FLAIR: a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery

Anatol Garioud[1] **Nicolas Gonthier**[2] **Loic Landrieu**[2,3] **Apolline De Wit**[1]
Marion Valette[1] **Marc Poupée**[4] **Sébastien Giordano**[1] **Boris Wattrelos**[1]

[1]French National Institute of Geographical and Forest Information (IGN)
[2]Univ Gustave Eiffel, IGN, ENSG, LASTIG
[3]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS
[4]National School of Geographic Sciences (ENSG)
{firstname.lastname}@ign.fr

A Appendix

A.1 Benefits of the multi-source approach

Sentinel-2 imagery are synergistic approach with VHR aerial images for land cover mapping, as each source has a unique advantage allowing to distinguishing nuanced semantic classes, a critical need in detailed geospatial analysis. Some of the main benefits of integrating Sentinel-2 are:

- **Increased spectral resolution:** Unlike aerial acquisitions that generally contain only four spectral bands (with a single one in the infrared), Sentinel-2 is furnished with a 10-band multispectral imager. This includes bands in the near-infrared spectrum, which prove essential for discerning vegetation phenology [1].
- **Multi-temporal resolution:** Sentinel-2 provides a consistent yearly time series. This capability allows our model to trace the temporal progression of each pixel's spectral response, proving invaluable in distinguishing between similar plant species, as depicted in Figure 2. As an illustrative example, while an “agricultural land” and a “herbaceous surface” might appear identical during specific times (exhibiting low herbaceous vegetation), the agricultural land remains barren of vegetation during other parts of the year. VHR aerial acquisitions, in contrast, are limited to single-date images.
- **Larger spatial context:** The coarser spatial resolution of Sentinel-2 (10 m) compared to aerial images (20 cm) provides an unexpected advantage. By offering a broader context, Sentinel-2 enables our model to harness wider receptive fields. Consequently, each 102x102m aerial patch is linked with a Sentinel-2 image time series spanning a 400x400m area.
- **Spectral Consistency:** The Sentinel-2 time series benefits from consistent spectral calibration, which aids in countering the radiometric inconsistencies introduced during the BD Ortho's correction process.

A.2 Sentinel-2 Time Series

Table 1 indicates the original bands acquired by the Sentinel-2 satellites and considered in the FLAIR dataset. The images were downloaded from the Sinergise API [2] as Level-2A products (Bottom-Of-

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

the-Atmosphere reflectances) which are atmospherically corrected using the Sen2Cor algorithm [3]. [1] Sentinel-2 sensor acquires images at 10, 20 and 60 m spatial resolutions. The 60 m bands mainly intended for atmospheric corrections are not taken into account and the 20 m bands are resampled during data retrieval to 10 m by the nearest interpolation method.

Table 1: Sentinel-2 spatial and spectral resolutions. Original spatial and spectral resolutions of Sentinel-2 images along with the correspondence between original band number and the distributed data.

Original Band number	FLAIR band number	Central wavelength (nm)	Bandwidth (nm)	Original Spatial resolution (m)	FLAIR Spatial resolution (m)
2	1	490	65	10	10
3	2	560	35	10	10
4	3	665	30	10	10
5	4	705	15	20	10
6	5	740	15	20	10
7	6	783	20	20	10
8	7	842	115	10	10
8a	8	865	20	20	10
11	9	1610	90	20	10
12	10	2190	180	20	10

Table 2 indicates the cloud & snow probability masks provided as separate files alongside the Sentinel-2 acquisitions. It should be noted that cloud detection in satellite images is a complex task because of the diversity of clouds (thin, scattered clouds). As a result, probability masks can contain errors, notably confusion with surfaces with a high albedo and close to the top of a cloud, as is the case with the roofs of industrial buildings.

Table 2: Provided cloud and snow masks.

Mask	FLAIR band number	Original Spatial resolution (m)	FLAIR Spatial resolution (m)
Snow probability (SNW)	1	20	10
Cloud probability (CLD)	2	20	10

Table 3 provides information about the number of dates included in the filtered Sentinel-2 time series for the train and test datasets. On average, each area is acquired on 55 dates over the course of a year by satellite imagery.

Table 3: Sentinel-2 Time series length. Number of acquisitions (dates) in the Sentinel-2 times series of one year (corresponding to the year of aerial imagery acquisition).

	acquisitions per super-area		
Sentinel-2 time series (1 year)	min	max	mean
train dataset	20	100	55
test dataset	20	114	55

Note that cloudy dates are not suppressed from the time series. Instead, the masks are provided and can be used to filter the cloudy dates if needed.

