HEDGEMENTATION = HEDGEROW SEGMENTATION: A REMOTE SENSING BENCHMARK

Nathan Senyard[1] **Salem Hamdani**[3*] **Astrid Zhang**[1*] **Derek Wang**[1]
Evan Shelhamer[1,4+] **Mathias Lécuyer**[1+] **Joséphine Gantois**[2+]
[1]UBC CS [2]UBC IRES & FRE [3] INSAT Tunisia [4]Vector Institute
[*]equal contribution [+]equal advising

ABSTRACT

We propose Hedgementation: a new benchmark to evaluate machine learning models for hedgerow mapping from remote sensing data at country scale and 10m^2 spatial resolution. We combine and harmonize multiple remote sensing data products and ground truth labels sourced from a hedgerow inventory in France. We measure the ability of three baseline models to generalize across spatial distance, and across climatic zones, a more explicitly challenging task. Our benchmark tests both supervised and self-supervised learning approaches for remote sensing, applied to tracking fine-scale features of high agricultural importance.

1 INTRODUCTION: HEDGEROWS AND THEIR AGRICULTURAL IMPORTANCE

Small patches of natural habitat like hedgerows play a critical role in agricultural landscapes, as they support a range of biodiversity outcomes, carbon storage, and ecosystem services (Biffi et al., 2023; Asbjornsen et al., 2014; Montgomery et al., 2020; Albrecht et al., 2020). It is well-known that hedgerows have seen sharp and ongoing declines, in line with shifts in agricultural systems (Baudry et al., 2000; de Menthière et al., 2023). Yet, little data exists on their remaining extent and dynamics, as it is costly to collect at large scales. Existing inventory efforts are geographically and temporally limited (two years of inventory in France and one year in England are the most prominent). Satellite imagery offers a promising solution to fill this data gap (Ha et al., 2019; Faucqueur et al., 2019), especially when combined with deep learning models (Liu et al., 2023; Muro et al., 2025). However progress is limited by the sparsity of labels, unknown or poor transferability of existing models to new regions, and lack of access and reproducibility of the existing learning methods.

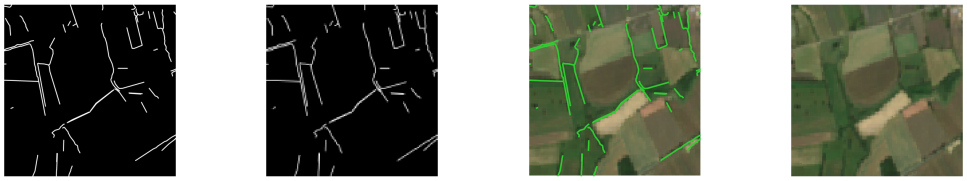
We propose a new benchmark called Hedgementation, to evaluate hedgerow detection from remote sensing data across France (§2). Hedgementation builds a dataset of imagery, embeddings, and labels: it curates and combines harmonized imagery from the Sentinel-2 satellite (ESA et al., 2025), embedding data from Alpha Earth Foundations (Brown et al., 2025) or AEF, and hedgerow labels from the BD Haie inventory produced by the French IGN institute (IGN, 2024) (§2.1). Based on this data, we design a benchmark for the hedgerow detection task. We partition the data to evaluate how models generalize across space and across agriculturally-relevant climatic zones (§2.2).

We evaluate three baselines on Hedgementation (§3): two deep networks, U-TAE (Garnot & Landrieu, 2021b) and FTW Kerner et al. (2024)), and one nearest neighbor model over the AEF embeddings. The initial results of the baselines show promise, with the best segmentation accuracy (intersection-over-union) reaching $>40\%$, and better performance in agricultural areas and in temperate zones, but underline the need for more progress to achieve accurate mapping.

2 HEDGEMENTATION BENCHMARK

2.1 THE HEDGEROW DATASET: IMAGERY, EMBEDDINGS, AND LABELS

We combine several data sources to create a hedgerow dataset in France, where imagery and labels are publicly available. The satellite imagery comes from the harmonized, surface reflectance, Sentinel-2 data product (ESA et al., 2025), accessed through Google Earth Engine (GEE) (Gorelick et al., 2017). Sentinel-2 provides 10-m resolution images of multispectral bands at 5-day intervals.



(a) Hedgerows rasterized at 0.5-m resolution. (b) Final label raster at 10-m resolution. (c) Final label raster overlaid on Sentinel-2 image. (d) Sentinel-2 image for the chosen patch.

Figure 1: Hedgerow label rasterization process, applied to one *patch*: spatial detail is preserved.

We use the 10 non-atmospheric bands, following prior work in self-supervised learning for remote sensing (Garnot & Landrieu, 2021a; Tseng et al., 2025). We gather all images acquired Sept. 2021–Oct. 2022, and drop images with more than 20% cloud cover.

The embeddings for Hedgementation are loaded from the AEF collection (Brown et al., 2025) on GEE. These are provided at a 10-m resolution in space, and at an annual resolution in time, where each "annualized" representation summarizes the full data over the calendar year. We collect the embeddings for 2022 for the highest overlap with the temporal interval of our satellite imagery. For any 10-m resolution pixel, we can thus load either the satellite imagery data product, with sensed spectral bands, or the embeddings data product, with learned bands.

We pair these remote sensing inputs with hedgerow labels from version 2 of the French hedgerow inventory BD Haie (IGN, 2024), which is unique in scope, accuracy, and public availability. These data are produced from aerial photo interpretation, validated by individual farmers in agricultural zones, and supplemented with automatic image segmentation in other areas (Commagnac et al., 2024). They are representative of the status of hedgerows in 2020-2022 (Appendix A.1).

The data cover all of France and contain 20,529,039 hedgerows, represented as line geometries with no width attribute. To match the 10-m resolution of the remote sensing inputs, we rasterize hedgerow lines through the following process, shown in Figure 1: we apply a buffer around each line to create 7m-wide hedgerows[1]; rasterize buffered lines at a 0.5-m resolution (Figure 1a); and downsample the resulting binary raster image to the target 10-m resolution by averaging (Figure 1b, 1c).

We split this national data into 1.28km x 1.28km square *patches*. This yields 603,564 data points (128x128 pixel images), which we eventually subsample to 2,995 for faster processing (§2.2).

To capture differences in vegetation type and agricultural systems, we assign an agriculturally-relevant climatic zone to each patch. We use data on Global Agro-Ecological Zones (FAO & IIASA, 2025), available at a 10-km resolution, extract the thermal regime underlying the classification, and assign to each patch the regime that it overlaps the most with. We further group patches into a temperate zone ("Temperate, cool" and "Temperate, moderately cool" regimes) and subtropic zone ("Subtropics, cool" and "Subtropics, moderately cool" regimes), which comprise $\sim$99% of the data.

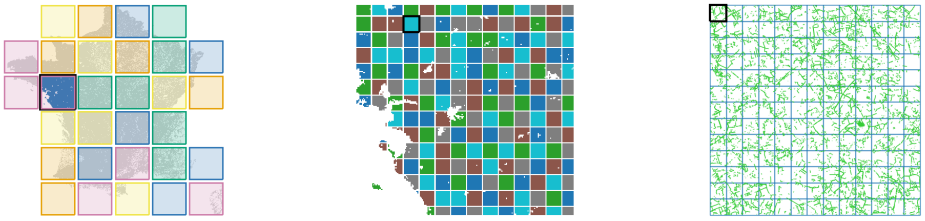
Finally, to evaluate whether the difficulty of the task varies inside and outside of agricultural zones, we use geospatial data on agricultural parcel boundaries across France (Cantelaube & Carles, 2014).

2.2 The Hedgerow Detection Task and its Evaluation

The Hedgementation benchmark uses the hedgerow dataset to evaluate semantic segmentation approaches to hedgerow detection and assess model generalization across space and climatic zones.

Assignment to train/val/test. We first group patches into cells of 12x12 patches separated by a 2-patch-wide buffer. We then randomly assign cells into one of five folds (see Figure 2b), and partition the folds into three for training (60% of the data), one for validation (20% of the data) and one for testing (20% of the data). The spatial buffer between cells ensures that no validation patch is directly adjacent to a training patch, and no test patch is directly adjacent to a training or validation patch. This is important as hedgerows and other landscape features can cross patch boundaries.

