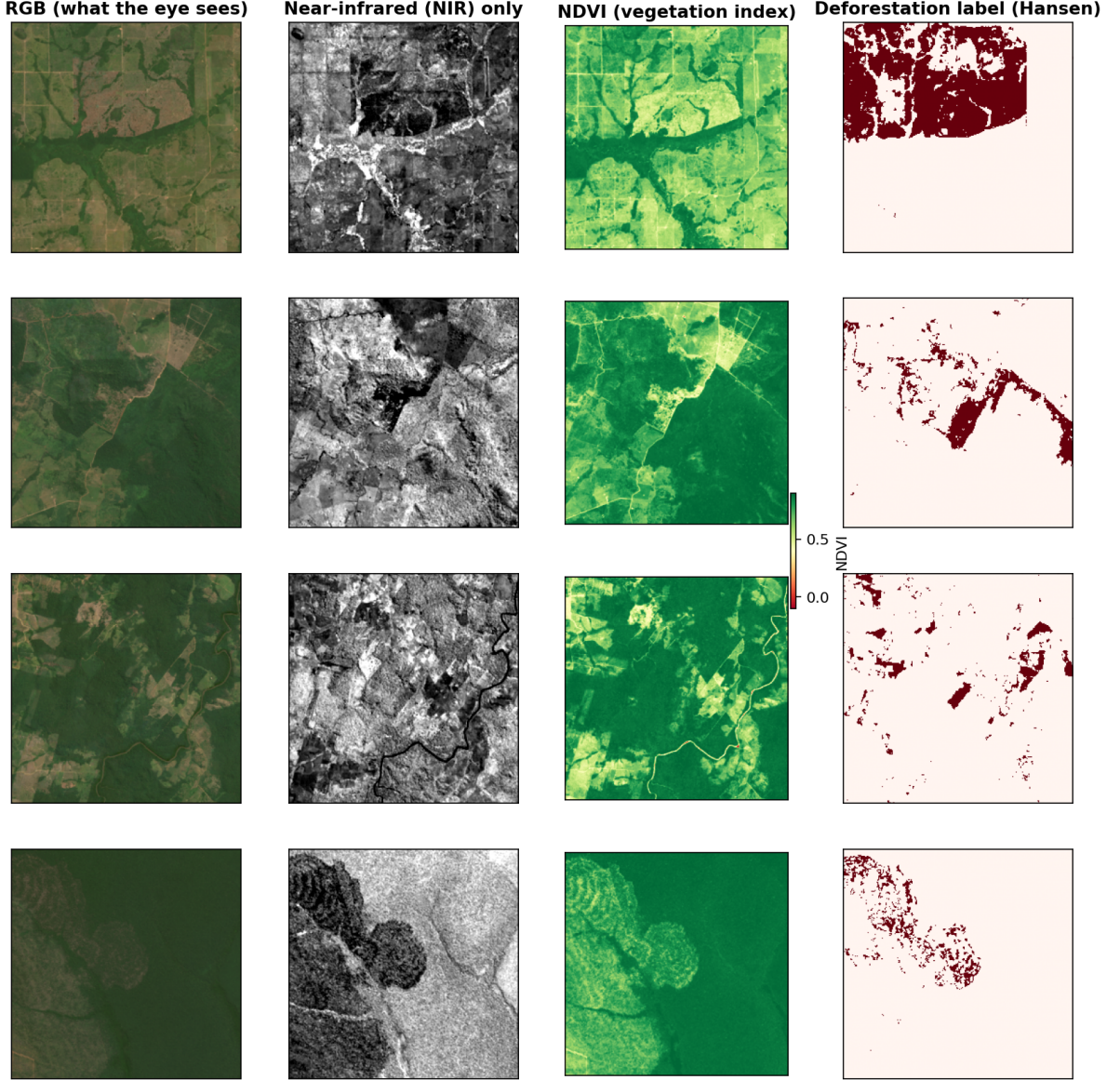


Fig. 1. Why near-infrared matters, on real patches. Left to right: the RGB image, the near-infrared band alone, the NDVI, and the Hansen label. Cleared ground is barely distinct in RGB but clearly separable in NIR and NDVI.

Each sample is a 256 by 256 patch (about 7.7 km on a side) sampled in the Amazon arc of deforestation, where the characteristic fishbone road pattern appears. I deliberately mix patch categories (Fig. 2): heavy, moderate and original low-loss patches, and uniformly-random patches with rivers, sandbanks and shadows, so the model also learns what is not loss. The final corpus is 3,100 patches: about 1,300 high-loss, 400 moderate, 300 original low-loss and 1,100 random. Even with the hotspot sampling, counted per pixel over the whole corpus, 95.6% of pixels are intact forest and only 4.4% are deforestation (Fig. 3), which is why the loss function, the attention gates, and the

use of IoU and F1 instead of accuracy are all necessary.

A fast acquisition engine made the multi-year, overlap-free dataset feasible. Candidates are seeded on Hansen loss pixels, a whole batch is screened with a single server-side reduction rather than one call per patch, accepted patches are exported in parallel, and any candidate overlapping an existing one is rejected by a per-axis (Chebyshev) test, so the corpus is overlap-free by construction. This turned days of sequential downloads into hours and grew the corpus from 1,300 to 3,100 non-overlapping patches spanning 2022 to 2024 (Fig. 4). Before the model, each patch is normalised by a fixed reflectance scale and clip,

Stratified dataset: example patches per sampling category (deforestation label overlaid in red)



Fig. 2. Example patches per sampling category, with the Hansen deforestation label overlaid in red.

and during training only it is lightly augmented with flips, 90-degree rotations and small brightness jitter, applied identically to both dates so no fake change is invented.