The spatial size of Sentinel-2 time series has been empirically determined and set to 40. Nevertheless, we provided in this dataset wider areas than the 40×40 used for our baseline. However, there is a limit of 110 pixels for edge patches. The choice of time series spatial size has an impact on the spatial context provided to both the U-TAE and U-Net branches through the *collapsed* fusion sub-module [5].

[1] More advanced algorithms [4] could be beneficial.

A.3 Semantic classes

Overall semantic class number of pixels and frequency of the FLAIR dataset are provided in Table 4. The class distribution in percentages of the train and test sets are presented in Figure 1. The detailed description of the original semantic classes is provided in Table 5.

The ground truth labels are based on photo-interpretation of the aerial imagery at 20 cm and has been manually produced by experts following a call for tenders from the IGN. An initial spatial multi-level image segmentation approach using PYRAM [6] was applied, simplifying the labeling at the small cluster level. This segmentation was modified interactively when deemed appropriate.

Table 4: Details about the semantic classes of the main nomenclature of the FLAIR dataset and their corresponding label values, frequency in pixels and percentage among the entire dataset.

	Class	Label Value	Pixels	%
	building	1	1,453,245,093	7.13
	pervious surface	2	1,495,168,513	7.33
	impervious surface	3	2,467,133,374	12.1
	bare soil	4	629,187,886	3.09
	water	5	922,004,548	4.52
	coniferous	6	873,397,479	4.28
	deciduous	7	3,531,567,944	17.32
	brushwood	8	1,284,640,813	6.3
	vineyard	9	612,965,642	3.01
	herbaceous vegetation	10	3,717,682,095	18.24
	agricultural land	11	2,541,274,397	12.47
	plowed land	12	703,518,642	3.45
	other	**>13**	153,055,302	0.75

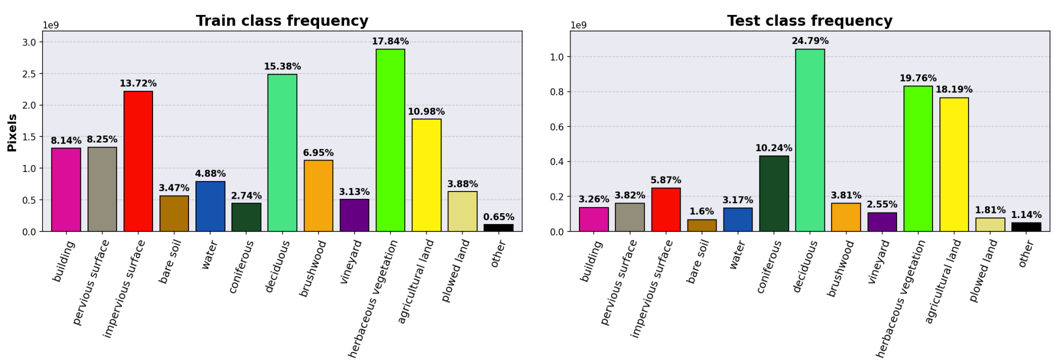

Figure 1: Class distribution of the train dataset (*left*) and test dataset (*right*).

A.4 Aerial imagery and spatial domains

Within a spatial domain, all aerial acquisitions are radiometrically corrected to reduce disparities in sunlight and contrast. Nonetheless, this homogenization is not applied equally across all the different spatial domains as can be seen in Figure 2. As opposed to satellite imagery, the pixel intensity in the image channels can therefore not be considered as a physical measure.

Table 5: Semantic classes of the FLAIR dataset.

Class description

Note: as previously stated, semantic classes are assigned on the cluster level. In a given aerial image, only observable objects are labeled, whereby temporal aspects are not taken into consideration.

Anthropized surfaces without vegetation (1, 2, 3, 13 and 18)

Class 1 – building includes not only buildings but also other types of constructions such as towers, agricultural silos, water towers and dams. Greenhouses (class 18) are an exception.
Class 2 – pervious surface defined as man-made bare soils covered with mineral materials (*e.g.* gravel, loose stones) and considered to be pervious. It includes pervious transport networks (*e.g.* gravel pathways, railways), quarries, landfills, building sites and coastal ripraps.
Class 3 – impervious surface is defined as man-made bare soils that are impervious due to their building materials (e.g. concrete, asphalt, cobblestones). It includes roadways, parking lots, and certain types of sports fields.
Class 13 – swimming pool is defined as man-made artificial (open-air) swimming pools. It is not included in class 5 (water).
Class 18 – greenhouse although it can be considered as a building, is given a distinct label. Greenhouses are a class of their own and are not part of class 1.

