

FRACTAL: An Ultra-Large-Scale Aerial Lidar Dataset for 3D Semantic Segmentation of Diverse Landscapes

Charles Gaydon

Michel Daab

Floryne Roche

Institut national de l'information géographique et forestière (IGN), France

fractal.dataset@ign.fr

Abstract

Mapping agencies are increasingly adopting Aerial Lidar Scanning (ALS) as a new tool to map buildings and other above-ground structures. Processing ALS data at scale requires efficient point classification methods that perform well over highly diverse territories. Large annotated Lidar datasets are needed to evaluate these classification methods, however, current Lidar benchmarks have restricted scope and often cover a single urban area. To bridge this data gap, we introduce the FRench ALS Clouds from TArgeted Landscapes (FRACTAL) dataset: an ultra-large-scale aerial Lidar dataset made of 100,000 dense point clouds with high quality labels for 7 semantic classes and spanning 250 km². FRACTAL achieves high spatial and semantic diversity by explicitly sampling rare classes and challenging landscapes from five different regions of France. We describe the data collection, annotation, and curation process of the dataset. We provide baseline semantic segmentation results using a state of the art 3D point cloud classification model. FRACTAL aims to support the development of 3D deep learning approaches for large-scale land monitoring.

1. Introduction

High-density Aerial Lidar Scanning (ALS) is now recognized as a powerful remote sensing modality to support important public actions such as ecological monitoring [17] and risk management (e.g., for floods [18] and forest fires [26] prevention). Detailed 3D mapping also contribute to the consolidation of existing geographic databases (e.g., by updating building footprints). The reduction in costs associated with the acquisition, storage, processing, and dissemination of Aerial Lidar Scanning (ALS) data has paved the way for its production at unprecedented scales.

In recent years, ALS has been increasingly used by public authorities at the regional and national levels. In a comprehensive report, [14] describe the availability of non-

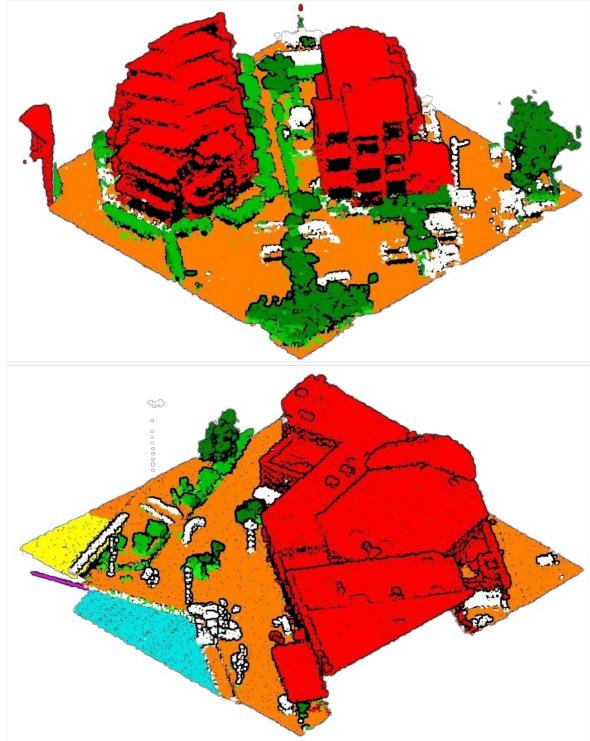


Figure 1. Random scenes from our FRACTAL dataset. Each point is colored according to its semantic class; ground (orange), vegetation (greens), building (red), water (cyan), bridge (yellow), permanent structure (purple), other (white).

commercial Lidar data across Europe: at least a dozen European countries have already performed nationwide Lidar acquisitions, with densities from 10 to 40 pts/m². In France, the national mapping agency (*Institut national de l'information géographique et forestière, IGN*) aims to map the entire French territory with high-density ALS point clouds (10 pulses/m², about 40 pts/m²) by 2026, in a program called Lidar HD (for "High Density") [10].

Point cloud classification is a prerequisite to most downstream use of Lidar products. For example, ground points

must be identified to produce accurate Digital Terrain Models (DTMs) and separated from vegetation points to normalize tree heights in forest applications, e.g., for forest biomass estimation. Buildings are reconstructed from building points for urban modeling, while city trees are mapped from clusters of vegetation points to manage urban biodiversity.

The complexity, volume, and lack of structure of Lidar data make manual labelling of Lidar data time and labor-intensive; existing commercial software all require some form of manual input to reach satisfying results [27]. Deep learning methods are increasingly available, especially since the publication of PointNet [20] in 2016, and of PointNet++ [21] a year later.

Researchers need large annotated Lidar datasets to develop and evaluate their 3D deep learning methods. This need motivates us to release the FRench ALS Clouds from TArgeted Landscapes, or FRACTAL. FRACTAL is a large dataset for the semantic segmentation of ALS data, that spans 250 km². FRACTAL is made of 100,000 distinct point clouds, each spanning 50 × 50 m and typically containing 20k to 40k labelled points (37 pts/m² on average).

The dataset is built from the French Lidar HD data [10], using a sampling scheme that explicitly concentrates rare classes, rare objects, and challenging landscapes. Being sampled from an initial area of 17,440 km² in 5 French regions, it is characterized by a high variety of scenes.

Point clouds are labelled with seven common semantic classes: ground, vegetation, building, water, bridge, permanent structure, and other. The labels were produced with automated processes and then verified and corrected by Lidar operators to achieve high quality classification. FRACTAL is overall the largest open benchmark dataset for 3D semantic segmentation, and the first to offer the diversity of landscapes inherent to large-scale land monitoring.

We make the following contributions:

- Define a general framework to catalog and sample diverse point clouds from large open ALS archives.
- Introduce FRACTAL: a benchmark dataset for 3D point cloud semantic segmentation, sampled from verified areas of the French Lidar HD data. The sampling has explicit consideration for spatial diversity and landscape diversity, making FRACTAL suitable for the evaluation of point cloud classification methods against the specific challenges of land monitoring.
- Set a baseline evaluation of segmentation performance on the dataset using a state-of-the-art deep point cloud segmentation model, to show its potential for benchmarking.

Section 2 lays out current ALS benchmarks and their shortcomings. Section 3 introduces the data sources used

to create FRACTAL. Section 4 presents the strategy for cataloguing and sampling data patches at scale, and highlights the high diversity of the resulting dataset. Section 5 describes the first experimental results on the benchmark.

2. Related work

ALS differs from ground-based Lidar in critical ways: airborne Lidar has nadir orientation instead of lateral orientation, and a lower but more homogeneous point density [24]. Most importantly, airborne Lidar can be used to collect data at a much larger scale than ground-based Lidar, which adds specific challenges to their classification. Vast territories the size of a country have diverse landscapes and vegetation, along with unique signs of human activity, leading to high intraclass heterogeneity. Additionally, geographic data are characterized by spatial autocorrelation and a long-tail distribution, resulting in a variety of rare scenes that may be spatially concentrated, i.e., globally rare but locally frequent, like wind turbines or greenhouses.

The specific nature and scope of ALS must be addressed in the evaluation of 3D point cloud semantic segmentation methods. Having large representative annotated datasets is critical, however, current benchmark datasets come up short: limited in size and spatial diversity, they cannot attest to the capacity of 3D deep learning models to deal with the challenges of large-scale land monitoring. In a comprehensive review of current Lidar benchmark datasets for 3D point cloud semantic segmentation, authors found 26 datasets [28]; only six of them have ALS data, which we report in Tab. 1 along with their proposed dataset (CENAGIS-ALS) and ours (FRACTAL).

Five of these benchmarks cover only a tiny area, not exceeding 2 km². This reduced scale contrasts with the larger support that other modalities enjoy in land monitoring, e.g., with massive land cover mapping benchmarks such as LandCover.ai [2], which covers 216 km² with VHR aerial images, and FLAIR [6], which spans 817 km² across 50 distinct spatial domains with both VHR aerial images and Sentinel 2 time series.

Furthermore, point densities in these ALS benchmarks is either too low for typical urban applications (3-4 pts/m² in LASDU [25], 4-7 pts/m² in ISPRS Vaihingen [19]) or unreasonably high (275 pts/m² in CENAGIS-ALS [28], 348 pts/m² in DublinCity [29], 800 pts/m² in Hessigheim 3D [16]).