B. The spatial data-leakage problem

Before the bi-temporal work, I had to fix a flaw in how the first model was evaluated. Because patches were sampled at random coordinates and the hotspot sampling revisited the same places, many patches shared ground, and a random train/test split then put near-

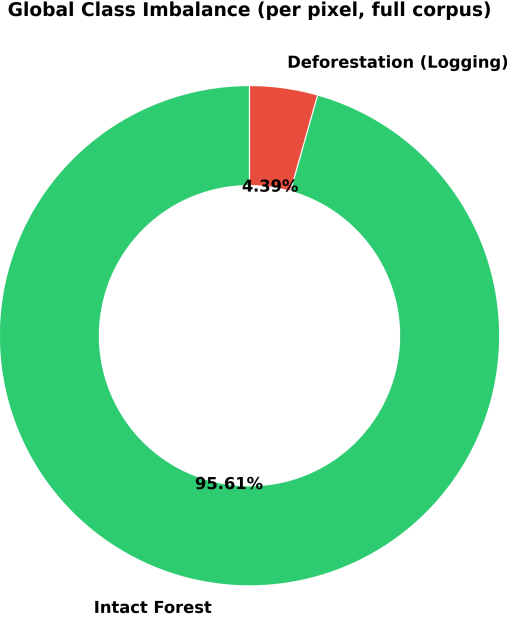


Fig. 3. Class imbalance, measured per pixel over the whole corpus: 95.6% intact forest against 4.4% deforestation.

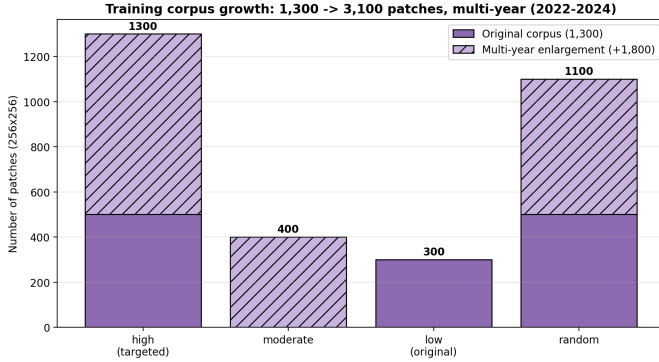


Fig. 4. Corpus growth from 1,300 to 3,100 non-overlapping patches.

TABLE II

The same single-image model on overlapping (contaminated) vs non-overlapping (clean) test patches: a +0.21 F1 inflation purely from leakage.

Test subset	IoU	F1	P	R
Contaminated	0.360	0.529	0.606	0.470
Clean	0.191	0.320	0.362	0.287

duplicate patches on both sides of the boundary. The model was rewarded at test time for recognising places it had trained on, a failure known as spatial autocorrelation leakage. Recovering every patch’s footprint and evaluating the same single-image model on overlapping versus non-overlapping test patches made the size of the problem concrete: 0.529 F1 on the contaminated half against only 0.320 on the clean half, a +0.21 gap (Table II, Fig. 5).

I removed it with a spatially-blocked split, following

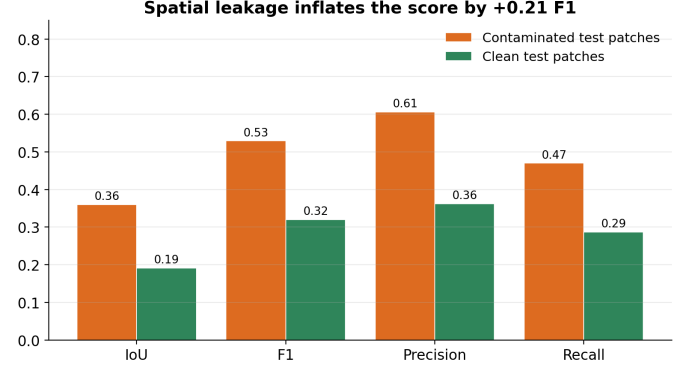


Fig. 5. Metric inflation from spatial leakage, on every metric.

the non-overlapping tiling of Sartor et al. [5]: the area is divided into 0.5 degree cells (about 55 km), whole cells go to train, validation or test, and any evaluation patch still touching a training patch across a cell border is dropped. This yields 2,480 training, 305 validation and 297 test patches, and every number reported from here on uses this honest split. The lesson, which shaped the rest of the work, was to distrust a number that looks too good and to measure leakage rather than assume its absence.