Natural areas without vegetation (4, 5 and 14)

Class 4 – bare soil defined as natural permanently bare soils. These natural soils remain without vegetation throughout the year and generally are covered with sand, pebbles, rocks or stones. Examples of natural bare soils are frequently found in coastal, mountainous and forested areas.
Class 5 – water is defined as areas covered by water, such as sea, rivers, lakes and ponds. An exception are swimming pools (class 13).
Class 14 – snow refers to surfaces covered by snow. It is an extremely rare class as the images are taken in the summertime and only very few regions in France are covered with snow year-round.

Woody natural vegetation surfaces (6, 7, 8, 15, 16 and 17)

Class 6 – coniferous, is defined as trees identifiable as coniferous (pines, firs, cedars, cypress trees, ...) and taller than 5 m.
Class 7 – deciduous is defined as trees identifiable as deciduous (oaks, beeches, birches, chestnuts, poplars, ...) and taller than 5 m.
Class 8 – brushwood refers to natural woody surfaces with a vegetation less than 5 m high. It includes short and young trees, brushwood, shrublands, mountain moors and abandoned agricultural lands.
Class 15 – clear-cut, is defined as forest areas, in which the trees have been cut down and harvested.
Class 16 – ligneous is an extremely rare class used to describe forest areas with a homogeneous representation of either coniferous or deciduous trees.
Class 17 – mixed is an extremely rare class used to describe forest areas with heterogeneous trees for which the types of trees (coniferous/ deciduous) cannot be determined with sufficient certainty.

Agricultural surfaces (9, 11 and 12)

Class 9 – vineyard despite being an agricultural use of the land, are assigned a class apart, a reason being their rather distinctive land cover characteristics.
Class 11 – agricultural land encompasses various different agricultural classes. For example, besides major crops, it also includes permanent and temporary grasslands with agricultural use. Vineyards (class 9) are not included in this class.
Class 12 – plowed land is defined as agricultural land with no visible vegetation (*e.g.* recently plowed and freshly harvested land).

Herbaceous surfaces (10)

Class 10 – herbaceous vegetation defines herbaceous surfaces that are not intensively exploited for agriculture purposes. This class includes ornamental lawns (e.g. gardens, public parks), recreational fields (*e.g.* used for sport), natural herbaceous areas in forested or mountainous areas, non-cultivated grass in agricultural areas or along transportation networks.

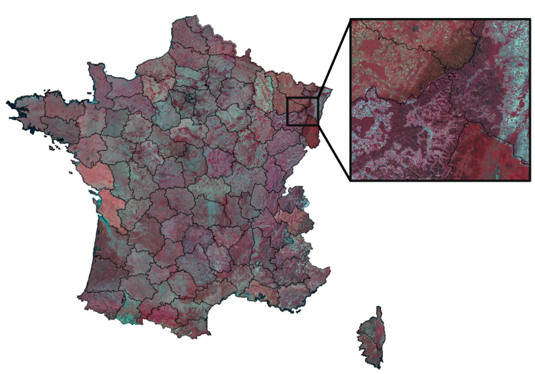

Figure 2: Radiometric discrepancies of the aerial images between domains. The 3 channels image displayed is a composite of Near-Infrared, Red and Green spectral information.

A.5 Benchmark architecture

A.5.1 U-Net (spatial/texture branch)

We choose a U-Net architecture [7] with a ResNet34 encoder backbone (pre-trained on the ImageNet dataset [8]) for a total of ≈ 24.4 M parameters and rely on the implementation available in the *segmentation-models-pytorch* library [9] and trained with the PyTorch lightning [10] framework. The architecture employed is illustrated in Figure 3.

