How the Choice of SLAM Algorithm Impacts Point Cloud Classification: An Experimental Evaluation

Antonio Matellon[1,2], Eleonora Maset[1], Alberto Beinat[1], Domenico Visintini[1]

[1] Polytechnic Department of Engineering and Architecture (DPIA), University of Udine, 33100 Udine, Italy
matellon.antonio@spes.uniud.it; (eleonora.maset, alberto.beinat, domenico.visintini)@uniud.it
[2] Department of Engineering and Architecture (DIA), University of Trieste, 34127 Trieste, Italy

Keywords: Classification, Point Clouds, Simultaneous Localization and Mapping, Mobile Robotics, Machine Learning

Abstract

This paper investigates the impact of different Simultaneous Localization and Mapping (SLAM) algorithms on semantic point cloud classification, by analyzing how reconstruction characteristics affect downstream classification tasks. Four SLAM approaches (*RTAB-Map*, *LIO-SAM*, *DLO*, and *LeGO-LOAM*) are used to process identical raw LiDAR and IMU data acquired by a sensorized mobile robotic platform, producing reconstructions with markedly different point density and geometric quality. A Random Forest classifier based on multi-scale geometric features is then applied to the SLAM-derived point clouds and to reference data acquired using Terrestrial Laser Scanning (TLS), assessing both in-domain classification performance and cross-dataset generalization. The results show that classification accuracy and class-wise reliability strongly depend on SLAM-induced point cloud characteristics, with denser reconstructions yielding higher performance. Cross-dataset experiments further reveal an asymmetric generalization behavior, whereby models trained on SLAM-derived point clouds transfer more robustly to denser datasets than models trained on TLS data. Feature importance analysis links this behavior to a shift toward coarser geometric descriptors as point density decreases. To support reproducibility and to encourage further investigations, including the evaluation of alternative classification strategies, the SLAM and TLS point clouds are released as an open-access dataset, together with manually annotated ground-truth labels.

1. Introduction

Simultaneous Localization and Mapping (SLAM) algorithms are increasingly used not only for robotic navigation, but also for autonomous three-dimensional mapping and reconstruction of real-world environments. The diffusion of SLAM-based portable laser scanners has enabled rapid and flexible data acquisition in complex indoor and outdoor scenarios, often providing higher efficiency and greater spatial completeness than traditional surveying techniques (Maset et al., 2021). These advantages, however, usually come at the expense of a reduction in accuracy, precision, and level of detail when compared to established approaches such as Terrestrial Laser Scanning (TLS) (Conti et al., 2024, Matellon et al., 2024). Therefore, integration strategies combining SLAM, TLS, and Unmanned Aerial Vehicle (UAV)-based photogrammetric data have also received attention to exploit the complementary strengths of surveying techniques (Chiabrando et al., 2019, Maset et al., 2022).

In many applications, data acquisition and point cloud generation are not the final outcome of a survey. In the architectural and built heritage domain, point clouds are increasingly exploited within scan-to-BIM workflows, where three-dimensional data form the basis for information-rich modeling of the built environment. In this context, semantic segmentation and classification of point clouds play a crucial role (Croce et al., 2021), as they support the automation or semi-automation of modeling pipelines. While these processing steps have been extensively investigated for TLS-derived point clouds, their application to SLAM-generated data remains comparatively underexplored.

The non-uniform density, higher noise levels, and trajectory-related artifacts typical of SLAM reconstructions introduce additional challenges with respect to static datasets. Moreover, as these properties vary across SLAM algorithms, they are expected to directly influence classification performance. For this reason, this study investigates how different SLAM approaches affect classification results by applying a consistent classification workflow to point clouds reconstructed from identical raw measurements using different SLAM algorithms. The objective is to quantitatively assess the impact of SLAM-induced point cloud characteristics on the reliability of semantic classification.

The remainder of the paper is structured as follows. Section 2 reviews the relevant state of the art. Section 3 describes the study area, the data acquisition setup, and the adopted methodology. Section 4 presents and discusses the experimental results, while Sect. 5 concludes the paper and outlines perspectives for future work.

2. Related Work

The semantic segmentation and classification of three-dimensional point clouds have long represented a central research topic within the fields of geomatics, computer vision, and digital built heritage. The primary aim is to partition raw point cloud data into homogeneous and semantically meaningful subsets, supporting the interpretation of the surveyed environment and subsequent modeling and analysis tasks.