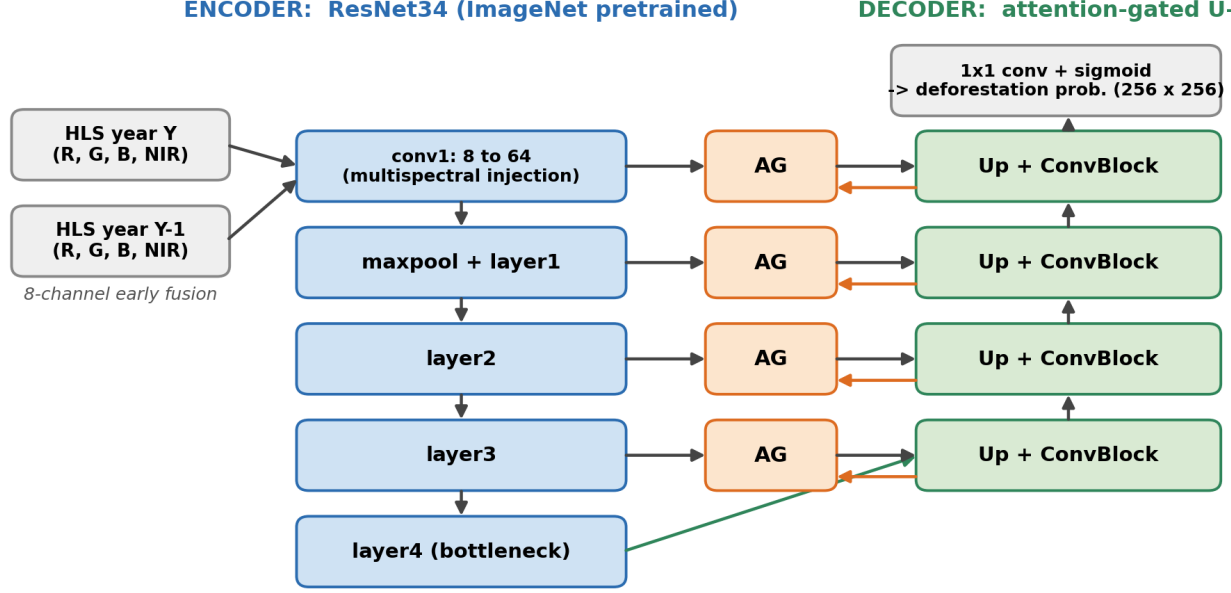
C. The bi-temporal model

The model is an Attention Res-UNet (Fig. 6). The encoder is a ResNet34 pretrained on ImageNet, which gives strong general-purpose edge and texture filters for free; the decoder is a U-Net decoder that upsamples back to full resolution, with an attention gate on every skip connection. Each gate computes, from the deeper decoder signal, a soft 0-to-1 map and multiplies the encoder features by it before concatenation, which pushes the dominant intact-forest background toward zero and lets the few loss pixels through. The encoder goes from the 256 by 256 input to an 8 by 8 by 512 bottleneck and the decoder mirrors it back; the network has about 24 million parameters, almost all in the pretrained encoder.

To make the network a true change detector, I feed it both years: for each patch I acquire the year Y-1 HLS composite over the identical footprint and stack it under the year Y composite, forming an eight-channel input. This is early fusion: I let one network learn the comparison, rather than running each year through a shared Siamese encoder and differencing the features. Early fusion is simpler, keeps the multispectral injection straightforward, and lets the pretrained encoder see both dates at once. The pretrained first convolution expects three channels, so I rebuild it for eight and initialise it from ImageNet, copying the RGB filters for each date and setting the NIR filter to their mean, so no channel starts blind. The before/after signal the network exploits is visible as a drop in NDVI between the two years (Fig. 7).

One more piece is needed to make the labels honest: a prior-year ignore mask. A pixel cleared in the three years before Y is correctly “no change in Y” but is visually bare,

Bi-temporal Attention Res-UNet



Attention gates (AG) use the deeper decoder signal to suppress the intact-forest background in each skip before concatenation.
Phase 1: encoder frozen, decoder trained. Phase 2: whole network fine-tuned.

Fig. 6. The bi-temporal Attention Res-UNet. The eight-channel input (year Y-1 stacked under year Y) feeds the pretrained ResNet34 encoder; an attention-gated decoder reconstructs the loss probability, with an attention gate on every skip connection.

so without care the model is punished for not flagging it. I download the full Hansen loss-year band per patch and exclude such already-cleared pixels, and any no-data pixel, from both the loss and the metrics, which also absorbs Hansen’s plus or minus one-year timing noise. Getting the two years pixel-aligned is essential: I reconstruct the year Y-1 export from each patch’s stored centre, exactly as year Y was exported, so the two composites share one grid; a one-pixel misalignment would inject spurious change everywhere.

D. Loss, training and evaluation

Because the positive class is rare, plain cross-entropy is minimised by predicting forest everywhere, so I use an equal combination of two losses that target the imbalance directly. Focal loss down-weights easy, confident pixels through a factor $(1 - p)^\gamma$ (with $\gamma = 2$ and a positive-class weight $\alpha = 0.8$) and so concentrates learning on the rare hard pixels; Dice loss optimises region overlap directly, which is close to optimising F1. Training is in two phases (Fig. 8): phase 1 freezes the encoder and trains only the decoder (learning rate 10^{-4}) so the randomly initialised decoder cannot corrupt the good pretrained features; phase 2 fine-tunes the whole network gently (10^{-5}). The optimizer is Adam, the batch size 8, and the checkpoint with the best validation loss is kept. Table III

TABLE III
Main training settings, shared by both models.

Setting	Value
Encoder	ResNet34 (ImageNet)
Loss	Focal + Dice (equal)
Optimizer	Adam
Learning rate phase 1 / 2	10^{-4} / 10^{-5}
Epochs phase 1 / 2	60 / 40
Batch size	8

lists the settings, identical between the two models for a fair comparison.

For evaluation I report IoU, F1, precision and recall, not accuracy (which already scores about 96% by predicting forest everywhere), pooled over all valid test pixels (micro). The valid mask is applied consistently in the loss and in every metric, so a prediction on an already-bare prior-year clearing is neither rewarded nor punished, which is what lets the precision figure reflect genuine new-loss detection. The decision threshold is chosen once on validation and applied unchanged to the test set; the F1 curve is shallow near its optimum (it changes by less than 0.01 between 0.5 and 0.6), so I report the validation-optimal 0.6 and note that the result is not sensitive to this choice. The project is a small Python package with

The before/after signal: deforestation appears as a drop in NDVI between the two years (red), which a single image cannot show