[1] 7 meters is the average width of hedgerows measured across Germany (Drexler & Don, 2024), and is the width used in hedgerow accounting calculations by the French government (de Menthière et al., 2023).



(a) Tiles for testing spatial generalization. Empty tiles are dropped.

(b) Cells with a spatial buffer, assigned to train/val/test sets.

(c) One cell of 1.28km x 1.28km patches (hedgerows are in green).

Figure 2: Dataset creation process: (2a) France is split into large tiles, randomly assigned to folds (shade colors) that determine near/far test sets for spatial generalization. (2b) Each tile is split into cells separated by a spatial buffer; cells are randomly assigned to train/val/test sets, via 5 folds (colors). (2c) Each cell is a 12x12 grid of 1.28km x 1.28km patches (individual data points).

Evaluation metric. To evaluate hedgerow detection performance, we binarize the continuous hedgerow labels ($y_i^r \in [0, 1]$ for pixel i from averaging higher-resolution binary pixels) into $y_i = 1$ iff $y_i^r > 0$. We measure the quality of our models' binary predictions $\hat{y}_i$ by the Intersection over Union (IoU) (Jaccard, 1901), a standard metric for pixel-wise classification tasks (Shelhamer* et al., 2016). IoU $= \frac{\sum_i (\hat{y}_i=1 \text{ and } y_i=1)}{\sum_i (\hat{y}_i=1 \text{ or } y_i=1)}$ with summation over all test pixels. This metric is invariant to scale and sparsity, which is relevant to our setting, and is a common metric for remote sensing tasks (Lacoste et al., 2024; Kikaki et al., 2024; Garnot & Landrieu, 2021a).

Spatial generalization. To assess spatial generalization, we split France into large *tiles* separated by a 28km buffer, as depicted in Figure 2a. We drop empty tiles, and randomly assign the 30 remaining tiles to five folds. Four folds provide train/val patches, and the test patches from those folds are potentially near training patches, thus constituting our *near test* set. From the remaining fold, we only keep test patches, which constitute our *far test* set. These test patches are at least 28km away from any training patch, and potentially much further. We can compare model performance on test patches from the *near* vs. *far test* set. We use cross-validation, with each fold serving as the *far* fold in turn, and document performance averages and variance across folds.

Overall dataset size. We subsample a 600 patches per fold stratified over each tile, splitting folds 3/1/1 across train/val/test. This creates our final dataset: 1,797 patches in train, 599 in valid, and 599 in test, totalling 2,995 patches. It contains 49,070,080 pixels, of which 4.95% are hedgerow pixels.

Climate zone generalization. To measure performance variability and generalization across different types of landscapes, we evaluate models separately on test patches in the temperate zone and on test patches in the subtropical zone. We also compare performance when training on temperate training patches only, subtropical training patches only, or both. Climatic zones are unequally represented, with temperate patches outnumbering subtropical ones 3-to-1. To remove confounding from training data size, we subsample all training sets for this analysis to the subtropical training set size.

3 Experiments: Baselines, Results, and Discussion

Models. We explore two approaches for Hedgementation and fit five models. First, we train deep learning models from scratch, using raw satellite imagery to predict hedgerows. We fit models from two machine learning for remote sensing papers: the PASTIS U-net with Temporal Attention Encoder (Garnot & Landrieu, 2021a) (PASTIS U-TAE), and the Fields of The Worlds model (Kerner et al., 2024) (FTW). Both models are trained from random initialization, though we experiment with fine-tuning from pre-training (Appendix D).

We use the PASTIS U-TAE as-is, altering only the output layer for binary classification, with no changes to model architecture or hyperparameters. The FTW model uses fewer spectral bands and only two input images from different times of the year. Appendix A.2 details its alteration.

Second, we use the AEF embeddings (Brown et al., 2025) with simpler models. These embeddings have been positively evaluated on vegetation and agriculture tasks (Ma et al., 2025; Houriez et al.,

	G0		G1		G2		G3		G4		Avg	
Method	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far
PASTIS U-TAE	0.409	0.417	0.408	0.392	0.416	0.341	0.425	0.374	0.393	0.420	**0.4102**	**0.3888**
FTW Model	0.32	0.35	0.31	0.34	0.30	0.22	0.30	0.29	0.31	0.27	**0.31**	**0.29**
AEF: Random Forest	0.236	0.228	0.232	0.245	0.243	0.212	0.231	0.249	0.232	0.249	**0.235**	**0.237**
AEF: Log. Reg.	0.206	0.198	0.202	0.214	0.211	0.172	0.200	0.217	0.209	0.193	**0.206**	**0.200**
AEF: KNN (k=3)	0.155	0.134	0.147	0.158	0.155	0.119	0.145	0.173	0.155	0.136	**0.151**	**0.157**

Table 1: IoU (higher = more accurate) on our Hedgementation benchmark across groups (G0, . . .).

Trained on	**Tested on**	**IoU**
Full dataset	Full dataset	0.416
Full dataset	Temp. only	0.424
Full dataset	Subtropic only	0.374
Subs. dataset	Full dataset	0.368
Subs. dataset	Temp. only	0.380
Subs. dataset	Subtropic only	0.316

Trained on	**Tested on**	**IoU**
Subtropic only	Full dataset	0.352
Subtropic only	Temp. only	0.357
Subtropic only	Subtropic only	0.331
Temperate only	Full dataset	0.357
Temperate only	Temp. only	0.372
Temperate only	Subtropic only	0.287

Trained on	**Tested on**	**IoU**
All pix (subs.)	All pix	0.410
All pix (subs.)	Agri. pix	0.459
All pix (subs.)	Non-Agri. pix	0.297
Agri. pix	All pix	0.378
Agri. pix	Agri. pix	0.457
Agri. pix	Non-Agri. pix	0.243

Table 2: PASTIS U-TAE generalization on climatic zones (left, center) and agricultural zones (right).

2025). We use them with: a kNN enhanced with 3×3 spatial context via concatenated embeddings and center weighting to reduce noise, with a $200k$ balanced training set restricting positive samples to high-confidence interior hedgerow pixels; a Random Forest; and a Logistic Regression. Details and hyper-parameter tuning are in Appendix D.

Analysis: Generalization over space. Table 1 compares baseline performance on the near and far test sets. Figure 4 and Appendix C show qualitative predictions on four patches. On average, the PASTIS and FTW models are better on nearer patches than on further ones (+0.02 IoU/more than 5% higher), as expected. kNN performs similarly on near and far. This might be due to the global training of the embedding. However, performance is weak overall (see Figure 4). The difference due to distance is minor, with significant variation across near/far tile groupings, with (G0, G4) yielding better performance on the far set. We thus turn to generalization over climatic zones.

Analysis: Generalization over climatic zones. Table 2 focuses on our best model (PASTIS U-TAE) to compare generalization across climate zones. Subtropical patches are consistently more difficult to predict, even for the model trained exclusively on these patches. The smallest gap is for the subtropic-only model (2.1%), and the largest gap is for the temperate-only model (7.0%). Interestingly, while training on a random subset does worse on subtropic tiles than the subtropic-only model, it does better on temperate tiles than the temperate-only model. We confirm that dataset size matters: the model trained on less data is much worse than the model trained on all the data.

Analysis: Performance in agricultural zones. Hedgerows are prevalent and particularly important in agricultural regions, where BD Haies label quality is also higher thanks to validation by individual farmers. We find that training on all pixels yields better results than restricting training to agricultural zones (Table 2). However, predicting hedgerows in agricultural landscapes is significantly easier: the IoU of the PASTIS U-TAE model reaches 45.9% of IoU on agricultural pixels, up from 41% on all pixels. This is a valuable and encouraging finding for applications that focus on these landscapes.

Discussion: usage and limitations. The hedgerow task is complementary to agricultural tasks such as field boundary mapping (Kerner et al., 2024) and crop type mapping (Garnot & Landrieu, 2021b; Rußwurm et al., 2019; Van Tricht, 2021). It is a spatial challenge due to the thin extents, diverse aspects, and various arrangements of hedgerows, and a generalization challenge due to the variations in the surrounding landscapes (complexity, tree density) across climate / agro-ecological zones.