The Dayton Annotated Lidar Earth Scan (DALES) [24] attempts to provide a level of detail more representative of real-world use cases, with 8 semantic classes and a point density of 50 pts/m². It is a significant improvement in scale compared to previous urban ALS benchmarks: covering an area of 10 km², DALES is the largest dataset of its kind.

Unfortunately, all ALS benchmarks mentioned so far, including DALES, are limited to single urban areas. This re-

Table 1. Benchmark datasets for 3D semantic segmentation of ALS data. We include the datasets listed by [28] along with their proposed benchmark dataset CENAGIS-ALS, and our own dataset FRACTAL. ALS: Aerial Lidar Scanning; ULS: Unmanned Lidar Scanning.

Reference	Year	Sensor Platform	Area (km ²)	Classes	Points	Density (pts/m ²)
Vaihingen (ISPRS) [19]	2012	ALS	0.1	9	1.16M	4-7
DublinCity [29]	2015	ALS	2	9 ∈ 7 ∈ 4	260M	348
LASDU [25]	2020	ALS	1.02	5	3.12M	3-4
DALES [24]	2020	ALS	10	8	500M	50
Hessigheim 3D [16]	2021	ULS	0.19	11	74M	800
OpenGF [22]	2021	ALS	47.7	2	542M	11
CENAGIS-ALS [28]	2023	ALS	2	49 ∈ 28 ∈ 7	550M	275
FRACTAL (Ours)	2024	ALS	250	7	9261M	37

sults in adjacent training and test data, which is likely to result in overestimated performance metrics given the high autocorrelation of geographic data. Even with reduced ambition (e.g., only considering urban contexts), it is impossible to draw conclusions about the spatial generalizability of the benchmarked methods with current ALS benchmarks. In addition, considering only a single, small urban area hides the complexity of regional scale land monitoring. Overall, researchers need datasets with better territorial representativeness.

OpenGF [22] is the one ALS benchmark that addresses these shortcomings. It is built from open large-scale ALS archives from 4 countries in 3 continents, via a careful selection of areas with high-quality point annotations in diverse landscapes (i.e., metropolis, small city, village, mountain). OpenGF is largest ALS benchmark to date, with 542M points over 50 km². Authors demonstrate the benefits of leveraging open Lidar assets to quickly create high-quality, diverse, large-scale benchmark datasets for 3D semantic classification. However, OpenGF is tailored for the specific task of ground/non-ground classification and was stripped of the semantic classes used in most applications (e.g., building, vegetation).

In summary, current ALS benchmarks for 3D point cloud segmentation are either limited in both volume and representativeness, and unsuitable to evaluate generalizability of methods, or are large and diverse enough but lack semantic depth (OpenGF). Our proposed dataset, FRACTAL, aims to address these shortcomings with: volume (250 km²), territorial representativeness (sampling from 17,440 km² in 5 distinct spatial domains), reduced influence of spatial autocorrelation (large contiguous test areas), and a semantic depth matching typical land monitoring applications (7 classes).

3. Data Collection

We achieve efficient, large scale dataset creation by leveraging an open ALS data archive: the Lidar HD data. To reflect real-world Lidar processing practices, clouds are colorized from national Very High Resolution (VHR) aerial

imagery from the ORTHO HR database.

3.1. ALS point clouds

The Lidar HD program [10] is a national initiative that aims to provide a 3D description of the French territory by 2026, using high-density ALS (10 pulses/m² or about 40 pts/m²). The data produced as part of this program (i.e., point clouds, Digital Terrain Models, Digital Surface Models) are made available as open data with extensive documentation [11].

The program covers mainland France and its overseas territories for a total of 550,000 km². It consists of four phases: data acquisition, storing, processing, and dissemination, along user support. The data acquisition and processing are sequenced in blocks of 50 × 50 km and must be compatible with a variety of Lidar sensors (Leica, Riegl, and Teledyne/Optech), acquisition seasons (leaf-on or leaf-off), and landscapes (e.g., urban, rural, mountains, seashores, overseas territories). The point clouds are disseminated with semantic segmentation labels from 11 classes: unclassified, ground, vegetation (low, medium, and high), buildings, water, bridge deck, permanent structures, artifact, synthetic. The specification of this nomenclature is detailed in Appendix D.1.

The classification of Lidar HD data comes in two flavors:

Classified Lidar HD Results from a fully automated classification process using commercial software and deep learning models.

Optimized Lidar HD Results from automatic classification followed by manual corrections. The annotations are then audited via visual inspection for approximately 10% of the delivered point clouds. All classification errors are listed and rated for severity, with particular attention to confusions in buildings, bridges and ground. Special importance is given to the exhaustivity of individual buildings, with a minimal recall of 99.9% in high-stakes areas (e.g., flood-prone, urban) and a minimal recall of 99.5% elsewhere. The classification is validated if it meets the requirements for real-world use; otherwise, it undergoes further manual corrections until it reaches the desired quality.

3.2. VHR aerial images

The ORTHO HR® [12] is a mosaic of VHR aerial images acquired during national surveys. The individual images are mapped onto a cartographic coordinate reference system and projected on the RGE ALTI Digital Terrain Model for orthorectification. ORTHO HR images have a high spatial resolution of 0.20 m, and near infrared, red, green, and blue channels. Radiometric processing methods, including equalization and global correction, are applied to obtain the final product.

Aerial surveys are not synchronized with Lidar HD acquisitions and a variable time lag of up to 3 years might separate them. Indeed, the aerial images are renewed every three years, whilst Lidar data are processed as soon as they are acquired. In the interim, buildings may have been constructed and vegetation may have expanded or been cleared. Moving elements such as vehicles are unlikely to be consistent. Also, the appearance of cultivated fields is likely to be inconsistent between modalities due to different seasons of acquisition. Despite these discrepancies, working with colorized point clouds is recommended as it systematically yields better segmentation models. For all practical purposes, we document the year of acquisition of images and point clouds as part of the dataset’s metadata.

4. FRACTAL: the dataset

4.1. Area of interest

At the time of dataset creation, Lidar HD data is available for about half of metropolitan France, mainly in its southern half. To have the highest possible annotation quality, we restrict our sampling to Optimized Lidar HD i.e., to the areas that went through human verification and audit of the classification. Fig. 2 presents the 5 spatial domains we consider, each spanning 3456 km² on average, and at least 100 km distant from one another. The spatial domains compose an area of interest of 17,440 km² in Southern France, acquired with a variety of Lidar sensor (documented in Appendix D.3).

To define a common setting for the benchmark of deep learning models, we reserve an area for testing in each spatial domain. Each of the 5 test areas span 210 km² on average, for a total of 1049 km². Having only a limited number of spatial domains with their own unique characteristics (seashores, mountains, field crops, etc.), test data and train data share the same spatial domains. Test areas are however contiguous, large, and distinct from the train areas. This enables spatial block validation and reduces the influence of spatial autocorrelation on model evaluation, which is critical when evaluating remote sensing solutions [15].

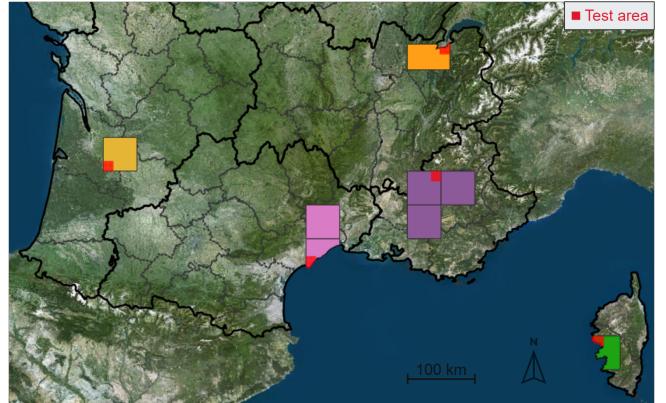


Figure 2. The five spatial domains composing the 17,440 km² of the area of interest. The 1,049 km² reserved to sample the test set are highlighted in red.

4.2. Sampling methodology

The sample unit is a 50 × 50 m square patch, which provides a compromise between detailed structures and long-range spatial dependencies. A square shape facilitates data extraction and visualization of data patches. Also, training models on square point clouds favors the use of non-overlapping tiling at inference time, thus gaining efficiency by limiting redundant predictions.