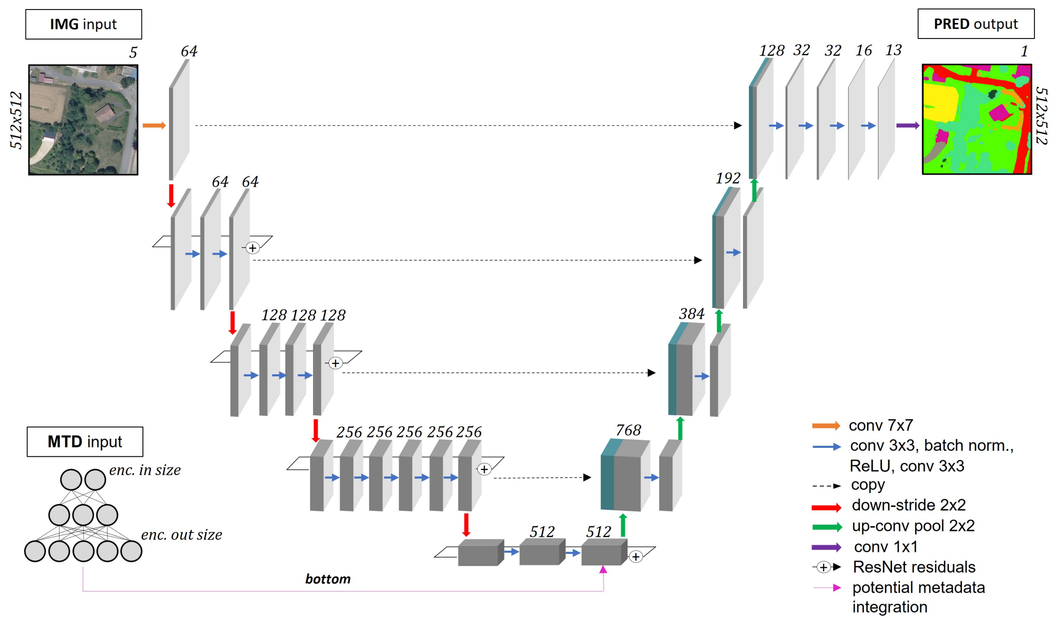

Figure 3: U-Net architecture used for the baseline. **IMG** = input image; **MTD** = input metadata; **PRED** = prediction output. One potential and traditional approach to integrate the metadata would be to add a Multi-layer Perceptron for encoding and add the output to the output of the last layer of the encoder or as an additional band to the IMG input.

Concerning the exploitation of metadata, a simple approach has been tested [11]. The strategies explored have a first step of metadata encoding: positional encoding of spatial and temporal information and one-hot-encoding for camera type and aerial image acquisition year. A shallow MLP with dropout (probability of 0.4) and ReLU activation is then defined to jointly encode the metadata and to provide a specified output size. Subsequently, multiple different integration strategies with the current ResNet34/U-Net segmentation architecture are possible. We have chosen a commonly employed strategy (depicted as '*bottom*' in Figure 3) consists in matching the MLP output size to the output size of the last layer of the ResNet34 encoder. The two vectors (encoded metadata and encoded images) can then be added and fed into the first layer of the architecture's decoder. Strategies following similar approaches that add the MLP encoded output at different positions in the architecture's encoder or decoder parts (*e.g.*, after the first input convolution layer, with the last decoder layer, or even added as a sixth channel to the input image) are possible. A positional encoding of size 32 is used specifically for encoding the geographical location information.

The exploitation of metadata deserves to be studied more by the computer vision community, as it could bring real gains by taking advantage of the specificity of remote sensing data.

A.5.2 Fusion module of the U-T&T model

A Fusion Module is employed within the U-T&T baseline model to integrate the feature maps from satellite time-series (with broader spatial extent) into the feature maps from the aerial imagery branch. The details of this module can be seen in Figure 4. Within the *Fusion Module*, two sub-modules (*cropped* and *collapsed*) have different purposes and focus on distinct aspects: the spatio-temporal information and the spatial context. This *Fusion Module* is applied to match with each feature maps of the U-Net encoder.

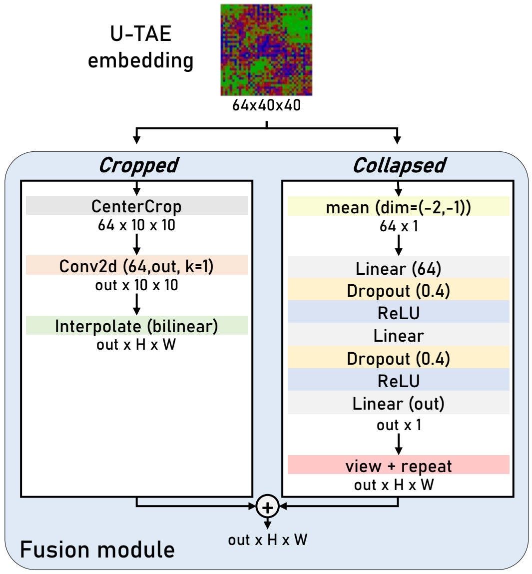

Figure 4: Fusion module. This module takes as input the last U-TAE embeddings. It is applied to each stage of the U-Net encoder feature maps. *out* corresponds to the channel size of the U-Net encoder feature map and *H* and *W* to the corresponding spatial dimensions.

A.5.3 Data augmentation

By introducing variance in the dataset, image data augmentation helps to prevent overfitting and provides trained models with enhanced generalization capabilities. For our baseline, only geometric transformations are explored using the *Albumentation* library. Vertical and horizontal flips, and random rotations of 0, 90, 180 or 270 degrees are tested. A data augmentation probability of 0.5 is used.