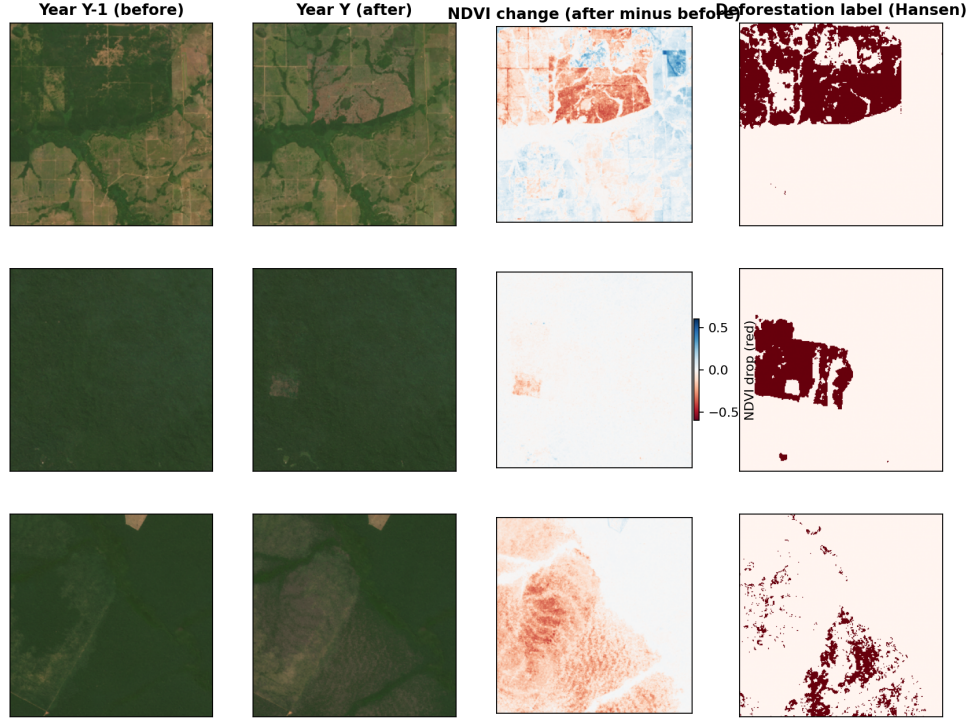


Fig. 7. Year Y-1, year Y, and the NDVI change on real patches. Deforestation appears as a drop in NDVI (red), the signal the bi-temporal model exploits.

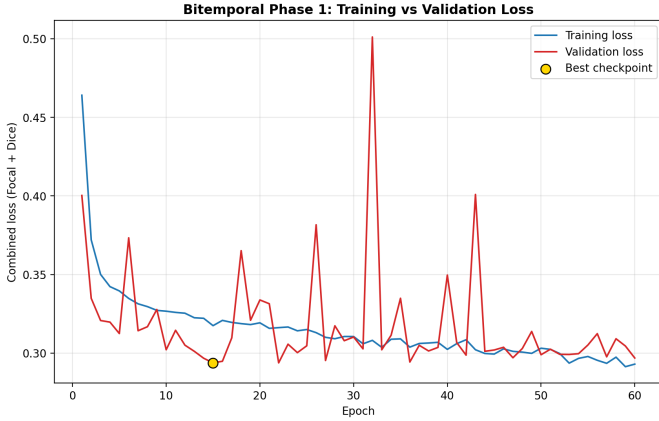


Fig. 8. Bi-temporal phase-1 training and validation loss; validation tracks training closely, which indicates the model is not overfitting.

a single configuration module as the source of truth, plus numbered command-line scripts, one per pipeline stage, so every figure and table here is regenerable with one command.

IV. Results and discussion

All metrics below are on the leakage-free, geographically disjoint test set of 297 patches, and all were reproduced from the code for this report.

TABLE IV

Single-image vs bi-temporal on the honest test set (micro metrics).

Model	IoU	F1	P	R
Single-image (4ch)	0.272	0.428	0.539	0.354
Bi-temporal (8ch)	0.376	0.547	0.710	0.445

A. Bi-temporal vs single-image

Feeding the model both years improves every metric, with a large gain in precision, exactly as expected: with two dates the model stops flagging already-bare ground as new loss. The single-image model reaches F1 0.428; the bi-temporal model reaches F1 0.547, a +0.12 improvement, with precision rising from 0.539 to 0.710 (Table IV, Fig. 9). This is the single largest gain in the project, and it comes from re-formulating the task rather than changing the network. The mechanism is the heart of the project: a single-image model sees bare ground and cannot tell whether it was cleared this year or years ago, so to obtain recall it must flag much of the bare ground, which costs precision; the bi-temporal model sees that the same ground was already bare the year before and leaves it alone.

To be honest about where the model stands, Fig. 10 shows its best and worst cases on unseen test patches. It is precise on large, geometric clearings (top, F1 near 1.0),

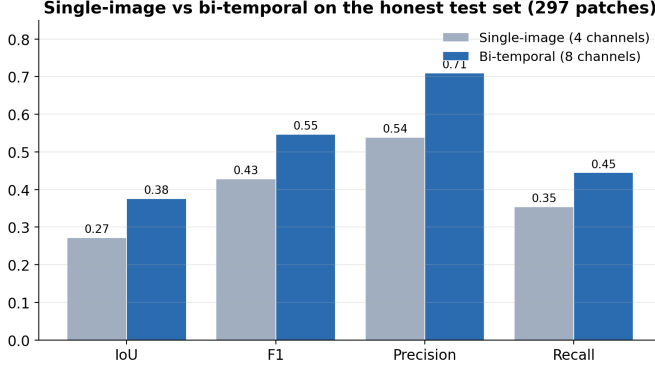


Fig. 9. Single-image vs bi-temporal: a gain on every metric, largest in precision.

TABLE V

Three-way agreement (micro F1 / IoU). The references agree with each other less than the model agrees with either.