We adopt the following design principles for sampling:

Class rebalancing: We explicitly address class imbalance, which is often detrimental to model training, by oversampling rarer classes such as water, bridge, and permanent structures.

Uncertainty sampling: We aim to increase the representation of scenes known to be challenging for classification models by oversampling specific landscapes and hard-scapes, such as mountainous areas, seashores, or complex urban scenes.

Spatial sampling as a proxy for landscape diversity: Given the spatial autocorrelation of geographic data, we aim for the broadest spatial distribution of patches. This maximizes diversity within each category of scenes, and globally.

Scenes of special interest need first to be identified as such to consider them for sampling. We define 19 scene types, which we list in Tab. 2. Twelve of them are related to the semantic classification of the point cloud and to the presence of certain objects within a semantic class. The seven other characterize patches in terms of landscapes, using heuristic definitions.

We adopt a simple yet effective sampling scheme consisting of targeted sampling followed by completion sampling:

Table 2. Scene types and their target minimal proportion for targeted sampling.

Motivation	Scene Type	Definition	Target (%)
Classes	BUILD	building ≥ 500 pts	8
	BUILD_GREENHOUSE	greenhouse (BD TOPO)	1 10
	BUILD_BIG	non-residential building (BD TOPO)	1
	BRIDGE	bridge ≥ 50 pts	5 5
	WATER	eau ≥ 50 pts	4 5
	WATER_SURFACE	water area (BD TOPO) & eau ≥ 50 pts	1
	PERMSTRUCT	permanent structure ≥ 50 pts	3
	PERMSTRUCT_PYLON	pylon (BD TOPO) & permanent structure ≥ 50 pts	1 5
	PERMSTRUCT_ANTENNA	antenna (BD TOPO) & permanent structure ≥ 50 pts	1
	OTHER	unclassified ≥ 250 pts	3
Landscapes	OTHER_PARKING	parking lot (BD TOPO) & unclassified ≥ 250 pts	1 5
	OTHER_HIGHWAY	highway (BD TOPO) & unclassified ≥ 400 pts	1
	FOREST	high vegetation $\geq 90\%$ of points	5
	HIGHSLOPE1	$35 \text{ m} \leq \text{elevation gain} < 45 \text{ m}$	2
	HIGHSLOPE2	elevation gain $\geq 45 \text{ m}$	2
	MOUNTAIN	elevation $\geq 1000 \text{ m}$	4 20
	WATER_ONLY	water ≥ 50 pts & ground = 0 pts	1
Seashore	SEASHORE	$-10 \leq \text{elevation} < 10 \text{ & water} \geq 50 \text{ & ground} \geq 100 \text{ pts}$	1
	URBAN	building $\geq 25\%$ of points	5

Targeted sampling: For each scene type, we sample patches with stratification on $1 \times 1 \text{ km}$ Lidar HD tiles, until we reach the target proportion for that scene type. Each sampling is performed independently, with replacement, to ensure that the spatial diversity within each group is optimal; duplicates are then dropped.

Completion sampling: The dataset is completed with stratification on $1 \times 1 \text{ km}$ Lidar HD tiles, this time without consideration for scene type. This is to ensure that we include more ordinary scenes such as field crops and rural areas, which are the majority.

Based on our own experience with training semantic segmentation models on Lidar HD data, we set a target dataset size of 250 km^2 , or 100,000 patches – a 70-fold reduction compared to the initial area of interest. This manageable size is suitable for common research practices and computing infrastructures.

For model benchmarking, we define a reference train/val/test split with an 80/10/10 ratio: 225 km^2 of data go to model training, of which 25 km^2 are reserved for in-training evaluation, and an additional 25 km^2 of data are sampled from test areas and kept for model evaluation. Train and test areas are sampled independently with the same target proportions of each scene type. Patches for in-training validation are sampled from the train areas, again with stratification on Lidar HD tiles. This strategy results in the widest possible spatial distribution for each scene type, and ensures that all landscapes are equally represented in the train, val, and test sets. See Appendix B.2 for the proportions of scene types in each set.

4.3. Spatial and landscape diversity in FRACTAL

Tab. 3 presents key figures of the dataset. Despite being 70 times smaller than the area of interest, the dataset incorporates data from all 17,440 initial Lidar HD tiles, ensuring comprehensive representation of its landscapes. With 9,261M points in 100,000 patches scattered over five vast spatial domains, FRACTAL has unprecedented territorial representativity, spatial diversity, and volume.

Table 3. Size, areas, and number of points in FRACTAL compared to the area of interest. r = ratio.

	Area of interest	\rightarrow	FRACTAL	r
Tiles	17440	\rightarrow	17440	1
Area (km^2)	17440	\rightarrow	250	70
Patches	6976000	\rightarrow	100000	
Points (M)	661998	\rightarrow	9261	71

The target proportions for each scene type are specified in Tab. 2. They are chosen so that about 20% of sampled patches consist of specific landscapes, such as seashores and forests, while about 30% of them contain specific structures and objects, such as highways and water surfaces. Half of the dataset is found by targeted sampling (47,253 patches, or 47.3%), then completed (52,747 patches, or 52.7%) until the target dataset size is reached. The target minimum proportion is achieved for all scene types except one: scenes with permanent structure points and with an antenna in the BD TOPO could not amount to 1% in the train set (target of $n = 1000$ patches, 830 were found) and test

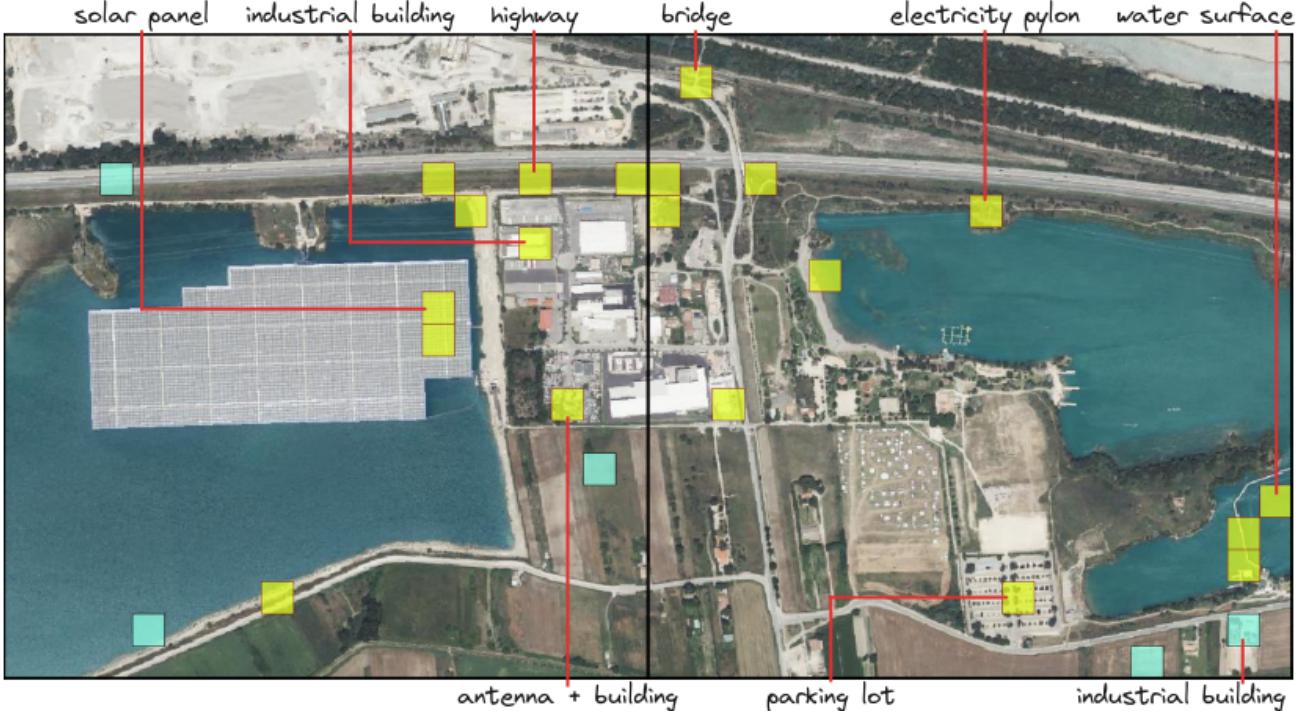


Figure 3. Enlarged view of two Lidar HD tiles and their sampled patches, most of which contain scenes of particular interest.

set (target of $n = 100$ patches, 65 were found).