A.6 Benchmark results

A.6.1 Official data split of the FLAIR dataset

The following per domain split of the data has been used for the experiments:

`TRAIN:`	D006, D007, D008, D009, D013, D016, D017, D021, D023, D030, D032, D033, D034, D035, D038, D041, D044, D046, D049, D051, D052, D055, D060, D063, D070, D072, D074, D078, D080, D081, D086, D091
`VALIDATION:`	D004, D014, D029, D031, D058, D066, D067, D077
`TEST:`	D015, D022, D026, D036, D061, D064, D068, D069, D071, D084

A.6.2 Extra results

Figure 5 illustrates the confusion matrix of the best U-T&T model. This confusion matrix is derived by combining all individual confusion matrices per patch and is normalized by rows. The analysis of the confusion matrix shows that the best U-T&T model achieves accurate predictions with minimal confusion in the majority of classes. However, when it comes to natural areas such as *bare soil* and *brushwood*, although there is improvement due to the use of Sentinel-2 time series data, a certain level of uncertainty remains. These classes exhibit some confusion with semantically similar classes, indicating the challenge of accurately distinguishing them.

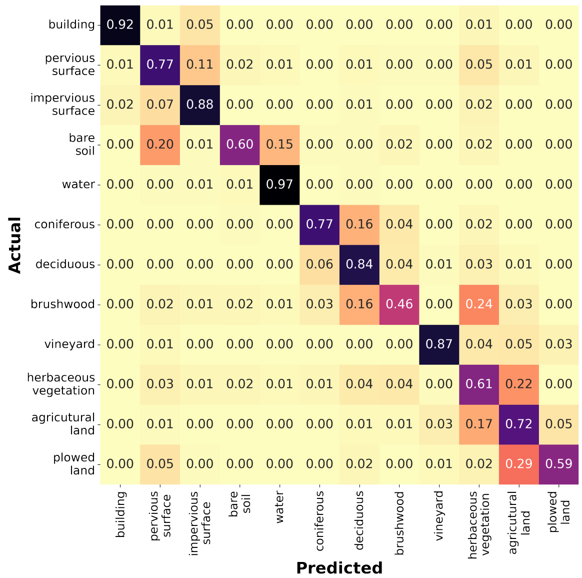

Figure 5: U-T&T best model confusion matrix of the test dataset. The matrix is normalized by rows.

More qualitative examples can be found in Figure 6.