The Hedgementation imagery and labels are derived from free and public data sources, for accessibility and reproducibility, and extensive quality checks support its labeling. As a benchmark with ground truth on the ground, it complements benchmarks with image-derived labels (Helber et al., 2019). We discuss additional related work, limitations, and next steps in Appendix B. While there are improvements to be made, this version of Hedgementation can already inform model development with its applied task and measures of spatial and agricultural generalization.

ACKNOWLEDGMENTS

We thank Perrine Porcher for discussions and experiments during her Vector internship about the PASTIS U-TAE model and its variations as applied to agricultural mapping.

REFERENCES

Matthias Albrecht, David Kleijn, Neal M Williams, Matthias Tschumi, Brett R Blaauw, Riccardo Bommarco, Alistair J Campbell, Matteo Dainese, Francis A Drummond, Martin H Entling, et al. The effectiveness of flower strips and hedgerows on pest control, pollination services and crop yield: a quantitative synthesis. *Ecology letters*, 23(10):1488–1498, 2020.

Heidi Asbjornsen, Virginia Hernandez-Santana, Matthew Liebman, Jules Bayala, Jiquan Chen, Matthew Helmers, CK Ong, and LA Schulte. Targeting perennial vegetation in agricultural landscapes for enhancing ecosystem services. *Renewable Agriculture and Food Systems*, 29(2):101–125, 2014.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19530–19540, 2025.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16772–16782, 2023.

Jacques Baudry, RGH Bunce, and Françoise Burel. Hedgerows: an international perspective on their origin, function and management. *Journal of environmental management*, 60(1):7–22, 2000.

Sofia Biffi, Pippa J Chapman, Richard P Grayson, and Guy Ziv. Planting hedgerows: Biomass carbon sequestration and contribution towards net-zero targets. *Science of the Total Environment*, 892:164482, 2023.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025.

Pierre Cantelaube and Marie Carles. Le registre parcellaire graphique: des données géographiques pour décrire la couverture du sol agricole. *NOV'AE-Ingénierie et savoir-faire innovants*, (spécial Cahier des techniques):58–64, 2014.

Loïc Commagnac, Frédéric Letouze, and Louise Le Bellec. Dispositif national de suivi des bocages : mise à jour du référentiel des haies établi lors de la phase 1. Rapport technique, Institut national de l'information géographique et forestière (IGN) / Office français de la biodiversité (OFB), 1 2024. URL `https://geoservices.ign.fr/sites/default/files/2025-06/Descriptif_de_contenu_et_limite_DSB_v2.pdf`. Version 1.0, Conventions MTECT–OFB–IGN.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Catherine de Menthière, Patrick Falcone, Vincent Piveteau, and Xavier Ory. La haie, levier de la planification écologique. *Rapport du CGAAER*, 22114:116, 2023.

Stefania Di Tommaso, Sherrie Wang, and David B. Lobell. Combining gedi and sentinel-2 for wall-to-wall mapping of tall and short crops. *Environmental Research Letters*, 16(12):125002, 2021.

Sophie Drexler and Axel Don. Carbon sequestration potential in hedgerow soils: Results from 23 sites in germany. *Geoderma*, 445:116878, 2024.

European Space Agency ESA, Copernicus, and Google Earth Engine. Harmonized sentinel-2 msi: Multispectral instrument, level-2a (sr), 2025. Accessed: 2025-08-01 from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED.

FAO and IIASA. Global agro-ecological zones version 4 (GAEZ v4). http://www.fao.org/gaez/, 2025. Accessed: November 12, 2025.

Loic Faucqueur, Nathalie Morin, Antoine Masse, Pierre-Yves Remy, Justine Hugé, Clémence Kenner, Fabrice Dazin, Baudouin Desclée, and Christophe Sannier. A new copernicus high resolution layer at pan-european scale: Small woody features. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI*, volume 11149, pp. 268–278. SPIE, 2019.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4872–4881, 2021a.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks, 2021b.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: 10.1016/j.rse.2017.06.031. URL https://doi.org/10.1016/j.rse.2017.06.031.

Thuan V Ha, Beyhan Y Amichev, Kenneth W Belcher, Murray J Bentham, Suren N Kulshreshtha, Colin P Laroque, and Ken CJ Van Rees. Shelterbelt agroforestry systems inventory and removal analyzed by object-based classification of satellite data in saskatchewan, canada. *Canadian Journal of Remote Sensing*, 45(2):246–263, 2019.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. Olmoearth: Stable latent image modeling for multimodal earth observation. *arXiv preprint arXiv:2511.13655*, 2025.

Luc Houriez, Sebastian Pilarski, Behzad Vahedi, Ali Ahmadalipour, Teo Honda Scully, Nicholas Aflitto, David Andre, Caroline Jaffe, Martha Wedner, Rich Mazzola, et al. Scalable geospatial data generation using alphaearth foundations model. *arXiv preprint arXiv:2508.11739*, 2025.

IGN and OFB. Les haies du dispositif de suivi des bocages. Rapport technique, Institut national de l'information géographique et forestière (IGN) / Office français de la biodiversité (OFB), 12 2020. URL https://geoservices.ign.fr/sites/default/files/2021-07/Descriptif_de_contenu_et_limite_DSB.pdf. Version 1.0.

Institut national de l'information géographique et forestière IGN. Bd haie v2 mars 2024, 2024. Accessed: 2025-08-01 from Link: https://geoservices.ign.fr/bdhaie.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. TerraMind: Large-scale generative multimodality for earth observation. In *ICCV*, 2025. URL https://arxiv.org/abs/2504.11171.

Hannah Kerner, Snehal Chaudhari, Aninda Ghosh, Caleb Robinson, Adeel Ahmad, Eddie Choi, Nathan Jacobs, Chris Holmes, Matthias Mohr, Rahul Dodhia, Juan M. Lavista Ferres, and Jennifer Marcus. Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation, 2024.

Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:39–54, 2024.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. GEO-Bench: Toward foundation models for earth monitoring. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Siyu Liu, Martin Brandt, Thomas Nord-Larsen, Jerome Chave, Florian Reiner, Nico Lang, Xiaoye Tong, Philippe Ciais, Christian Igel, Adrian Pascual, et al. The overlooked contribution of trees outside forests to tree cover and woody biomass across europe. *Science Advances*, 9(37): eadh4097, 2023.

Yuchi Ma, Yawen Shen, Anu Swatantran, and David B Lobell. Harvesting alphaearth: Benchmarking the geospatial foundation model for agricultural downstream tasks. *arXiv preprint arXiv:2601.00857*, 2025.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pp. 15630–15649. PMLR, 2022.

Ian Montgomery, Tancredi Caruso, and Neil Reid. Hedgerows as ecosystems: service delivery, management, and restoration. *Annual Review of Ecology, Evolution, and Systematics*, 51(1):81–102, 2020.

Javier Muro, Lukas Blickensdörfer, Axel Don, Anna Köber, Sarah Asam, Marcel Schwieder, and Stefan Erasmi. Hedgerow mapping with high resolution satellite imagery to support policy initiatives at national level. *Remote Sensing of Environment*, 328:114870, 2025.

Planet Labs PBC. Planet application program interface: In space for life on earth, 2020–2022. URL `https://api.planet.com`.

Arjun Rao and Esther Rolf. Using multiple input modalities can improve data-efficiency and ood generalization for ml with satellite imagery. In *TerraBytes-Towards global datasets and models for Earth Observation*, pp. 166–188. PMLR, 2025.

Marc Rußwurm, Sébastien Lefèvre, and Marco Körner. BreizhCrops: A satellite time series dataset for crop type identification. In *Proceedings of the International Conference on Machine Learning Time Series Workshop*, volume 3, 2019.

Evan Shelhamer*, Jonathan Long*, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *PAMI*, 2016.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *Forty-second International Conference on Machine Learning*, 2025.

Kristof Van Tricht. Mapping crops at global scale! what works and what doesn't? `https://blog.vito.be/remotesensing/worldcereal-benchmarking`, 2021. Accessed: 2023-07-31.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2204–2214, 2025.