Fig. 3 presents a visualization of the sampling in two Lidar HD tiles, where 25 patches are selected from the 800 possible patches. Targeted sampling (yellow) over-concentrates specific landscapes such as water surfaces, specific classes such as buildings, and specific human-made structures such as parking lots and electricity pylons. On the other hand, four of the five patches selected by completion sampling (cyan) do not contain any predefined scene type. This illustrates the scarcity of complex scenes in geographical data and the need to explicitly target them to obtain a challenging and diverse dataset. See Appendix B.3 for a similar illustration at a larger scale.

Fig. 4 illustrates how our sampling increases the presence of targeted scene types in FRACTAL. The gain is most important for rarer classes, with up to a 22-fold concentration of scenes with a bridge. All scene types of interest end up with better representation in FRACTAL; except for high-slope scenes: most of them are spatially concentrated in mountainous areas, and thus slightly undersampled by completion sampling.

Additional figures are reported in the appendices: proportions of each semantic class (B.1) and prevalence of scene types in train, val, and test sets (B.2).

4.4. Data extraction and colorization

From the sampled 50×50 m geometric patches, we extract the corresponding ALS point clouds. We colorize point clouds using near infrared, red, green and blue channels from the ORTHO HR images, based on the vertical alignment of points and pixels. Colorization is done regardless of potential obstructions: for instance, a ground point beneath a tree will be colored with the spectral information of the tree above it.

Point clouds are made available in the LAZ 1.4 format defined by the American Society for Photogrammetry and Remote Sensing (ASPRS) [23]. The following dimensions are relevant for semantic segmentation: xyz coordinates in Lambert 93 spatial reference system (EPSG:2154), intensity (16-bits encoding), return number and number of returns, scan angle, and near-infrared, red, green and blue (16-bits encoding). The Lidar HD classification nomenclature, detailed in Appendix D.1, is left intact in the data patches.

Except for its colorization, point clouds in FRACTAL inherit the characteristics of the Lidar HD data, and we refer users to Section 2.2 of their official product description [11] for more information.

4.5. Limitations and recommendations

FRACTAL is sampled from an open ALS archive, and shares its limitations. The dataset only includes data that underwent human corrections and auditing (i.e. ‘Optimized

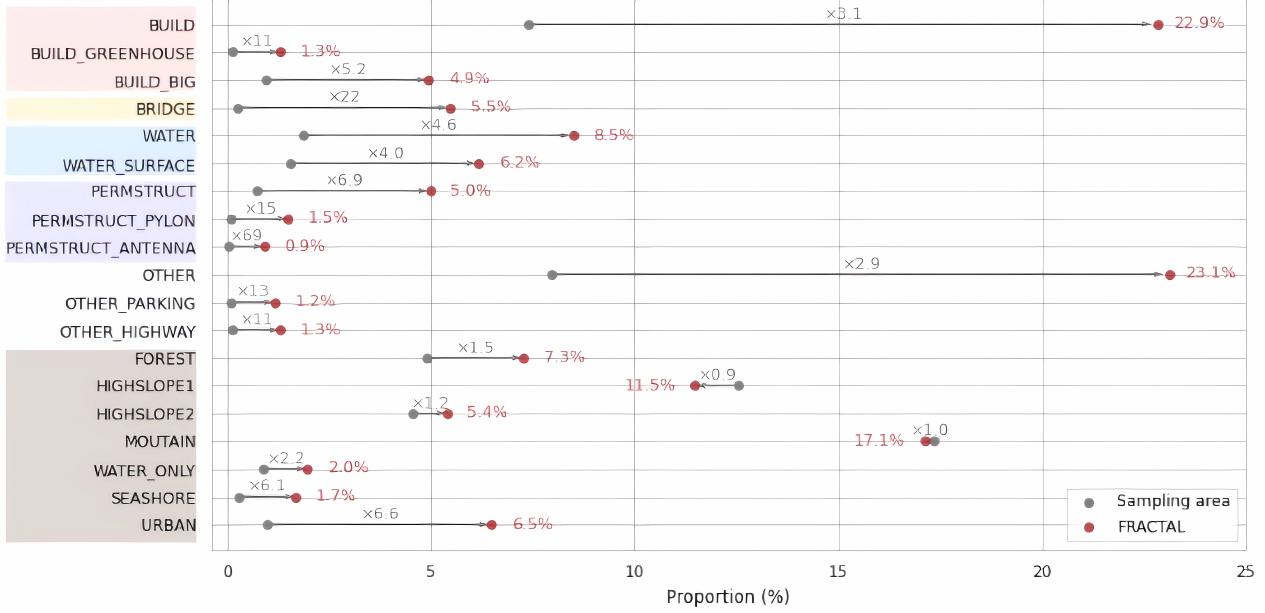


Figure 4. Proportional change of proportion of scene types in FRACTAL after sampling.

Lidar HD”), however, the specification of the Lidar HD classification tolerates some errors. This tolerance is class specific; for instance, a perfect 100% recall for individual buildings is not mandatory (as previously stated, see Section 3.1).

Beyond errors, the very definition of “unclassified” points leads to potential confusions with other classes. This class is semantically ill-defined and may include points with blurred boundaries with adjacent classes such as building, ground, and vegetation. These imperfections are expected to have a negligible impact for model evaluation, but users should be aware that they exist when inspecting the data.

Point clouds are colorized with non-simultaneous aerial imagery (as explained in Section 3.2), which can lead to color artifacts (typically for moving objects such as vehicles) and to misalignment between colors and shapes, which can blur object boundaries (e.g., roof points colored by a nearby tree). We observed only marginal impact on segmentation performances; overall, the colorization of the point clouds benefits segmentation models.

Using FRACTAL to train models for production should be done with caution. The dataset covers 5 spatial domains from 5 southern regions of metropolitan France. While large and diverse, it represents only a fraction of the French territory and is not representative of its full diversity regarding either landscapes or human-made structures. Furthermore, domain shifts are frequent in aerial images due to different acquisition conditions and downstream data processing. When using models trained on the dataset to predict a classification in unseen regions, one should make the adequate verification and not assume its capacity to generalize.

Similarly, considering new Lidar sensors should be done carefully. ALS data of comparable point densities (about 40 pts/m²) are expected to have consistent geometric characteristics, but users should still assess the model’s accuracy when predicting from alternative 3D data sources.

5. Experimental results

5.1. Task and metrics

FRACTAL is a benchmark dataset for semantic segmentation of ALS scenes into 7 semantic classes: other, ground, vegetation, building, water, bridge, and permanent structure. This reduced nomenclature is adapted from the Lidar HD nomenclature, which has 11 classes (see Appendix D.1). Low, medium, and high vegetation are grouped together since they differ only in height above ground. Artifact and synthetic points are simply filtered out. The “unclassified” ASPRS class is renamed “other” for clarity. See Appendix D.2 for a table of correspondence between the two nomenclatures.

We adopt the evaluation practices of similar Lidar datasets (e.g., DALES [24]) and use the mean Intersection over Union (mIoU) as our main metric. The “other” class is imperfect (see Section 4.5), but we decide not to exclude it from metric calculations since it contains well-defined objects that are relevant to scene segmentation (e.g., boats, containers). Considering the size of the dataset, imperfections of the “other” class are expected to have a negligible impact in model evaluation.

For consistency with other benchmarks, we also report Overall Accuracy (OA). All metrics should be computed

Table 4. Baseline test IoUs and OA

Class	IoU	OA
other	47.5	54.9
ground	91.9	97.7
vegetation	93.8	95.6
building	90.4	93.7
water	90.1	92.6
bridge	65.2	96.1
permanent structure	63.5	76.6
Macro Average	77.5	86.7

from the confusion matrix accumulated over all point clouds from the test set.