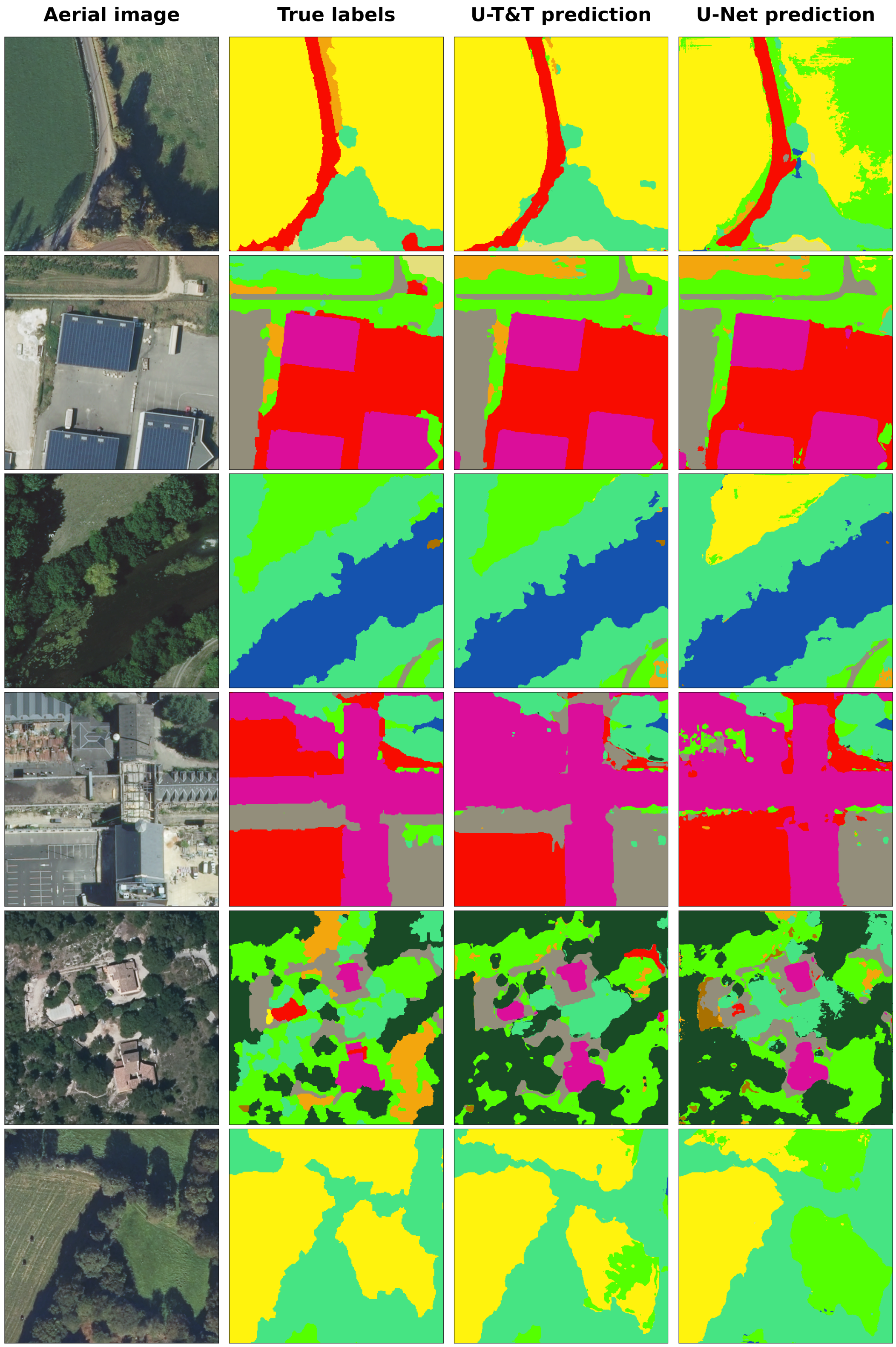

Figure 6: Illustration of patch-wise results. Random results on the FLAIR dataset for the multimodal approach U-T&T than the standard U-Net model.

References

[1] Félix Quinton and Loic Landrieu. Crop rotation modeling for deep learning-based parcel classification from satellite time series. *Remote Sensing*, 2021.

[2] Sentinel Hub, Sinergise Ltd. https://www.sentinel-hub.com. [Online; Accessed: 02-May-2023].

[3] Magdalena Main-Knorn, Bringfried Pflug, Jerome Louis, Vincent Debaecker, Uwe Müller-Wilm, and Ferran Gascon. Sen2Cor for Sentinel-2. In *Image and Signal Processing for Remote Sensing XXIII*, 2017.

[4] SentinelHub. Sentinel2 Cloud Detector. https://github.com/sentinel-hub/sentinel2-cloud-detector, 2023. [Online; accessed 10-August-2023].

[5] Anatol Garioud, Apolline De Wit, Marc Poupée, Marion Valette, Sébastien Giordano, and Boris Wattrelos. Flair #2: textural and temporal information for semantic segmentation from multi-source optical imagery. *arXiv*, 2023.

[6] Laurent Guigues, Jean Pierre Cocquerez, and Hervé Le Men. Scale-sets Image Analysis. *International Journal of Computer Vision*, 2006.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 2009.

[9] Pavel Iakubovskii. Segmentation models Pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[10] PyTorch Lightning. https://www.pytorchlightning.ai. [Online; Accessed: 10-October-2022].

[11] Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. Flair #1: semantic segmentation and domain adaptation dataset. *arXiv*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4 and Annexes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sections 4 and 3 and Annexes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5, we report the standard deviation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 7 about Sentinel-2 data.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 4
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section A.1
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 6
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

B Datasheet for FLAIR dataset

B.1 Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a particular gap that needed to be filled? Please provide a description.

- The FLAIR dataset is created to train and evaluate models that can predict very-high-resolution land cover maps from diverse data sources with heterogeneous spatial, temporal, and spectral resolutions. The main gap we are addressing is the lack of large-scale data with high-definition annotations.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

- This dataset is presented by the French National Institute of Geographical and Forest Information (IGN), a French public state administrative establishment aiming to produce and maintain geographical information for France. The IGN has the mission to document and measure land-cover on French territory and provides referential geographical datasets, including very-high-resolution aerial images and topographic maps. IGN produces reference data and carries out innovation, research and teaching activities. As part of its innovation activities, the IGN provides the FLAIR dataset to democratize access to large-scale open powerful machine learning models through the research and development of open-source resources.