Pairing	F1	IoU
model vs Hansen	0.504	0.337
model vs MapBiomass	0.533	0.363
Hansen vs MapBiomass	0.488	0.323

tracing the blocky pasture and crop fields almost exactly, and it fails on small, scattered, irregular disturbances (bottom, F1 near 0.0), which are also the cases where the Hansen reference is least certain about the exact pixel and year. The visual errors and the metric gap therefore share a root cause, and showing both ends openly is more useful than a single average.

B. How much of the error is reference noise?

Because Hansen is a model and not ground truth, part of the apparent error is the reference being wrong rather than the model. To separate the two I derived an independent loss map from MapBiomass [7] (forest in year $Y-1$ and not forest in year Y) and compared three pairwise agreements over all 297 test patches (Table V, Fig. 11). The model agrees with Hansen at F1 0.504 and with MapBiomass at 0.533, while the two independent references agree with each other at only 0.488. The model is therefore at least as consistent with each reference as the references are with one another, and it agrees most with MapBiomass, which it never trained on. The residual disagreement is dominated by reference noise and by definitional differences, Hansen measuring tree-canopy loss and MapBiomass a land-use transition, not by model error. This sets a ceiling: even a perfect detector could not reach an agreement of one against an imperfect reference.

C. Five-year evolution and forecast, 2021–2025

The operational point of the project is to estimate deforestation year by year, including a year with no reference yet, so I applied the bi-temporal model across a five-year span at three study sites. Each year’s prediction is

TABLE VI

Newly deforested area per year, summed over the three evolution sites (hectares). 2025 has no Hansen reference.

Year	Model (ha)	Hansen (ha)
2022	336	309
2023	343	86
2024	1039	1045
2025 (forecast)	672	n/a

TABLE VII

Cumulative 2022–2024 agreement with Hansen (F1), by site, before and after the bi-temporal re-formulation.

Site	Single-image	Bi-temporal
A	0.14	0.64
B	0.69	0.82
C	0.31	0.51

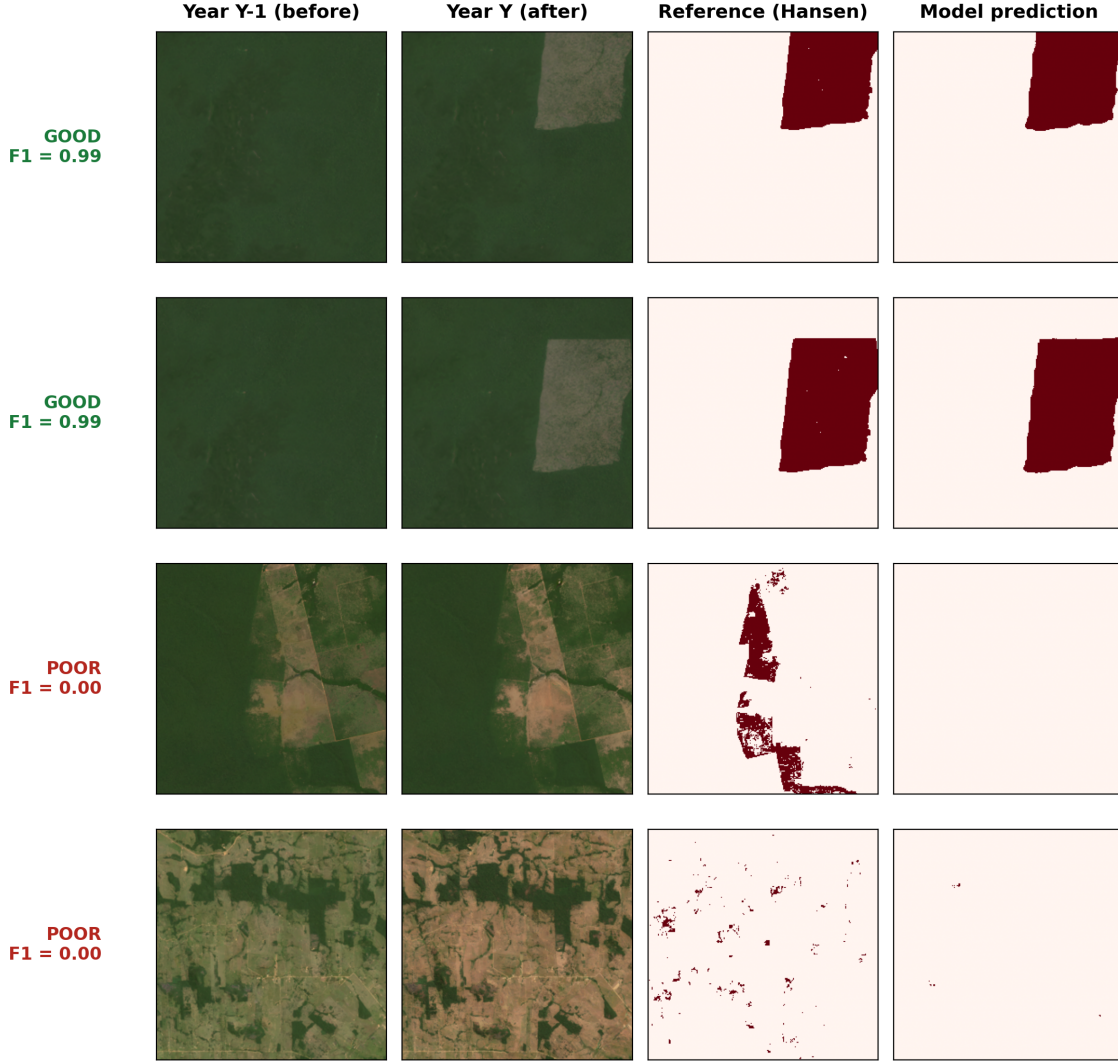
a direct change detection, model(year Y , year $Y-1$), and the per-year detections accumulate into a running map. To be transparent about what is checkable, Fig. 12 lays out, for one site, the satellite image, the model’s cumulative prediction, and the Hansen reference, one column per year: 2021 to 2024 can be compared directly against the reference (the model tracks it closely, for example 1,318 ha predicted against 1,364 ha reference by 2024), while 2025 has no Hansen reference and is shown clearly as the model’s forecast, continuing the trend rather than inventing a discontinuity.