5.2. Baseline model

A baseline performance is established using Myria3D [7]. Myria3D is a 3D deep learning library developed at IGN, which leverages Pytorch-Lightning [4], and the the 3D deep learning library PyTorch-Geometric [5]. It was explicitly built for the semantic segmentation of Lidar HD data. Scalability informed its design, including the choice of its neural architecture: RandLa-Net [9].

Since the PointNet++ architecture [21] succeeded to the ground-breaking PointNet [20] to operate directly on un-ordered point clouds, there were many attempts to improve over point-based architectures, characterized by PointNet-like operations hierarchically organized in a U-shaped architecture. Conceptually simple, RandLa-Net makes some interesting additions to PointNet++. It uses a lightweight module for local spatial encoding and achieves performance gains thanks to random sampling. Importantly, its authors demonstrated its segmentation accuracy on large-scale outdoor Lidar benchmark datasets like SemanticKITTI [1] and Semantic 3D [8].

To produce the first experimental results on the benchmark, we kept the default hyperparameters of [3d-deep-learning-library]. Details about the processing of colorized Lidar HD tiles, about hyperparameters (optimizer, learning rate, scheduler, early stopping, etc.), and about infrastructure, are given in Appendix C.4.

5.3. Results

The baseline model achieves a test mIoU of 77.5% and a test OA of 96.1%. Tab. 4 reports the IoU and OA for each class. We observe highly accurate results for the three most common classes: ground, vegetation, and building have an IoU above 90%, with a maximum of 93.8% for vegetation. The "water" class has an IoU of 90.1%, despite its extreme initial rarity (0.6%) in the area of interest. For the "bridge" and "permanent structure" classes, also initially rare (0.01% each), the model achieves decent performance with an IoU above 60%. The baseline underperforms for the "other"

class compared to the six other ones, but still achieves an IoU of 47.5%, which is remarkable considering the fuzzy definition of this class.

Additional metrics (precision, recall, and F1 score) are reported in Appendix C.1. Confusions matrices are also given (Appendix C.2). Finally, the qualitative performance of the baseline model can be assessed from examples of model predictions (Appendix C.3).

6. Conclusion

We present FRACTAL, an ultra-large-scale benchmark dataset for the 3D semantic segmentation of ALS point clouds. The dataset is based on high quality open data from the French Lidar HD acquisition program. It is a distillation of a larger initial area of 17,440 km² in five French regions. We show that a simple targeted sampling with spatial stratification at all levels preserves the diversity of regional-scale volumes of Lidar data. FRACTAL's size is compatible with deep learning research practices, and it is the largest Lidar benchmark dataset to date, with 9,261 million points in 100,000 point clouds and a total span of 250 km². While other ALS benchmark datasets typically cover a single urban area, the large diversity of urban and rural landscapes in FRACTAL is representative of the challenges of 3D semantic segmentation for land monitoring. Thanks to class rebalancing, our methodology opens the possibility of robust assessment of semantic segmentation performance, even for the initially rare classes water, bridge, and permanent structure. The baseline evaluation of a 3D neural network (RandLa-Net) further demonstrates the quality of the dataset for model evaluation. We invite the research community to benchmark both state-of-the-art and novel methods against this dataset. We hope that FRACTAL will advance the field of deep learning for airborne Lidar and ultimately benefit public Lidar-based 3D mapping programs for land monitoring.

Code and data access

The dataset is made available on the HuggingFace platform as [IGNF/FRACTAL](#). The sampling, extraction, and colorization of point clouds were conducted with the Patch Catalog Sampling (PaCaSam) code repository available at: github.com/IGNF/pacasam. The baseline model was trained with the Myria3D code repository available at github.com/IGNF/myria3d. Its weights are made available on HuggingFace as [IGNF/FRACTAL-LidarHD_7cl_randlanet](#).

All assets are released under permissive open licences.

Acknowledgements

The authors thank Léa Vauchier for her code reviews of the data engineering code, and Marouane Zellou for main-

taining the in-house high-performance computing server, and Matthieu Porte and Anatol Garioud and Nicolas Audibert for reviewing this data paper. The authors also thank Anatol Garioud for paving the way for the open release of deep learning datasets for land monitoring at IGN with the FLAIR dataset [6].

Authors' contribution

C.G. established the sampling framework. C.G. and F.R defined the data descriptors and their target proportions. C.G. implemented the tools for patch sampling and data extraction. F.R. and M.D. defined the specifications for the Lidar Patch Catalog, and M.D. implemented it. C.G. implemented and evaluated the deep learning baseline. C.G. wrote the data paper and released the dataset.

References

- [1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 8
- [2] Adrian Boguszewski, Dominik Batorski, Natalia Ziembankowska, Tomasz Dziedzic, and Anna Zambrzycka. Land-cover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1102–1110, June 2021. 2
- [3] Comet ML. Comet [ML platform]. www.comet.com, 2024. 17
- [4] William Falcon and The PyTorch Lightning team. PyTorch Lightning [software]. www.github.com/Lightning-AI/lightning, 2019. 8
- [5] Matthias Fey and Jan E Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 8
- [6] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 9
- [7] Charles Gaydon. Myria3D: Deep learning for the semantic segmentation of aerial Lidar point clouds [software]. www.github.com/IGNF/myria3d, 2022. 8
- [8] Timo Hackel, N Savinov, L Ladicky, Jan D Wegner, K Schindler, and M Pollefeys. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-1-W1, pages 91–98, 2017. 8
- [9] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLa-Net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020. 8
- [10] Institut national de l'information géographique et forestière (IGN). Lidar HD [Database]. www.geoservices.ign.fr/documentation/donnees/alti/lidarhd, 2023. 1, 2, 3
- [11] Institut national de l'information géographique et forestière (IGN). LiDAR HD version 1.0 - descriptif de contenu des nuages de points LiDAR. www.geoservices.ign.fr/sites/default/files/2023-10/DC_LiDAR_HD_1-0_PTS.pdf, 10 2023. 3, 6, 18
- [12] Institut national de l'information géographique et forestière (IGN). ORTHO HR [Database]. www.geoservices.ign.fr/bdortho, 1 2023. 4
- [13] Institut national de l'information géographique et forestière (IGN). BD TOPO® [Database]. www.geoservices.ign.fr/bdtopo, 2024. 12
- [14] Georgia Kakoulaki, Ana Martinez, and Florio Petro. Non-commercial Light Detection and Ranging (LiDAR) data in Europe. Technical report, Joint Research Commission, 2021. 1
- [15] Teja Kattenborn, Felix Schiefer, Julian Frey, Hannes Feilhauer, Miguel D. Mahecha, and Carsten F. Dormann. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5, 8 2022. 4
- [16] Michael Kölle, Dominik Laupheimer, Stefan Schmohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1:100001, 10 2021. 2, 3
- [17] Markus Melin, Aurélie C Shapiro, and Paul Glover-Kapfer. LiDAR for ecology and conservation - WWF conservation technology series (3). Technical report, WWF-UK, 2017. 1
- [18] Nur Atirah Muhadi, Ahmad Fikri Abdullah, Siti Khairunniza Bejo, Muhammad Razif Mahadi, and Ana Mijic. The use of LiDAR-derived DEM in flood applications: a review. *Remote Sensing*, 12, 7 2020. 1
- [19] Joachim Niemeyer, Franz Rottensteiner, and Uwe Soergel. Contextual classification of lidar data and building object detection in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:152–165, 1 2014. 2, 3
- [20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 8
- [21] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 8
- [22] Nannan Qin, Weikai Tan, Lingfei Ma, Dedong Zhang, and Jonathan Li. OpenGF: An ultra-large-scale ground filtering dataset built upon open ALS point clouds around the world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1082–1091, 2021. 3
- [23] The American Society for Photogrammetry and Remote Sensing. LAS specification 1.4 - R15. Technical report, The American Society for Photogrammetry and Remote Sensing, 2019. 6
- [24] Nina Varney, Vijayan K Asari, and Quinn Graehling. DALES: A large-scale aerial LiDAR data set for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 186–187, 2020. 2, 3, 7
- [25] Zhen Ye, Yusheng Xu, Rong Huang, Xiaohua Tong, Xin Li, Xiangfeng Liu, Kuifeng Luan, Ludwig Hoegner, and Uwe Stilla. LASDU: A large-scale aerial LiDAR dataset for semantic labeling in dense urban areas. *ISPRS International Journal of Geo-Information*, 9, 7 2020. 2, 3
- [26] M Yebra, S Marselis, A van Dijk, G Cary, and Y Chen. Using Lidar for forest and fuel structure mapping: Options, bene-