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

- The funding of the FLAIR dataset is 100% public. This work was sponsored by the Ministry of Ecological Transition (more specifically the Directorate for Planning, Housing and Nature *Direction générale de l'aménagement, du logement et de la nature*) and the Fund for the transformation of public action (*Fonds pour la transformation de l'action publique*) from the Minister of the Civil Service. The IGN is funded by the French Ministry of Ecological Transition and the French Ministry of Agriculture.

Q4 **Any other comments?**

- [N/A]

B.2 Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

- We provide aerial image with corresponding land cover segmentation along with Sentinel-2 satellite image time series around each aerial patch. The acquisitions are taken from 916 unique areas distributed across 50 French spatial domains (**départements**), covering approximately 817 km^2. The test labels will be released at the end of the second challenge hosted on CodaLab. We made our baseline codes openly available on the FLAIR GitHub page (`https://github.com/IGNF/FLAIR-2-AI-Challenge`).

Q6 **How many instances are there in total (of each type, if appropriate)?**

- We provide 77,762 triplet aerial image, Sentinel-2 time and land cover segmentation. The FLAIR dataset encompasses 20,384,841,728 annotated pixels at a spatial resolution of 0.20 m from aerial imagery with a 19 classes land cover. For each area, a comprehensive one-year record of Sentinel-2 acquisitions is also provided. A further overview of the statistics may be seen in the following annexes.

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

- The FLAIR dataset, derived from a larger dataset obtained by IGN for cartographic production upon the request of the French government, serves as a representative sample encompassing approximately one-third of the available data. While the complete dataset covers 64 spatial domains, the FLAIR dataset focuses on 50 domains by excluding contiguous domains and intra-domain areas. Nevertheless, the selected 50 domains offer comprehensive representation in terms of land cover classes, acquisition dates, and macro-climates, and encompass the metadata associated with the entire dataset. The expertise of IGN was leveraged to ensure the selection of a dataset that is representative and informative.

Q8 **What data does each instance consist of?**

- Each instance consists of an aerial image. Each image is 512×512 in size with a resolution of 20 cm per pixel, and feature 4 spectral channels: red, blue, green, and near-infrared along with an elevation value as fifth channel. Each patch is associated with a satellite image time series from the Sentinel-2 constellation (Drusch et al., 2012) of size 40×40 with a 10 m pixel resolution, centered around the aerial image. Each pixel from the Sentinel-2 sequences is characterized by 10 spectral bands.

Q9 **Is there a label or target associated with each instance?**

- [Yes] We provide a complete pixel-precise land cover segmentation per image (19 classes).

Q10 **Is any information missing from individual instances?**

- [No]

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

- [No]

Q12 **Are there recommended data splits (e.g., training, development/validation, testing)?**

- Yes, we provide data splits for reproducing the results of the baselines. The test split has been explicitly selected to address the complex domain shifts of geospatial data.

Q13 **Are there any errors, sources of noise, or redundancies in the dataset?**

- As the annotations are made through visual interpretation with quality control, some errors are unavoidable, especially for classes that are visually hard to distinguish. Internal quality control with multiple annotations has been performed to limit such errors. There are no redundancies in the dataset, each image covers a distinct area.

Q14 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

- This dataset is self-contained and will be stored and distributed by the IGN, a public institute. The dataset is under the Open Licence 2.0 of Etalab.

Q15 **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?**

- [No] . The building class does not contain information that would not be available in other open-access sources, such as the cadaster. We have specifically avoided high-risk areas such as military installations or nuclear plants.

Q16 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

- [No]

Q17 **Does the dataset relate to people?**

- The dataset may feature pedestrian or individuals, but the resolution of 20cm/pixel and the aerial perspective is not sufficient to recognize them uniquely.

Q18 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

- [No]

Q19 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

- [No] . The resolution of 20cm/pixel and the aerial perspective is insufficient to recognize them uniquely.

Q20 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

- [No]

Q21 **Any other comments?**

- [No]

B.3 Collection Process

Q22 **How was the data associated with each instance acquired?**

- The aerial images are sampled from the ORTHO HR® imagery collection. It is a mosaic of all the individual images taken during an aerial survey done by IGN and mapped onto a cartographic coordinate reference system. The individual images are projected to the RGE ALTI® DTM, which provides solely the altitude of the ground.
- The Sentinel-2 time series were downloaded from the Sinergise Sentinel-Hub API as Level-2A products (see annexes for more information).

Q23 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

- The IGN selected several acquisition companies through a call for tender with strict specifications.

Q24 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

- The sampling strategy involved class frequency, acquisition dates distribution, radiometric histogram analysis and geographical location spread. The final sampling based on these comprehensive variables was made manually by experts at the IGN.