A Data and Methods

A.1 Hedgerow Labels

Labels come from a nationwide inventory of hedgerows in France that was recently updated. The original inventory, "BD Haies V1", is obtained by aggregation and linearization of two data sources: (1) hedgerows identified in agricultural landscapes as part of compulsory reporting for the *Registre Parcellaire Graphique* (RPG), which covers more than 90% of the agriculture zone (Cantelaube & Carles, 2014)— identification of these hedgerows uses pre-processed aerial photographs, dating from 2011-2014 depending on the region, and validated by individual farmers; and (2) hedgerows identified by automatic image segmentation and photo interpretation, as part of the vegetation layer of the national topographic database BD TOPO—identification of these hedgerows uses imagery from 2004-2015 depending on the region, is not restricted to the agricultural zone, but is restricted to longer (25m+) and higher (1,30m+) hedgerows (IGN & OFB, 2020). BD Haies V1 is thus representative of the status of hedgerows around 2004-2014. The updated version, "BD Haies V2", used in this paper, is constructed by revision of BD Haies V1 and involves two main data sources: (1) hedgerows identified in agricultural landscapes as part of more recent iterations of the RPG, now involving imagery from 2020-2022 depending on the region; and (2) numerical surface models from corresponding years that serve to further validate the continued existence of V1 hedgerows (Commagnac et al., 2024). BD Haies V2 is thus representative of the status of hedgerows around 2020-2022.

A.2 Pre-processing Inputs for the Fields of the World Model

FTW uses input satellite imagery in a slightly different format than PASTIS, such that it cannot ingest as is the data provided by Hedgementation. We pre-process input data as follows for use by the FTW model:

- **Scene date selection:** FTW uses imagery from two temporal windows per geographical zone, corresponding to the planting and harvest seasons, with dates adapted to each country's agricultural calendar (e.g., in France, March–June for window A and August–September for window B). Basde on the Sentinel-2 imagery extracted for Hedgementation, each window contains between 30 and 60 satellite images. For each window (A and B), a single representative image is selected based on the median acquisition date within the interval, reducing sensitivity to irregular acquisitions and transient noise.
- **Cloud coverage:** No modification is applied.
- **Spatial dimensions:** The original spatial dimensions of the generated dataset are preserved.
- **Spatial resolution:** No modification is applied.
- **Spectral bands:** Only the B04, B03, B02, and B08 bands are extracted from the generated dataset patches and stacked following the same procedure used in FTW.

Figure 3 illustrates the FTW input data processing. Window A corresponds to the median composite of Sentinel-2 images acquired during the planting season (March–June 2022), while Window B corresponds to the median composite of images acquired during the harvest season (August–October 2022). Each window is represented by a four-band multispectral image covering the same agricultural area.

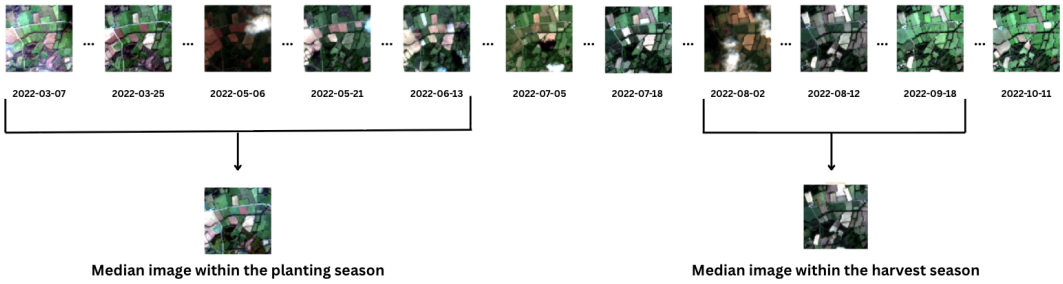


Figure 3: Input data for the FTW model: the model requires two input images taken from two characteristic times of year. We collect the necessary imagery for the model from the image time-series collected for our Hedgementation dataset.

B Related Work, Limitations, and Next Steps

Related Work. Pixel-wise prediction tasks, like our hedgerow task, are valuable for benchmarking not only recognition (what) but localization (where). As evidence of their importance, the popular GEO-Bench suite of tasks includes both image classification and pixel classification (Lacoste et al., 2024). Pixel-wise benchmarks are likewise included in the comparisons of pre-trained models for remote sensing (or "foundation models") (Cong et al., 2022; Tseng et al., 2025; Astruc et al., 2025; Brown et al., 2025; Waldmann et al., 2025; Jakubik et al., 2025).

Systematic hedgerow detection is not a new goal, and capacities have increased with the availability of remote sensing and machine learning tools. Yet, relatively few papers propose and evaluate approaches to map hedgerows at regional scales where data does not yet exist, in part due to limitations in label data or resource availability. One such example, Ha et al. (2019), use Sentinel imagery and object-based classification to map shelterbelts across Saskatchewan in 2016. Another, Muro et al. (2025), use high-resolution PlanetScope imagery and semantic segmentation (standard U-net with a ResNet backbone) to map hedgerows across one state in Germany and evaluate transferability to other states. Related initiatives exist to map wooded landscape features more broadly. In particular, the Copernicus Land Monitoring Service produces the Small Woody Features data layer that spans much of Europe and includes hedgerows, though accuracy of hedgerow detection specifically is not known (Faucqueur et al., 2019). And Liu et al. (2023) map all trees outside forests in Europe and evaluates transferability to boreal and temperate zones of North America, using PlanetScope imagery and semantic segmentation (U-Net with and EfficientNet-B4 backbone), though LiDAR-derived height data are used in lieu of annotations. A key limitation is that performance is not comparable across models, and workflows are generally not reproducible, which limits expansion of modeling approaches beyond their original spatial and temporal extent.

Our task is complementary to other agricultural segmentation tasks that focus on groups of points, like the recognition of crop types and fields Garnot & Landrieu (2021a), because hedgerow segmentation focuses on lines. In this way it is more related to the task of mapping field boundaries (Kerner et al., 2024). At the same time, the hedgerow task is distinct: not all hedgerows are field boundaries, and not all field boundaries are hedgerows.

Limitations and Next Steps.

Label Quality and Consistency. The BD Haie data from which we derive our benchmark labels is intended to be exhaustive and accurate. However, it may still be subject to noise and inconsistency, especially due to differences in hedgerow definition and in the detection process inside and outside of agricultural areas, use of different years of imagery across the country, changes in methodology across years, and lack of independent accuracy assessment. To verify the quality and consistency of annotations, we will examine higher-resolution imagery provided by Planet PBC (2020–2022) and check for label noise by monitoring for outliers in the training loss.

Scope in Space, Time, and Modality. Hedgementation covers the one country of France, in the one time period of 2021–2022, and imagery from one satellite mission and optical modality: Sentinel-2. We will next extend the benchmark in geography in modality. For geography, we will propagate

then verify labels in related agricultural regions of the UK and Canada (Houriez et al., 2025). For modality, we will combine our optical imagery with an additional modality such as SAR or LIDAR, as they are sensitive to height. These sensors have been incorporated into models for recognizing tall and short crops (Di Tommaso et al., 2021) and could help differentiate hedgerows from other landscape features like fences. Furthermore multi-modal data has been shown to help with data efficiency and generalization (Rao & Rolf, 2025), and we have already confirmed that generalization across agro-ecological zones is a challenge.

Prediction Quality. The accuracies achieved by our baselines are promising but insufficient for agricultural and ecological analysis. We will strive to improve performance by transfer learning from supervised pre-training (Garnot & Landrieu, 2021b; Bastani et al., 2023) or self-supervised pre-training (Herzog et al., 2025), more sophisticated sampling for curriculum learning (Mindermann et al., 2022) and countering imbalance, and further tuning of the architecture and optimization.

C Qualitative Results: Predictions

Figure 4 shows four patches (first row) and predictions from the PASTIS, FTW, kNN, Random Forest, and Logistic Regression models on those in rows 2, 3, 4, 5, and 6 respectively. True Positives are shown in green, False Negatives are shown in yellow, and False Positives are shown in magenta.