- fits, requirements and costs. Technical report, Bushfire and Natural Hazards CRC, 2015. 1
- [27] Kin Yen. Automated LiDAR extraction software. Technical report, Caltran's Division of Research, Innovation and System Information, 2021. 2
- [28] P. Zachar, K. Bakuła, and W. Ostrowski. CENAGIS-ALS Benchmark - new proposal for dense ALS benchmark based on the review of datasets and benchmarks for 3D point cloud segmentation. volume 48, pages 227–234. International Society for Photogrammetry and Remote Sensing, 10 2023. 2, 3
- [29] S. M. Iman Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Rogerio Eduardo da Silva, Morteza Rahbar, and Aljosa Smolic. DublinCity: Annotated LiDAR point cloud and its applications. 9 2019. 2, 3

Appendices of the FRACTAL datapaper

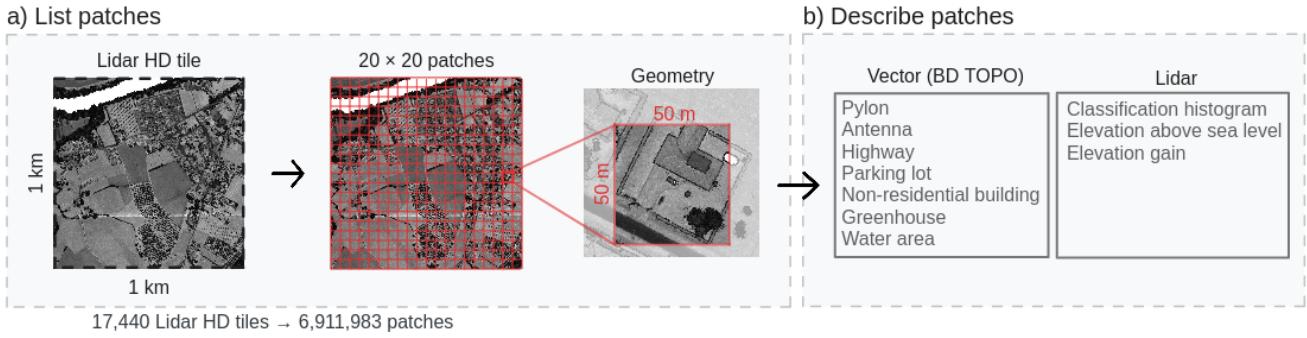
A. Data representation for patch sampling

A.1. Considerations for data cataloguing

We present the data representation used for sampling, which takes the form of a PostGIS database. Cataloging data boils down to two steps: listing and describing, which we illustrate in Appendix A.2. We start by regularly dividing the area of interest into 50×50 m patches. We then describe all patches with two sets of descriptors: Lidar descriptors and vector descriptors. Lidar descriptors are derived from the Lidar data patches themselves: number of points in each semantic class, elevation (above sea level), and elevation gain. Vector descriptors are derived by cross-referencing the patches with the BD TOPO® [13], a 3D vector description of the French territory and its infrastructures (buildings, roads, etc.). We include flags indicating the presence of the following objects: greenhouses, antennas, pylons, highways, parking lots, water surfaces, greenhouses, pylons, antennas, and industrial, commercial, or agricultural buildings.

Cataloguing with a PostGIS database gives us the full power of SQL syntax to select data and to further characterize each Lidar patch with bespoke SQL queries that combine simple descriptors into more complex ones (see Tab. 2).

A.2. Data cataloguing



Data cataloguing boils down to a) listing and b) describing all Lidar patches, prior to sampling.

B. Dataset content

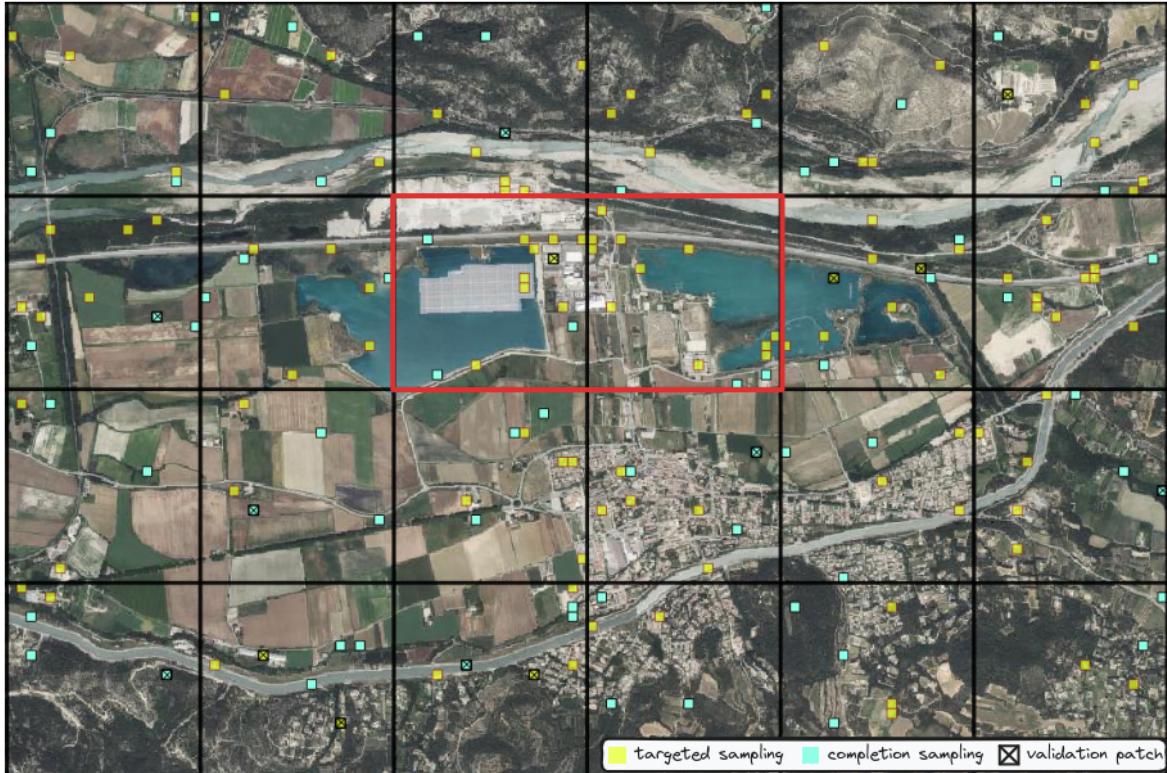
B.1. Number of points by class in the area of interest, in the dataset, and in the train, val and test sets

Name	Area of Interest		FRACTAL		Train		Val		Test	
	Points (M)	%	Points (M)	%	Points (M)	%	Points (M)	%	Points (M)	%
Other	2,137	0.32	53	0.57	41	0.56	5	0.53	6	0.66
Ground	258,751	39.1	3,625	39.1	2,874	39.0	359	39.1	391	40.5
Vegetation	394,425	59.6	5,248	56.7	4,203	57.0	523	56.9	523	54.1
Building	4,939	0.75	264	2.85	206	2.80	26	2.80	32	3.33
Water	1,504	0.23	54	0.59	38	0.52	5	0.49	12	1.20
Bridge	47	0.01	12	0.13	9	0.13	1	0.10	2	0.16
Permanent structure	33	0.01	3	0.04	3	0.04	0	0.04	0	0.03
Total	661998	-	9261	-	7376	-	919	-	966	-

B.2. Proportions of scene types in the train, val, and test sets

Descriptor	Train (%)	Val (%)	Test (%)
BUILD	22.5	22.8	25.7
BUILD_BIG	5	4.8	4.9
BUILD_GREENHOUSE	1.3	1.3	1.2
BRIDGE	5.5	5.4	5.4
WATER	8.2	8.1	11
WATER_SURFACE	6	5.9	7.7
PERMSTRUCT	5	4.9	4.6
PERMSTRUCT_ANTENNA	0.9	0.9	0.6
PERMSTRUCT_PYLON	1.5	1.4	1.4
OTHER	22.7	22.6	27.5
OTHER_HIGHWAY	1.3	1.2	1.7
OTHER_PARKING	1.1	1.1	1.5
FOREST	7.3	7.4	7.1
HIGHSLOPE1	11.5	11.8	11
HIGHSLOPE2	5.5	5.4	4.8
MOUTAIN	17.4	17.6	14.5
WATER_ONLY	1.7	1.7	4.5
SEASHORE	1.4	1.4	4.5
URBAN	6.4	6.5	7.4