Q25 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

- IGN contracted geography experts from the private sector selected through a public call for tender to annotate the dataset. The quality control of the dataset was carried out by geography experts affiliated with IGN. The creation of the dataset was facilitated by researchers and developers employed by IGN under their work contracts.

Q26 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**

- The collection of aerial imagery spanned from 2018 to 2021, which coincides with the duration required for an aerial survey to encompass the entirety of the French territory. Annotations were then applied to the aerial images, aligning with the same time frame. Subsequently, the dataset was created in 2022 after the final processing for both the aerial imagery and annotations.

Q27 **Were any ethical review processes conducted (e.g., by an institutional review board)?**

- [No]

Q28 **Does the dataset relate to people?**

- [No]

Q29 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

- [N/A]

Q30 **Were the individuals in question notified about the data collection?**

- [N/A]

Q31 **Did the individuals in question consent to the collection and use of their data?**

- [N/A]

Q32 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

- [N/A]

Q33 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**

- [No]

Q34 **Any other comments?**

- [No]

B.4 Preprocessing, Cleaning, and/or Labeling

Q35 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

- [No]

Q36 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

- [No]

Q37 **Is the software used to preprocess/clean/label the instances available?**

- [No]

Q38 **Any other comments?**

- [No]

B.5 Uses

Q39 **Has the dataset been used for any tasks already?**

- The optical images of FLAIR train split were used for two data challenges ran in 2022 and 2023 by IGN.
- Marsocci et al., 2023 used a subset of FLAIR for to evaluate techniques for unsupervised domain adaptation.

Q40 **Is there a repository that links to any or all papers or systems that use the dataset?**

- [Yes] . We propose below a list of scientific publications and systems that use FLAIR dataset:

    - Garioud et al., 2022 provides a technical description of the FLAIR aerial imagery dataset;
    - Garioud et al., 2023 provides insight on the multimodal fusion of aerial and satellite imagery;
    - Marsocci et al., 2023 experiments remote sensing unsupervised domain adaptation using geographical coordinates on a subset of the FLAIR dataset.

Q41 **What (other) tasks could the dataset be used for?**

- We encourage future researchers to use FLAIR dataset for several tasks. Particularly, we see applications in land cover segmentation and multimodal fusion. Due to the breadth of the data, it also offers a unique opportunity for pre-training of models for other geospatial analysis tasks with low resource, such as object detection, super-resolution, or change detection.

Q42 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

- This dataset is geographically limited to metropolitan France. Although France's territory is quite diverse, featuring oceanic, continental, Mediterranean, and mountainous bioclimatic regions, it does not contain tropical or desert areas.
- The FLAIR dataset's reliance on purely optical data may limit the applicability of the models trained on it to regions with pervasive cloud cover.

Q43 **Are there tasks for which the dataset should not be used?**

- [No] .

Q44 **Any other comments?**

- [No] .

B.6 Distribution

Q45 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

- [Yes] the dataset will be open-source.

Q46 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?**

- The data will be available through *.zip* files available on the FLAIR project page hosted on GitHub (`https://ignf.github.io/FLAIR/`).

Q47 **When will the dataset be distributed?**

- All data with the exception of the test split is presently accessible by registering for an ongoing challenge hosted on Codalab. The entire dataset, including the test split, will be released under an open-source license on the FLAIR project page in early October 2023.

Q48 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- [Yes] . The data is governed by the Open Licence 2.0 of Etalab (`https://www.etalab.gouv.fr/wp-content/uploads/2018/11/open-licence.pdf`).

Q49 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

- [No]

Q50 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

- [No]

Q51 **Any other comments?**

- [No]

B.7 Maintenance

Q52 **Who will be supporting/hosting/maintaining the dataset?**

- IGN will support hosting of the dataset and metadata.

Q53 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

- `ai-challenge@ign.fr`

Q54 **Is there an erratum?**

- [No] . There is no erratum for our initial release. Errata will be documented as future releases on the dataset website.

Q55 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

- Additional modalities (*e.g.*, supplementary satellite, aerial, UAV-based imagery) may be added to the FLAIR dataset.

Q56 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

- N/A

Q57 **Will older versions of the dataset continue to be supported/hosted/maintained?**

- [Yes] . We are dedicated to providing ongoing support for the FLAIR dataset.

Q58 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

- Proposed extensions or corrections to the FLAIR dataset may be submitted to the providers for consideration. The IGN will assess the feasibility of incorporating the suggested modifications, considering factors such as data licensing, maintenance requirements, and relevance.

Q59 **Any other comments?**

- [No] .