D Supplementary Results: Generalization over Space

D.1 PASTIS UTAE

PASTIS UTAE on Hedgementation								
Experiment	**Test(Near)**				**Test(Far)**			
Far Group	**IoU**	**Precision**	**Recall**	**F1**	**IoU**	**Precision**	**Recall**	**F1**
G0	0.409	0.529	0.643	0.58	0.417	0.524	0.67	0.588
G1	0.408	0.526	0.645	0.579	0.392	0.48	0.683	0.563
G2	0.416	0.544	0.638	0.587	0.341	0.551	0.473	0.509
G3	0.425	0.537	0.672	0.597	0.374	0.499	0.599	0.544
G4	0.393	0.523	0.613	0.565	0.42	0.568	0.617	0.591
Avg	**0.4102**	**0.5318**	**0.6422**	**0.5816**	**0.3888**	**0.5244**	**0.6084**	**0.559**

Table 3: Results on the Far and Near test sets, when trained on the Near train set, for all tile-folds.

Table 3 shows the performance of the PASTIS U-TAE model on our Hedgementation benchmark. Focusing on the IoU, our prefered metric, we see that mean IoU is lower when evaluating on the far group than the near group with a 0.021% gap. However, this trend is less consistent at the level of individual folds and models. Two models, G0 and G4, actually perform better on their far group than on the near group. Models G1, G2 and G3 all perform worse on their near group, by 1.6%, 7.5%, and 5.1% IoU respectively.

Model Architecture	**Weights**	**Layers Trained**	**IoU**	**Precision**	**Recall**	**F1**
PASTIS U-TAE	Random Init	All	0.415	0.533	0.651	0.586
PASTIS U-TAE	PASTIS	All	0.414	0.551	0.624	0.585
PASTIS U-TAE	PASTIS	out_conv	0.081	0.242	0.108	0.149

Table 4: Pretrained vs. Randomly Initialized Weights

In addition, we experiment with fine-tuning the weights of the original PASTIS model for our semantic segmentation task. We test two approaches: fine-tuning all the layers of the model; and fine-tuning the final convolutional block. All hyperparameters except for the initial weights were shared with our other experiments. Neither approach yield an improvement over training from randomly initialized weights. Training all layers performs almost exactly on-par with the results of

training from scratch, while training only the final convolutional block yields significantly inferior results. Additional experiments to explore this further are an interesting avenue for future work.

D.2 FIELDS OF THE WORLD

Table 5 shows the results of run the FTW model on our extrapolation over space task. We observe that the far group is slightly harder to predict on average, but the difference is not very large. The performance of this model is lower than that of our best model, the PASTIS U-TAE §D.1.

FTW on Hedgementation								
Experiment	**Test (Near)**				**Test (Far)**			
Far Group	**IoU**	**Precision**	**Recall**	**F1**	**IoU**	**Precision**	**Recall**	**F1**
G0	0.32	0.67	0.38	0.49	0.35	0.65	0.42	0.51
G1	0.31	0.60	0.39	0.47	0.34	0.59	0.44	0.50
G2	0.30	0.70	0.34	0.46	0.22	0.67	0.25	0.36
G3	0.30	0.69	0.35	0.46	0.29	0.74	0.33	0.45
G4	0.31	0.66	0.38	0.48	0.27	0.65	0.31	0.42
Avg	**0.31**	**0.66**	**0.37**	**0.47**	**0.29**	**0.66**	**0.35**	**0.45**

Table 5: Results by far group on Hedgementation (Near vs. Far test sets).

Fine-tuning however can help the FTW model: as we see in Table 6, initializing with ImageNet weights provides a modest but consistent improvement over random initialization, with IoU increasing from 0.333 to 0.359 and F1 from 0.500 to 0.529. This suggests that low-level visual features learned on natural images transfer usefully to satellite imagery even across this domain gap.

Model Architecture	**Weights**	**Layers Trained**	**IoU**	**Precision**	**Recall**	**F1**
FTW	Random Init	All	0.333	0.607	0.425	0.500
FTW	ImageNet	All	0.359	0.575	0.489	0.529

Table 6: Pretrained vs. Randomly Initialized Weights for FTW

D.3 KNN + EMBEDDINGS

We use pre-trained, publicly released AEF embeddings Brown et al. (2025) to train a k-Nearest Neighbours (kNN) model for the hedgerow detection task. We tune the number of neighbours $k \in \{1, 3, 5, \ldots, 25\}$ and the training set size, constructed by randomly sampling a balanced set of hedgerow and non-hedgerow pixels from the training data. We explore training set sizes ranging from very small samples (e.g., 1 positive / 1 negative) to larger balanced subsets.

We first fix $k = 1$ and evaluate performance as a function of training set size. Performance improves as more training samples are added, confirming the expected behavior of kNN. We then fix the training set size to 100 positive / 100 negative samples and tune k. We find that $k = 5$ achieves the best overall performance under this setup (Table 7).

Experiment	**Test (Near)**				**Test (Far)**			
Far Group	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
G0	0.109	0.114	0.774	0.190	0.094	0.097	0.764	0.164
G1	0.096	0.099	0.807	0.169	0.121	0.128	0.787	0.208
G2	0.122	0.129	0.795	0.208	0.090	0.094	0.783	0.159
G3	0.108	0.112	0.783	0.187	0.151	0.160	0.796	0.251
G4	0.115	0.120	0.815	0.196	0.094	0.096	0.821	0.165
Avg	**0.110**	**0.115**	**0.795**	**0.190**	**0.110**	**0.115**	**0.790**	**0.189**

Table 7: KNN on Hedgementation (k=5, 100 positive/ 100 negative).

To further improve the kNN baseline for hedgerow detection, we incorporate local spatial context by representing each pixel with a concatenation of embeddings from its 3×3 neighborhood. The embeddings in this 3×3 grid are concatenated in a fixed spatial order, forming a feature vector that captures local structure. Furthermore, to reduce label noise, we only select positive samples that are interior hedgerow pixels: that is, where both the center and neighboring pixels are labeled as positive. We then add spatial weighting by scaling the center embedding prior to distance computation by a factor. Moderate upweighting improves performance by reducing the influence of noisy neighbors, with a center weight of 6 achieving the best results. Larger weights provide no additional benefit, indicating a trade-off between center emphasis and contextual information. With this representation, $k = 3$ achieves the best performance, while larger values degrade performance due to increased averaging over heterogeneous neighborhoods. Increasing the number of training samples up to 100,000 positive and negative samples, while enforcing per-tile caps to maintain diversity, improves IoU from approximately 0.11 to 0.15 (Table 8). The similar performance on both Near and Far test sets shows that AEF embeddings, trained on the whole world, favor generalization over space compared to training from scratch.

KNN on Hedgementation (k=3, 100,000 positive/ 100,000 negative)								
Experiment	**Test (Near)**				**Test (Far)**			
Far Group	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
G0	0.155	0.166	0.735	0.257	0.134	0.141	0.729	0.222
G1	0.147	0.156	0.738	0.244	0.158	0.170	0.728	0.260
G2	0.155	0.166	0.740	0.256	0.119	0.125	0.728	0.205
G3	0.145	0.153	0.736	0.255	0.173	0.187	0.735	0.230
G4	0.155	0.164	0.751	0.196	0.136	0.144	0.749	0.165
Avg	**0.151**	**0.161**	**0.740**	**0.251**	**0.147**	**0.157**	**0.733**	**0.245**

Table 8: Performance of the best kNN configuration on the Near and Far test sets. Results are reported across Far Groups (G0–G4), with averages shown in the last row. The model uses 3×3 neighborhood concatenation, center weighting of 6, $k = 3$, and 100,000 positive/100,000 negative training samples.

D.4 Random Forest + Embeddings

We also fit a Random Forest model on the usefulness of AEF embeddings Brown et al. (2025) for the hedgerow detection task. We tune the following hyper-parameters: number of estimators (N_{est}), minimum samples per leaf ($N_{min\_leaf\_samples}$), maximum depth of each tree ($N_{max\_tree\_depth}$), and the probability threshold for prediction ($proba\_threshold$).

We also set class weights to balanced, to counteract the imbalance of our dataset (overwhelming number of negatives compared to positives). Balancing class weights assigns a higher weight to the minority class (positives) and a lower weight to the majority class (negatives) when fitting the model. Without balanced weights, our model just predicts negative for all points to achieve great accuracy, compromising IoU.