B.3. Patches in FRACTAL on a subset area of 4×6 Lidar HD tiles



The 182 selected patches cover only 1.9% of the 24 km² (52-fold reduction). Patches selected via targeted sampling (yellow) are more present in complex urban areas. Patches from completion (cyan) sampling are homogeneously distributed. Validation patches (crossed in black) are homogeneously distributed across Lidar HD tiles and scene types. The two tiles framed in red are the ones displayed in Fig. 3

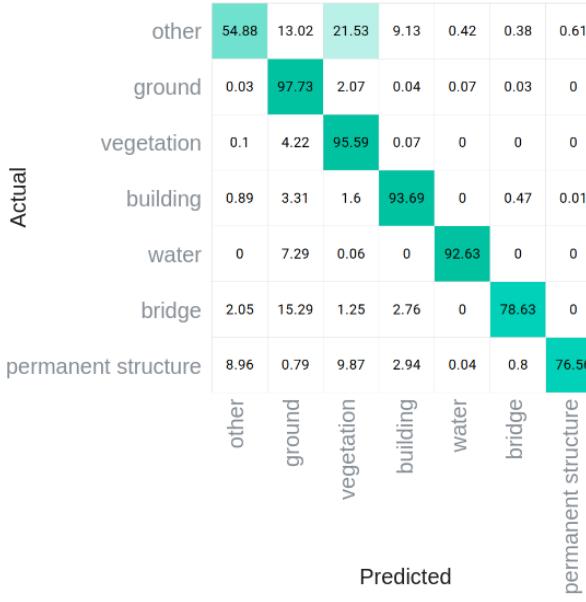
C. Baseline model evaluation

C.1. Baseline test IoU, precision, recall, and F1 score, for each semantic class and macro-averaged

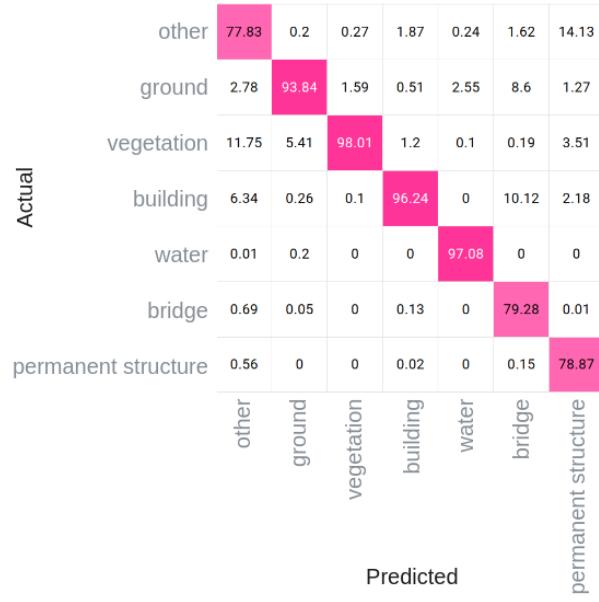
Class	IoU	Accuracy	Precision	Recall	F1 Score
other	47.5	54.9	77.8	54.9	64.4
ground	91.9	97.7	93.8	97.7	95.8
vegetation	93.8	95.6	98.0	95.6	96.8
building	90.4	93.7	96.2	93.7	95.0
water	90.1	92.6	97.1	92.6	94.8
bridge	65.2	96.1	79.3	78.6	79.0
permanent structure	63.5	76.6	78.9	76.6	77.7
Macro Average	77.5	86.7	88.7	84.2	86.2

C.2. Baseline test confusion matrices normalized by rows (a) and columns (b)

(a) Row-normalized (recall)



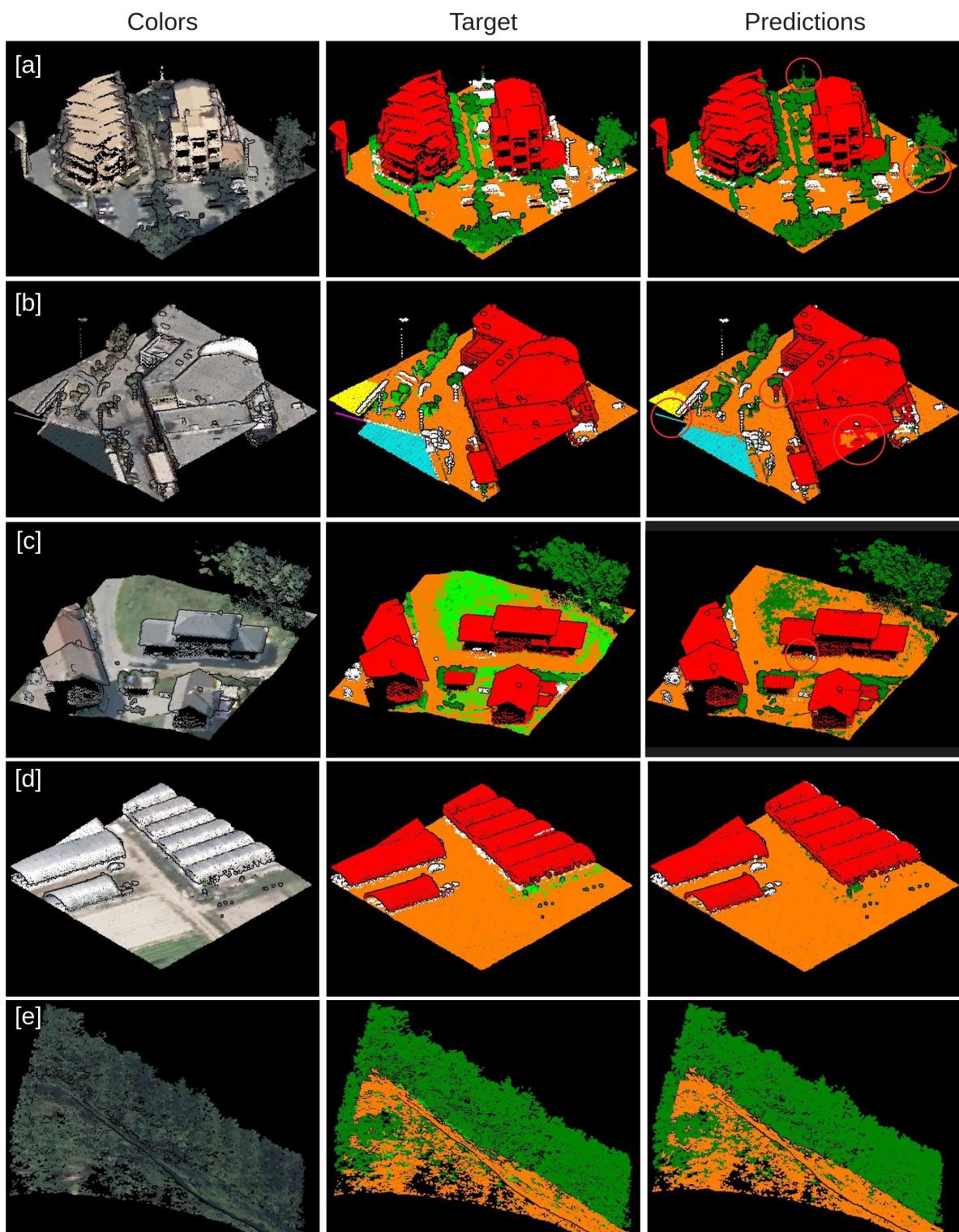
(a) Column-normalized (precision)



Diagonal values in row-normalized matrix correspond to recall. Diagonal values in column-normalized matrix correspond to precision.