Aggregated over the three evolution sites, the newly detected area per year tracks the reference where it can be checked (Table VI). Both the model and Hansen place the clear bulk of the loss in 2024 (about 1,040 ha by either count) and a smaller amount in 2022; the one honest discrepancy is 2023, where the model detects more than Hansen records, a year the reference treats as very quiet. The 2025 row is the model’s forecast. At the per-site level, agreement with Hansen on the cumulative 2022 to 2024 map improved sharply once the model was bi-temporal (Table VII), from as low as 0.14 to 0.64 at the hardest site, and the same picture holds across sites (Fig. 13). This is not proof the 2025 map is correct, but it is consistent with the validated years, and consistency is the most one can ask for in the absence of ground truth.

D. Error analysis and reproducibility

Per-patch F1 is strongly bimodal (Fig. 14): most patches score close to 1.0 or close to 0.0, with few in between. The high scores are large geometric clearings; the low scores are small, thin or scattered loss. Because the area-weighted majority of deforestation is in the large clearings, the model captures most of the hectares actually cleared even though its per-patch F1 averages lower than the micro figure. The failures concentrate on sub-hectare and linear disturbances below the effective resolution at which a 30 m composite and a noisy annual label can agree.

Bi-temporal model on unseen test patches: best and worst cases



The model is precise on large, clear clearings (top) and misses small, scattered disturbances (bottom), where the Hansen reference is also least certain.

Fig. 10. Bi-temporal model on unseen test patches: best (top) and worst (bottom) cases. Columns: year Y-1, year Y, Hansen reference, model prediction.

Reproducibility was treated as part of the engineering. A pytest suite of sixteen tests covers the utilities, the metric formulas, the losses, the model and the data pipeline, and a final test re-runs the trained model on the honest test set and confirms it still scores F1 0.55 (Fig. 15). A one-command script re-scores both models and prints a table comparing the reproduced metrics against the reported ones, and a self-contained kit reproduces the exact headline numbers offline.

Is an F1 of 0.55 good? For this task, I argue yes. The very high values in some papers are for static forest/non-forest mapping on balanced benchmarks [2], [3], [11], an easier problem; the realistic target for annual-loss change detection is around 0.80 [9], and a large part of the remaining gap is that the references themselves agree only

about 0.49 with each other. The main limitations, none hidden in the numbers I report, are the reliance on labels that are themselves models, the single training region, and recall on small clearings.

V. Conclusions and future work

I set out to detect yearly deforestation in the Amazon and to track its evolution, and the result I am most confident about is the bi-temporal model evaluated honestly. The most valuable work was not adding model capacity but fixing the foundations: defining the problem as change detection, removing the spatial leakage, and enlarging the dataset. With those in place, the bi-temporal reformulation raised the honest F1 from 0.43 to 0.55 with a large precision gain, the MapBiomass cross-check showed

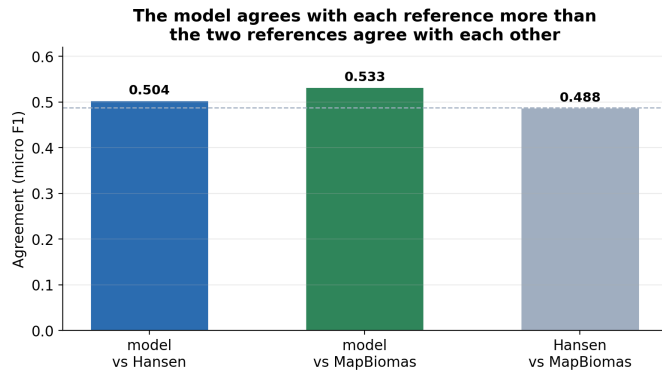


Fig. 11. Three-way agreement: the two independent references agree with each other less than the model agrees with either of them.

most of the residual disagreement is reference noise, and the model produced a coherent 2021 to 2025 evolution including a forecast year. The single most useful lesson was methodological: my early results looked good largely because the evaluation was flawed, and learning to distrust them changed how I will approach any evaluation. I would rather submit that honest 0.55, with its failures shown openly, than a higher number I could not defend.

The clearest next steps follow from the limitations: validate against PRODES and DETER [8] and a small set of human-checked examples; enlarge and diversify the dataset across regions and years [1], [2]; try a Siamese change-detection architecture [9]; and add Sentinel-1 SAR for the cloudiest periods [5], [10], which would fill exactly the gaps optical-only leaves. Nothing in the method is specific to the Amazon: the pipeline takes any area and any pair of years and the model retrains in hours.

Use of generative AI

Generative AI tools (Claude) were used as a support tool, in a directed and supervised way, and are declared here in full, consistent with my progress reports. I used them to help organise the literature, to assist with implementing the acquisition, training, evaluation and analysis scripts to my specifications, to help run and document the spatial-leakage audit, and to improve the writing, clarity and formatting of this report. The definition of the problem, the experimental design, the key decisions and the interpretation of the results are my own. I have reviewed and validated every output and take full responsibility for the content.

Acknowledgements

I thank my tutor, Mehmet Oguz Mulayim, for his guidance throughout the project and for the feedback that led me to look again at the foundations of the work.

References

- [1] I. Md Jelas, M. A. Zulkifley, M. Abdullah, and M. Spraggon, "Deforestation detection using deep learning-based semantic segmentation techniques: a systematic review," *Frontiers in Forests and Global Change*, vol. 7, 1300060, 2024.