We use 100,000 positive and negative samples to compare with the kNN, and the whole training set to maximize performance. We tune parameters sequentially. We first fix the probability threshold at 0.5 and search over the following space: N_{est} between 100 and 1000, $N_{min\_leaf\_samples}$ between 1 and 5, $N_{max\_tree\_depth}$ in $[20, 40, \infty]$, and pick the combination with highest validation IoU: $N_{est} = 500$, $N_{min\_leaf\_samples} = 2$, $N_{max\_tree\_depth} = \infty$. (We note that $N_{est} = 1000$ achieves a slightly better IoU than $N_{est} = 500$, around 0.007% higher, but the training time is 50% longer for such a minor gain).

We then tune the probability threshold ($proba\_threshold$) between 0.1 and 0.9. The best $proba\_threshold$ is different for the two training set variations, because one uses a down-sampled balanced dataset while the other uses the entire unbalanced training set. We conduct the space-extrapolation experiments using the best probability thresholds for the two variations respectively, with $proba\_threshold = 0.7$ for the balanced 100,000 positive / 100,000 negative samples model, and $proba\_threshold = 0.2$ for the one that uses the entire training set.

Random Forest on Hedgementation (100,000 Pos / Neg Samples)								
N_{est} = 500, $N_{min\_leaf\_samples}$ = 2, $N_{max\_tree\_depth}$ = ∞, proba_threshold=0.7								
Experiment	**Test (Near)**				**Test (Far)**			
Far Group	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
G0	0.226	0.324	0.501	0.389	0.213	0.282	0.526	0.373
G1	0.221	0.323	0.492	0.388	0.233	0.308	0.558	0.382
G2	0.231	0.317	0.533	0.394	0.205	0.312	0.443	0.377
G3	0.218	0.311	0.506	0.384	0.239	0.346	0.500	0.387
G4	0.221	0.312	0.499	0.379	0.236	0.360	0.538	0.433
Avg	**0.223**	**0.317**	**0.506**	**0.387**	**0.225**	**0.321**	**0.513**	**0.391**

Table 9: Performance of the Random Forest configuration on the Near and Far test sets using balanced 100,000 positive/negative samples. Results are reported across Far Groups (G0–G4), with averages shown in the last row. The model uses 500 trees, minimum samples per leaf as 2, infinite maximum tree depth, probability threshold as 0.7, and 100,000 positive / 100,000 negative samples.

Table 9 shows results with 100,000 positive and 100,000 negative training samples, which is directly comparable to the kNN. The random forest classifier produces much better IoU than the best kNN with the same number of samples, outperforming it on average by 7% points of IoU. It also has a 16% increase on precision (but sees a 20% drop in recall) leading to an overall better F1. As with the kNN, the performance is almost identical for near and far test sets across groups, showing that the AEF embeddings enable good generalization over space.

Random Forest on Hedgementation								
N_{est} = 500, $N_{min\_leaf\_samples}$ = 2, $N_{max\_tree\_depth}$ = ∞, proba_threshold=0.2								
Experiment	**Test (Near)**				**Test (Far)**			
Far Group	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
G0	0.236	0.358	0.471	0.407	0.228	0.328	0.488	0.397
G1	0.232	0.356	0.463	0.407	0.245	0.346	0.513	0.400
G2	0.243	0.356	0.497	0.412	0.212	0.348	0.417	0.387
G3	0.231	0.350	0.473	0.406	0.249	0.369	0.475	0.404
G4	0.232	0.349	0.467	0.396	0.249	0.380	0.509	0.456
Avg	**0.235**	**0.354**	**0.474**	**0.406**	**0.237**	**0.354**	**0.480**	**0.409**

Table 10: Performance of the best Random Forest configuration on the Near and Far test sets. Results are reported across Far Groups (G0–G4), with averages shown in the last row. The model uses 500 trees, minimum samples per leaf as 2, infinite maximum tree depth, probability threshold as 0.2, and the entire training set.

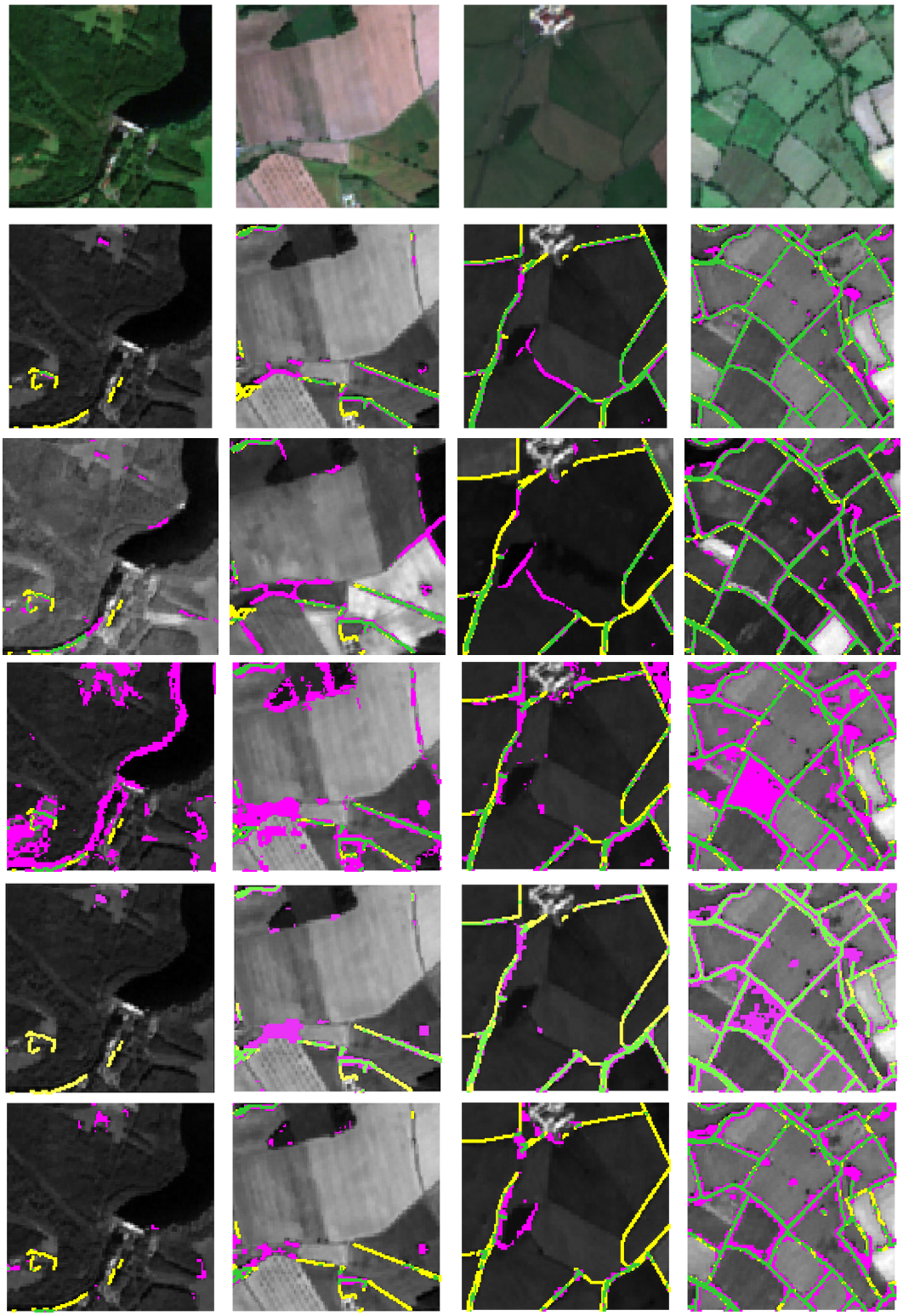


Figure 4: Qualitative examples of predictions for our best model in each class. We show input RGB patches (top), and predictions with TP=green, FN=yellow, and FP=magenta on following rows, in order for PASTIS, FTW, kNN (k=3), Random Forest and Logistic Regression.

Table 10 shows results using the full training set, yielding a 1.2% point improvement in IoU compared to a $200k$ balanced training set. The performance reaches an IoU of 0.237 (on test Far), a significant improvement over the kNN, but still far from deep learning models trained from scratch: our best model, the PASTIS U-TAE reaches 38.8% IoU on test Far, and 41% on test Near. However, compared to the PASTIS U-TAE model, the Random Forest model on AEF embeddings fully generalizes over space: the Near and Far IoUs are almost identical (a difference of 0.2% point change in IoU, slightly better on the Far test), compared to a 2.1% point drop on the Far test for the PASTIS model.