Vegetation, ground, and building have low relative confusion, as expected from their high IoU. Permanent structure has its highest confusion with vegetation, due to the vertical nature and intricate geometries shared by both classes e.g., between pylons and trees. High interclass confusion is also found between classes that may have imprecise geometric boundaries. For example, bridge points are misclassified as ground, of which they are often an extension. Points of class other have high confusion with vegetation and ground points. This is mainly due to the Lidar HD classification specification, which favors specificity over recall for the ground and vegetation classes. As a result, the model may be penalized for accurately identifying ground and vegetation whose target class is other. Note that with only 0.6% of all points in FRACTAL, class other is a small minority, and these errors have a negligible effect on the measured performance for the ground and vegetation classes.

C.3. Input cloud, target classification, and prediction of baseline model for a random subset of patches



Test patches are selected at random and without cherry-picking, among test patches with at least 10k points (at least 4 pts/m²), to match the following scene types: a) OTHER_PARKING, b) WATER and BRIDGE, c) URBAN, d) BUILD_GREENHOUSE, and e) HIGHSLOPE1. Color scheme is: other (white), ground (orange), vegetation (green), building (red), water (cyan), bridge (yellow), permanent structure (purple). Predictions errors are circled in red.

C.4. Preprocessing and training hyperparameters

Lidar is unstructured; how we feed clouds to a model is important, and has a lot of degrees of freedom. We first subsample clouds via a grid with a voxel size of 0.25 m. A maximal budget of 40,000 points per cloud is allocated, enforced when needed via a random subsampling. Clouds are horizontally centered by subtracting their average position along the x and y axes, and vertically aligned by subtracting their minimal position along the z axis. Centered coordinates are then divided by 25 (meters) to be homogeneously brought between -1 and 1. We apply several generic data augmentations, namely: Random Scale (x0.8 to x1.25 factor), Random Jitter (-0.05 m to +0.05 m along each axis), Random Translate (-1 m to +1 m), Random Flip (along x and y axes).

In terms of features, the following dimensions are included: x, y, z, reflectance, echo number, number of echos. Number of echos and echo number are normalized by constants to be between 0 and 1. Reflectance is log-normalized, standardized, and values above three standard deviations are clamped. The point cloud is colorized, and we also consider the near infrared, red, green, blue dimensions. We set all colors to 0 for points with an echo number above 1, as a very basic occlusion model. Furthermore, we compute the Normalized Difference Vegetation Index (NDVI). We also calculate a new “color intensity” feature as the average of red, green, and blue channels. Colors, color intensity, and NDVI, are divided by a constant to be between 0 and 1. Average color is normalized similarly to the reflectance, i.e., by log-normalization, standardization, and truncation of values above three standard deviations.

Training is supervised with Cross-Entropy loss and the Adam optimizer. The learning rate is 0.004, with a reduction strategy (ReduceLROnPlateau) that halves the learning rate with a patience of 20 epochs and a cooldown of 5 epochs after each reduction. A batch size of 10 is used. We train the model for at least 100 epochs and retain the model that minimizes the validation loss.

Training is conducted with 6 NVIDIA Tesla V100 GPUs, each equipped with 32 GB of memory, using Pytorch-Lightning’s Distributed Data Parallel (DDP) strategy. The approximate learning time is 30 minutes per epoch. We log metrics using Comet, a machine learning experiment tracking tool [3].

D. Data sources specifications

D.1. Specifications of the Lidar HD classification

RGB	Value	Name	Content
255,255,255	1	Unclassified	All points that do not belong in any of the other classes. For example, it includes vehicles, animals or people, temporary objects, wood piles, etc.
255,128,0	2	Ground	Points located on the surface of natural and artificial ground. Bridge decks are not part of this class.
0,255,0	3	Low vegetation	Trees, shrubs and low vegetation (e.g. bushes, ferns, reeds, etc.).
0,193,0	4	Medium vegetation	Vegetation at ground level (less than 20 cm high, typically grass) is classified as vegetation only if there are enough ground points locally for ground modeling. The class also includes cultivated trees (orchards, vineyards, field crops, etc.). Vegetation is further divided into 3 classes based on height above ground: low ($\leq 0.5\text{m}$), medium (between 0.5 m and 1.5 m), and high ($\geq 1.5\text{ m}$) vegetation.
0,134,0	5	High vegetation	
255,0,0	6	Building	A building is defined as a permanent structure with an area greater than 10 m^2 . Besides residential buildings it includes monuments, castles, mills, water towers, lighthouses, industrial chimneys, ramparts and fortifications. The specification also includes roofs and façades, as well as chimneys, dormer windows, skylights and balconies. As the permanent nature of a structure cannot be established from lidar data only, this class may include lightweight structures without walls such as garden sheds, bungalows, market canopies or awnings.
0,225,225	9	Water	All points located on the surface of rivers, bodies of water, sea, or ocean.
255,255,0	17	Bridge deck	A bridge is an engineering structure passing over one or more elements of the road, rail or waterway network. Point on bridge decks are included, while structural elements such as piers and parapets are assigned to the "Unclassified" class. Very high structural elements ($\geq 5\text{ m}$ above deck level) such as piers and abutments are classified as "Permanent structures". Tunneled passages (including nozzles, which are openings in the ground generally to allow water to drain) are considered part of the ground and therefore excluded.
128,0,64	64	Permanent structures	All aboveground objects other than buildings, vegetation and bridges, that are identified as perennial and of such a nature that they characterize the landscape. This class includes (but is not limited to): wind turbines, cable cars, telecommunication antennas, electricity distribution networks (cables and pylons), bridge elements above the deck (cables, piers, etc.).
64,0,128	65	Artifact	All points that do not correspond to an actual object or terrain.
255,0,255	66	Synthetic	Artificial points created under bridges and on water surfaces to have coherent digital models.

Refer to the Lidar HD product description [1] for more specifications.

Lidar HD	→	FRACTAL
Unclassified	→	Other
Ground	→	Ground
Low vegetation		
Medium vegetation	→	Vegetation
High vegetation		
Buildings	→	Building
Water	→	Water
Bridge deck	→	Bridge
Permanent structures	→	Permanent structure
Artifact		-
Synthetic		

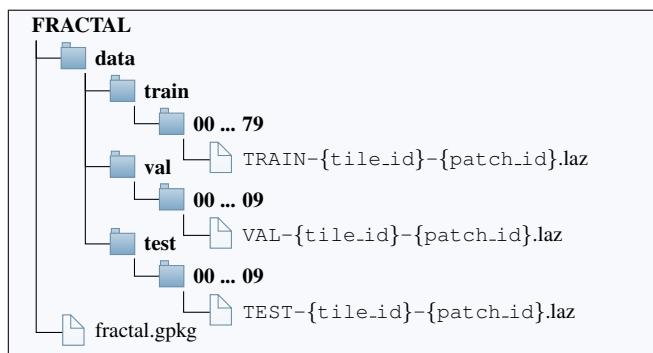
D.2. Relation between the Lidar HD classification nomenclature and its adaptation in FRACTAL

D.3. Sensors used to acquire Lidar HD data in each of the 8 50 × 50 km blocks represented in FRACTAL

Block	Sensor	Subcontractor(s)	
		Acquisition	Classification
GN	CityMapper 2H: Hyperion 2+	APEI + Avineon	Avineon
MQ	RIEGL VQ1560 II-S	Eurosense + SFS	Eurosense + SFS
MP	RIEGL VQ1560 II	Eurosense	Sintégra
PK	RIEGL VQ1560 II	Eurosense	Eurosense
PO	Leica CityMapper-2	Avineon + APEI	Avineon + APEI
PP	Leica CityMapper-2	Avineon + APEI	Eurosense + SFS
QO	RIEGL VQ780 II-S	Sintégra + Bluesky	Sintégra + Bluesky
UT	RIEGL VQ780 II-S	Sintégra + Bluesky	Avineon + APEI

Data acquisition and its classification were often conducted by distinct subcontractors, which we also report.

E. Structure of files and directories



The point clouds are organized by set (train, val, and test), in 100 subdirectories of 1000 point clouds each. The naming convention defines `tile_id` as the X and Y kilometer northwest coordinates of the Lidar HD tile in the Lambert 93 projection (EPSG:2154), and `patch_id` is a unique, arbitrary patch identifier. For instance, patch `TEST-0744_6246-006804306.laz` belongs to the test set, was extracted from Lidar HD tile whose top left coordinates are X = 0744 and Y = 6246, and its unique identifier is 006804306.

The geometries of all patches and their descriptions are provided in a metadata file: `fractal.gpkg`.