- [2] J. Eliason, R. Holubec, and C. Mosk, "Comparing Deep Learning Approaches to Image Segmentation for Monitoring Deforestation," 2023.
- [3] D. John and C. Zhang, "An attention-based U-Net for detecting deforestation within satellite sensor imagery," *Int. J. Applied Earth Observation and Geoinformation*, vol. 107, 102685, 2022.
- [4] J. Soundararajan, A. Kalukin, J. Malof, and D. Xu, "Deep Learning-Driven Multi-Temporal Detection for Deforestation and Forest Fire Detection," *Remote Sensing*, vol. 17, no. 14, 2333, 2025.
- [5] G. Sartor, M. Salis, S. Pinardi, O. Saracik, and R. Meo, "Deep Learning tools to support deforestation monitoring in the Ivory Coast using SAR and Optical satellite imagery," *arXiv:2409.11186*, 2024.
- [6] M. C. Hansen et al., "High-Resolution Global Maps of 21st-Century Forest Cover Change," *Science*, vol. 342, pp. 850–853, 2013.
- [7] C. M. Souza et al., "MapBiomass Collection 10 Land Use and Land Cover (Brazil)," Google Earth Engine asset, mapbiomas-public.
- [8] INPE, "PRODES and DETER deforestation monitoring of the Brazilian Legal Amazon," TerraBrasilis portal.
- [9] "Identification of Deforestation Areas in the Amazon Rainforest using Change Detection Models," *arXiv:2512.08075*, 2025.
- [10] J. N. Hansen, E. T. A. Mitchard, and S. King, "Assessing Forest/Non-Forest Separability Using Sentinel-1 C-Band Synthetic Aperture Radar," *Remote Sensing*, vol. 12, no. 11, 1899, 2020.
- [11] S.-H. Lee et al., "Classification of Landscape Affected by Deforestation Using High-Resolution Remote Sensing Data and Deep-Learning Techniques," *Remote Sensing*, vol. 12, no. 20, 3372, 2020.
- [12] J. Irvin et al., "ForestNet: Classifying Drivers of Deforestation in Indonesia using Deep Learning on Satellite Imagery," *NeurIPS 2020 Workshop*, *arXiv:2011.05479*, 2020.
- [13] E. Kinnebrew et al., "Biases and limitations of Global Forest Change and author-generated land cover maps in detecting deforestation in the Amazon," *PLoS ONE*, vol. 17, no. 7, e0268970, 2022.
- [14] R. Tropek et al., "Comment on 'High-resolution global maps of 21st-century forest cover change'," *Science*, vol. 344, no. 6187, 981, 2014.

Appendix A

Finer points and peculiar findings

Reproducibility. A one-command script (`00_reproduce.py`) re-scores both models on the honest test set and prints a reproduced-vs-reported table; on the run for this report it printed OK for all eight figures. A self-contained kit (the 297 test patches plus the trained weights) reproduces the exact headline numbers offline, without the full dataset or an Earth Engine account.

Why the bi-temporal precision is trustworthy. The prior-year ignore mask must be applied at evaluation, not only at training. When I first packaged the reproduction kit without passing the per-patch year, the mask was skipped and the bi-temporal precision dropped from 0.71 to 0.57 while recall stayed the same, because already-bare old clearings were wrongly counted as false positives. Restoring the mask recovers 0.71 exactly.

A latent year-labelling bug. During the bi-temporal work a consistency check uncovered that part of the original corpus had the wrong target year recorded (a 2023 loss map labelled as 2024), which would have corrupted the prior-year mask. It was only findable because every patch stores its coordinates and year as metadata; I verified the fix by checking that around 99% of the positive pixels of each subset fall on its recorded year.

Five-year deforestation prediction at study site 1 (2021-2025)