D.5 Logistic Regression + Embeddings

Finally, we fit a Logistic Regression model using the AEF Brown et al. (2025) embeddings. We use an ℓ_2 regularization using the *saga* solver with regularization parameter $C = 1.0$. The training data is constructed by sampling up to 1000 pixels per tile, followed by global balancing between hedgerow and non-hedgerow classes.

We perform hyper-parameter tuning over the prediction threshold $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. We observe that performance improves as τ increases up to $\tau = 0.8$, and degrades beyond this point. Table 11 shows that at $\tau = 0.8$, the model reaches an IoU of 20% on both test Near and Far, close to (but lower than) the Random Forest. Once again, using the AEF embedding shows strong generalization over space (a 0.6% drop of IoU between Near and Far tests, against 2.1% for PASTIS' U-TAE).

Logistic Regression on Hedgementation								
Experiment	**Test (Near)**				**Test (Far)**			
Far Group	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
G0	0.206	0.297	0.446	0.325	0.198	0.253	0.495	0.310
G1	0.202	0.279	0.453	0.318	0.214	0.309	0.499	0.337
G2	0.211	0.296	0.465	0.330	0.172	0.253	0.371	0.277
G3	0.200	0.283	0.451	0.317	0.217	0.340	0.407	0.335
G4	0.209	0.292	0.473	0.327	0.193	0.257	0.485	0.309
Avg	**0.206**	**0.289**	**0.457**	**0.323**	**0.200**	**0.283**	**0.450**	**0.314**

Table 11: Performance of logistic regression on the Near and Far test sets across Far Groups (G0–G4).

E Supplementary Results: Generalization over Agriculturally Relevant Climatic Zones

E.1 PASTIS UTAE

Table 12 shows the results of the PASTIS U-TAE model on the Hedgementation task evaluating generalizations over agriculturally-relevant climate zones. We make three observations.

First, the size of the training set has a large impact on model performance, with a 4.8% point drop in IoU on a uniformly subsampled dataset the size of the subtropic dataset. We thus subsample all datasets to the smallest size to make performance comparable.

Second, subtropic patches are consistently the most difficult to segment, even for models trained exclusively on those patches. As one may expect, the difference is at it's smallest for the subtropic-only model, where it is only 2.1%, and at it's largest for the temperate-only trained model, where it is 7.0%.

Third, while the downsampled model performs worse on subtropic tiles than the subtropic-only model, it performs better on temperate tiles than the temperate-only model. It also performs best on the full test set. This may indicate that subtropic tiles contain important information useful to generalize to subtropic tiles, while a diversity of climate zones is important for genralization.

Architecture	Trained On	Tested On	IoU	Precision	Recall	F1
PASTIS	Full Dataset	Full Dataset	0.416	0.538	0.646	0.587
PASTIS	Full Dataset	Temp. Only	0.424	0.542	0.661	0.596
PASTIS	Full Dataset	Subtropic Only	0.374	0.518	0.574	0.544
PASTIS	Subtropic Only	Full Dataset	0.352	0.453	0.613	0.521
PASTIS	Subtropic Only	Temp. Only	0.357	0.46	0.615	0.526
PASTIS	Subtropic Only	Subtropic Only	0.331	0.422	0.604	0.497
PASTIS	Temperate Only	Full Dataset	0.357	0.456	0.621	0.526
PASTIS	Temperate Only	Temp. Only	0.372	0.47	0.642	0.542
PASTIS	Temperate Only	Subtropic Only	0.287	0.392	0.52	0.447
PASTIS	Subs. Dataset	Full Dataset	0.368	0.468	0.634	0.538
PASTIS	Subs. Dataset	Temp. Only	0.38	0.478	0.648	0.55
PASTIS	Subs. Dataset	Subtropic Only	0.316	0.419	0.563	0.48

Table 12: Results across climate zones for models trained all, only subtropic, and only temperate patches. All datasets after the first group are random subsets matching the size of the subtropic only dataset (the smallest subset).

E.2 Fields of The Wold

Table 13 shows the impact of climate zones on the FTW model. Again, subtropic patches are consistently the most difficult across all models, even for the model trained exclusively on subtropic patches. The gap between subtropic and temperate performance varies by training set. As expected, this difference is smallest for the subtropic-only model, where it is only 2.3% (0.184 vs. 0.161 IoU), and largest for the temperate-only model, where it reaches 2.1% (0.1234 vs. 0.102 IoU).

Interestingly, the subsampled (but complete) training set outperforms significantly outperforms all other models. For instance, it outperforms the subtropic-only model even when testing on subtropic tiles (0.271 vs. 0.161 IoU) and also outperforms the temperate-only model on temperate tiles (0.331 vs. 0.1234 IoU). This suggests that exposure to diverse geographic contexts during training provides generalization benefits that outweigh the reduction in per-class sample count.

Architecture	Trained On	Tested On	IoU	Precision	Recall	F1
FTW	Full Dataset	Full Dataset	0.333	0.6069	0.4246	0.4996
FTW	Full Dataset	Temp. Only	0.3432	0.616	0.4365	0.511
FTW	Full Dataset	Subtropic Only	0.2839	0.5583	0.3661	0.4422
FTW	Subtropic Only	Full Dataset	0.1799	0.6061	0.2037	0.3049
FTW	Subtropic Only	Temp. Only	0.184	0.6194	0.2074	0.3107
FTW	Subtropic Only	Subtropic Only	0.161	0.5463	0.1859	0.2774
FTW	Temperate Only	Full Dataset	0.1197	0.6994	0.1262	0.2138
FTW	Temperate Only	Temp. Only	0.1234	0.7174	0.1297	0.2197
FTW	Temperate Only	Subtropic Only	0.102	0.6104	0.1091	0.1852
FTW	Subs. Dataset	Full Dataset	0.3208	0.6237	0.3978	0.4857
FTW	Subs. Dataset	Temp. Only	0.331	0.632	0.41	0.4974
FTW	Subs. Dataset	Subtropic Only	0.271	0.5782	0.3378	0.4264

Table 13: FTW results by THZ Class on Hedgementation. All datasets after the first group are random subsets matching the size of the subtropic only dataset (the smallest subset).

F Supplementary Results: Performance in Agricultural Geographies

To distinguish agricultural areas from non-agricultural areas, we use data from the *Registre Parcellaire Graphique* (RPG), which contains spatial polygons demarcating agricultural parcels across more than 90% of the agricultural zone (Cantelaube & Carles, 2014). By overlapping our patches

Pixels	Total Pixels	% of Total	Hedgerow Pixels	% of Hedgerow
All	49,070,080	100%	2,430,758	100%
Agri. + 10M Buffer	32,737,848	67%	2,010,352	83%
Agricultural	31,143,086	63%	1,751,315	71%
Non-Agri.	17,926,994	36%	679,443	28%
Non-Agri. - 10M Buffer	16,332,232	33%	420,406	17%

Table 14: Tables showing the proportion of all pixels, and hedgerow pixels specifically, included by agricultural masks.

Figure 5: The construction of the RPG agriculture mask. The mask is shown in yellow, and the 10-m buffer is shown in orange.

with these polygons and rasterizing the result into a binary mask, we can assign every pixel in our dataset as either agricultural or non-agricultural. This information is relevant for multiple reasons. First, we expect the BD Haies labels to have higher quality in agricultural zones, where they are validated by individual farmers. In addition, hedgerows are prevalent in agricultural areas, and particularly important to monitor there. This lets us focus on our models' error in agricultural zones explicitly, and measure any distribution differences in hedgerow labels and their detection through satellite images in agricultural vs. non agricultural landscapes.

Since many hedgerows are found adjacent to agricultural areas rather than directly within them, we also experiment with buffering the border of all agricultural polygons by a constant 10 meters before rasterizing, in order to include such hedgerows.

We see that 63%, a significant majority, of pixels in our dataset are classified as falling within agricultural areas when referencing the RPG data. The proportion is even larger when we consider specifically hedgerow pixels, of which 71% are found in agricultural areas. Hedgerow pixels are also very likely to be situated directly on the border of agricultural areas: adding the 10 meter buffer increases overall pixel coverage by 4%(from 63% to 67%), but hedgerow coverage by 12% (from 71% to 83%).

To assess variation in hedgerow segmentation performance in agricultural vs. non-agricultural landscapes, we evaluate our models on these subsets of the test set (with and without buffers), and compare the metrics. We also assess the importance of agricultural vs. non-agricultural pixels to training by training models exclusively on those subsets (in this case, we mask the loss of excluded pixels by multiplying it by zero), and comparing their results. This however introduces a confounder as it varies the size of the training set, which we observe the have a significant impact on performance (Appendix D). We also train models by randomly masking out the loss from pixels until the number of pixels used is the same as the agriculture-only experiment. This way, all models "see" the same number of pixels during training, and we can better isolate the effects to the distribution of training pixels.

F.1 PASTIS UTAE

Table 16 shows the performance of models trained on a mix of all pixels or just agricultural pixels, on all subsets of pixels. Table 16 focuses on training on non-ag pixels. We make two main observations.

First, training on all pixels always yields better results, though the effect is small compared to ag only pixels. Since the size of dataset matters, it seems like the best strategy is to train on all available data.

Model	Trained On	Tested On	IoU	Presc.	Recall	F1
PASTIS	All Pixels	All Pixels	0.416	0.538	0.646	0.587
PASTIS	All Pixels (subs.)	All Pixels	0.41	0.507	0.68	0.581
PASTIS	Agri. Pixels	All Pixels	0.378	0.448	0.706	0.548
PASTIS	Agri. + Buffer	All Pixels	0.394	0.492	0.665	0.566
PASTIS	All Pixels	Agri. Pixels	0.462	0.576	0.7	0.632
PASTIS	All Pixels (subs.)	Agri. Pixels	0.459	0.555	0.726	0.629
PASTIS	Agri. Pixels	Agri. Pixels	0.457	0.545	0.739	0.627
PASTIS	Agri. + Buffer	Agri. Pixels	0.452	0.555	0.709	0.622
PASTIS	All Pixels	Agri. + Buffer	0.457	0.572	0.694	0.627
PASTIS	All Pixels (subs.)	Agri. + Buffer	0.454	0.55	0.722	0.624
PASTIS	Agri. Pixels	Agri. + Buffer	0.451	0.538	0.737	0.622
PASTIS	Agri. + Buffer	Agri. + Buffer	0.447	0.551	0.704	0.618
PASTIS	All Pixels	Non-Agri. Pixels	0.303	0.433	0.502	0.465
PASTIS	All Pixels (subs.)	Non-Agri. Pixels	0.297	0.39	0.556	0.458
PASTIS	Agri. Pixels	Non-Agri. Pixels	0.243	0.286	0.617	0.391
PASTIS	Agri. + Buffer	Non-Agri. Pixels	0.273	0.354	0.546	0.429
PASTIS	All Pixels (subs.)	Non-Agri. - Buffer	0.234	0.356	0.406	0.379
PASTIS	All Pixels (subs.)	Non-Agri. - Buffer	0.232	0.316	0.466	0.376
PASTIS	Agri. Pixels	Non-Agri. - Buffer	0.179	0.21	0.549	0.304
PASTIS	Agri. + Buffer	Non-Agri. - Buffer	0.206	0.27	0.465	0.342

Table 15: Results across different pixel sets for models trained on all pixels, Agri. pixels, and Agri. pixels + Buffer.

Model	Trained On	Tested On	IoU	Presc.	Recall	F1
PASTIS	All Pixels(subs.)	All Pixels	0.39	0.483	0.67	0.561
PASTIS	Non-Agri. Pixels	All Pixels	0.205	0.232	0.641	0.34
PASTIS	Non-Agri. - Buffer	All Pixels	0.242	0.3	0.555	0.39
PASTIS	All Pixels (subs.)	Agri. Pixels	0.435	0.527	0.713	0.606
PASTIS	Non-Agri. Pixels	Agri. Pixels	0.191	0.209	0.689	0.321
PASTIS	Non-Agri. - Buffer	Agri. Pixels	0.236	0.281	0.593	0.382
PASTIS	All Pixels (subs.)	Agri. + Buffer	0.431	0.524	0.709	0.603
PASTIS	Non-Agri. Pixels	Agri. + Buffer	0.204	0.225	0.683	0.339
PASTIS	Non-Agri. - Buffer	Agri. + Buffer	0.204	0.225	0.683	0.339
PASTIS	All Pixels (subs.)	Non-Agri. Pixels	0.287	0.374	0.553	0.446
PASTIS	Non-Agri. Pixels	Non-Agri. Pixels	0.278	0.379	0.511	0.435
PASTIS	Non-Agri. - Buffer	Non-Agri. Pixels	0.267	0.393	0.453	0.421
PASTIS	All Pixels (subs.)	Non-Agri. - Buffer	0.225	0.301	0.472	0.368
PASTIS	Non-Agri. Pixels	Non-Agri. - Buffer	0.215	0.303	0.425	0.354
PASTIS	Non-Agri. - Buffer	Non-Agri. - Buffer	0.21	0.314	0.386	0.347

Table 16: Results across different pixel sets for models trained on all pixels, Non-Agri. pixels, and Non-Agri. pixels - Buffer.

Second, hedgerows in agricultural landscapes are significantly easier to map: the IoU of the PASTIS U-TAE model reaches 45.9% of IoU on agricultural pixels, up from 41% on all pixels. This is a valuable and encouraging findings for applications that focus on these landscapes.

F.2 Fields of The World

A similar experiment with the FTW model yields qualitatively consistent findings, as shown on Table 17. All IoUs are lower though, as the model is less performant that the PASTIS U-TAE on this task.

Model	Trained On	Tested On	IoU	Prec	R	F1
FTW	All Pixels	All Pixels	0.333	0.607	0.425	0.500
FTW	Agri. + Buffer	All Pixels	0.310	0.608	0.388	0.474
FTW	Agri. - Buffer	All Pixels	0.326	0.601	0.415	0.491
FTW	Non-Agri. + Buffer	All Pixels	0.126	0.728	0.132	0.224
FTW	Non-Agri. - Buffer	All Pixels	0.201	0.648	0.226	0.335
FTW	All Pixels	Agri. + Buffer	0.367	0.639	0.462	0.537
FTW	Agri. + Buffer	Agri. + Buffer	0.345	0.658	0.421	0.513
FTW	Agri. - Buffer	Agri. + Buffer	0.360	0.655	0.444	0.529
FTW	Non-Agri. + Buffer	Agri. + Buffer	0.140	0.752	0.146	0.245
FTW	Non-Agri. - Buffer	Agri. + Buffer	0.222	0.671	0.249	0.363
FTW	All Pixels	Non-Agri. + Buffer	0.173	0.403	0.233	0.296
FTW	Agri. + Buffer	Non-Agri. + Buffer	0.156	0.350	0.221	0.271
FTW	Agri. - Buffer	Non-Agri. + Buffer	0.180	0.356	0.268	0.306
FTW	Non-Agri. + Buffer	Non-Agri. + Buffer	0.059	0.528	0.062	0.111
FTW	Non-Agri. - Buffer	Non-Agri. + Buffer	0.097	0.465	0.110	0.177
FTW	All Pixels	Agri. - Buffer	0.373	0.648	0.468	0.544
FTW	Agri. + Buffer	Agri. - Buffer	0.352	0.668	0.426	0.521
FTW	Agri. - Buffer	Agri. - Buffer	0.367	0.669	0.448	0.537
FTW	Non-Agri. + Buffer	Agri. - Buffer	0.143	0.760	0.150	0.250
FTW	Non-Agri. - Buffer	Agri. - Buffer	0.224	0.675	0.251	0.366
FTW	All Pixels	Non-Agri. - Buffer	0.231	0.482	0.307	0.375
FTW	Agri. + Buffer	Non-Agri. - Buffer	0.210	0.444	0.284	0.346
FTW	Agri. - Buffer	Non-Agri. - Buffer	0.229	0.437	0.325	0.373
FTW	Non-Agri. + Buffer	Non-Agri. - Buffer	0.081	0.609	0.086	0.150
FTW	Non-Agri. - Buffer	Non-Agri. - Buffer	0.140	0.553	0.158	0.246

Table 17: FTW model results across different training and testing pixel masks.