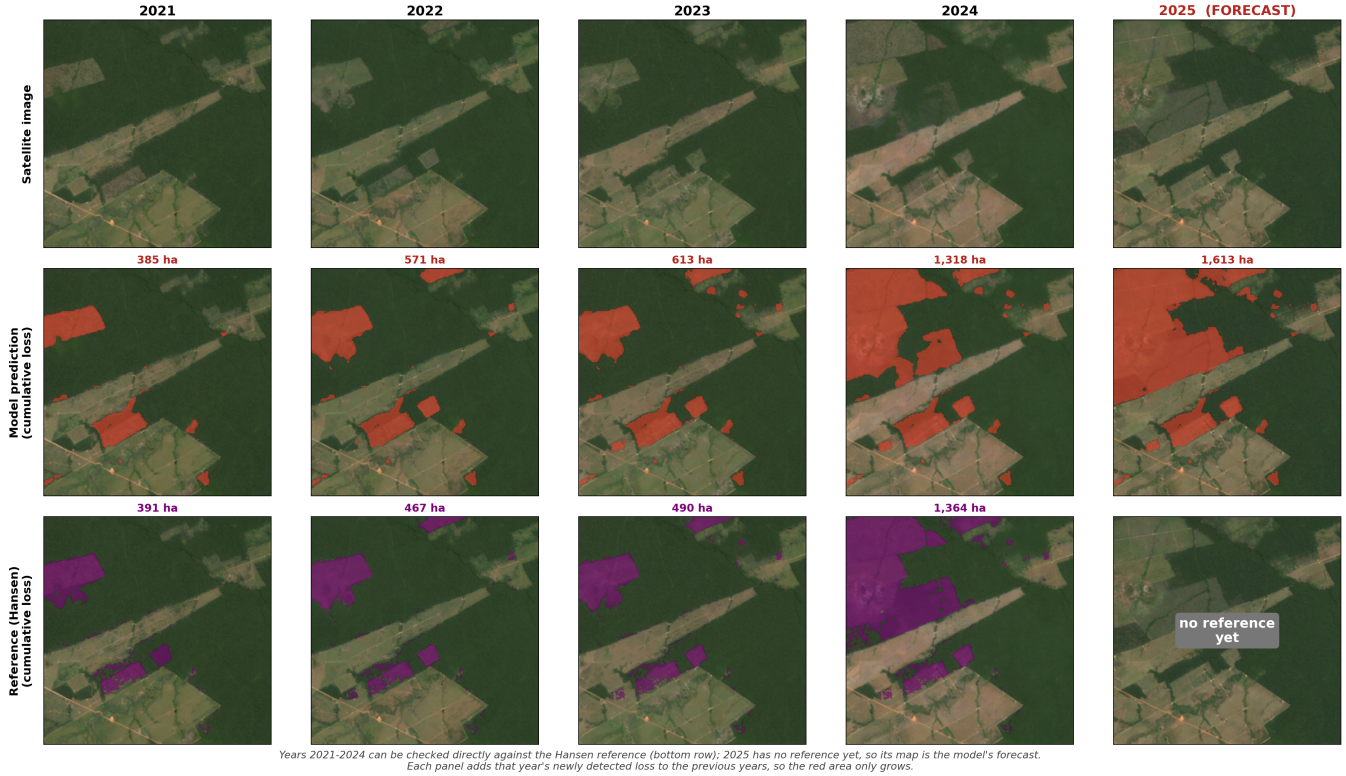


Fig. 12. Transparent five-year prediction at a study site (2021–2025). Top: satellite image. Middle: model cumulative prediction. Bottom: Hansen reference. 2021–2024 are checkable against Hansen; 2025 has no reference and is the model's forecast.

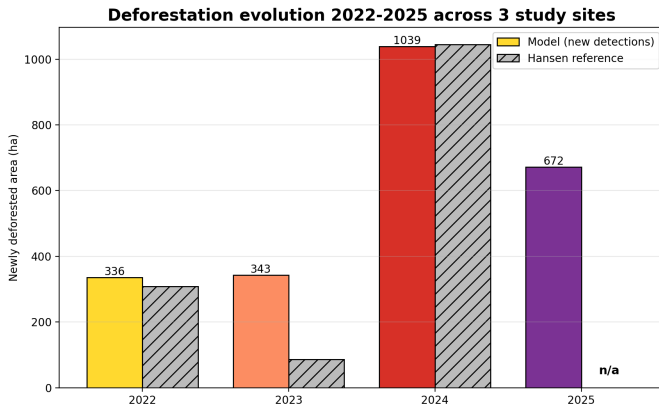


Fig. 13. Newly deforested area per year across the three sites (model vs Hansen). 2025 has no reference and is model-only.

Micro versus per-patch metrics. I report micro metrics, pooling every valid test pixel before computing the score once, which is sound for a 96/4 imbalance. A per-patch average is far more pessimistic (about 0.47, and only 0.28 for the second-reference agreements) because the many small, near-empty patches dominate it. The gap is a property of the aggregation, not of the model.

The HLS fusion synergy. Sentinel-2 and Landsat each revisit a point every few days, but clouds remove most passes; combining them roughly doubles the clear obser-

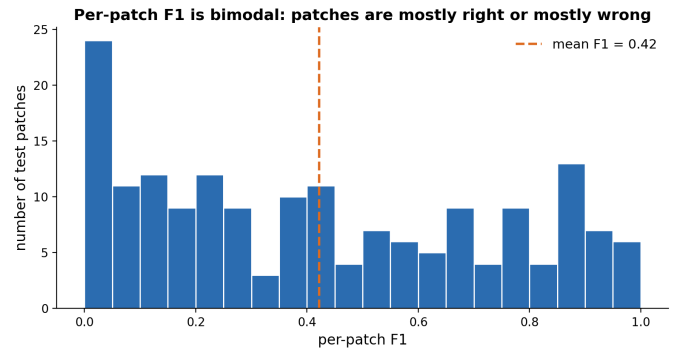


Fig. 14. Distribution of per-patch F1: bimodal, so a patch tends to be either almost entirely captured or almost entirely missed.

ventions feeding the annual median, which in the cloudiest months is the difference between a usable composite and a gap. This is the single reason the project can stay optical-only without radar.

Compute and footprint. The whole project runs on free data and a single consumer GPU. Each model trains in a couple of hours and the acquisition runs on Earth Engine's servers, so the monetary and carbon cost are small; inference on a new year is seconds per site, which is what makes ongoing monitoring cheap.

A statistical caveat. The honest test set is 297 patches, enough to separate the two models clearly, but the per-site

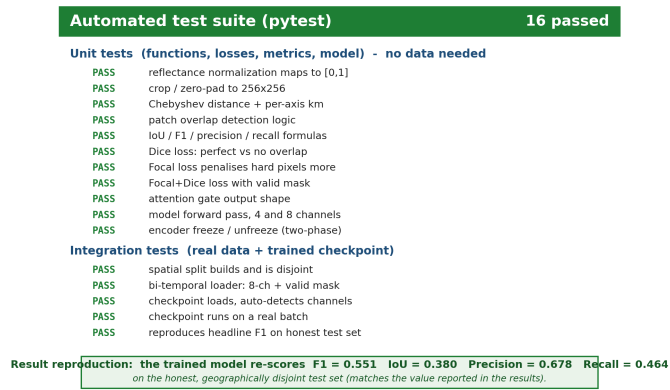


Fig. 15. The automated test suite: sixteen tests pass and the reproduction test re-scores F1 0.55 on the honest test set.

evolution numbers, computed on three sites, are illustrative rather than a precise regional estimate. Enlarging the test set and adding human-checked patches is the first thing I would do next.