Early approaches proposed in the literature are predominantly based on geometric and topological criteria, exploiting local point cloud properties such as surface orientation, curvature, point density, and neighborhood relationships. Region-growing, model-fitting, and clustering-based algorithms have demonstrated good performance, particularly in

structured environments (Grilli et al., 2017). Region-growing methods are widely used to extract planar surfaces from point clouds of historic buildings, while model-fitting techniques enable the segmentation of both simple and irregular geometric elements. Moreover, model-fitting and clustering-based algorithms support surface defect detection and deformation analysis by evaluating deviations between the point cloud and fitted geometries (Yang et al., 2023). However, these methods mainly address geometric segmentation, i.e., the partitioning of point clouds into clusters sharing similar geometric characteristics, and offer limited capabilities when semantic interpretation is required.

As a result, increasing attention has been devoted to approaches capable of incorporating semantic information into the segmentation process. The development of Machine Learning (ML) techniques has significantly influenced point cloud semantic segmentation and classification. Supervised ML models, such as Random Forest (RF) and Support Vector Machines (SVM), have been extensively investigated, particularly in the context of TLS and photogrammetric point clouds for architectural and cultural heritage applications (Grilli and Remondino, 2020). Several studies emphasize the importance of selecting discriminative features and appropriate computation scales to optimize classification performance (Grilli et al., 2019). In many cases, classification strategies are designed to be integrated within scan-to-BIM workflows, supporting the automated or semi-automated generation of parametric building elements (Moyano et al., 2021). Hierarchical classification approaches have also been proposed, in which an initial coarse classification is refined through subsequent stages to identify more detailed architectural elements (Ceccarelli et al., 2023). Such multi-level strategies have demonstrated their potential to enhance both the robustness and the interpretability of classification results, particularly in complex and heterogeneous heritage scenarios.

More recently, a growing body of literature has focused on Deep Learning (DL)-based methods, evaluating different neural network architectures and performing comparative analyses both among DL models and between traditional ML and DL approaches (Matrone et al., 2020, Pierdicca et al., 2020). These studies consistently indicate that no single methodology clearly outperforms the others across the heterogeneous scenarios typical of architectural and built heritage applications. Instead, classification performance is strongly influenced by the data characteristics and the required level of semantic detail. As a result, the selection of an appropriate classification strategy remains closely tied to the specific requirements of the intended application and the objectives of the downstream modeling process.

All the aforementioned studies primarily rely on point clouds acquired through TLS and photogrammetric techniques, including both terrestrial and UAV-based surveys. Despite the substantial progress achieved in semantic segmentation and classification, the application of these methods to data acquired using SLAM-based systems is still limited. Some studies have explored the adaptation of segmentation and classification workflows originally developed for TLS data to SLAM-derived point clouds, suggesting that semantic segmentation performance is not severely affected by differences in point cloud characteristics. In particular, variations in accuracy and level of detail appear to have a limited impact on classification outcomes, without significantly compromising the results (De Geyter et al., 2022). Other works have demonstrated that SLAM-based devices, such as the BLK2GO laser scanner, can produce point clouds suitable for semantic analysis, achieving promising classification results when using RF-based models (Franzini et al., 2023).

Nevertheless, a systematic comparison of point clouds generated by different SLAM algorithms with respect to classification performance is still largely missing. Most existing studies do not explicitly investigate how variations in the adopted SLAM technology influence downstream semantic processing stages, leaving a gap in the understanding of the relationship between data acquisition strategies and advanced point cloud interpretation. Within this context, the present work aims to contribute to a critical assessment of the potential and limitations of SLAM-derived point clouds for semantic classification tasks, also in comparison with TLS reference data, by explicitly analyzing how classification performance varies as a function of the adopted SLAM algorithm when applied to the same input dataset.

3. Materials and Methods

3.1 Data Acquisition Setup

To assess the influence of SLAM algorithms on point cloud classification, the dataset presented in (Tiozzo Fasiolo et al., 2023) was adopted in this study. Data were acquired using the mobile robotic platform Scout 2.0 by AgileX Robotics, a wheeled ground robot designed for outdoor environments (Fig. 1). The platform was equipped with a Velodyne VLP-16 Light Detection and Ranging (LiDAR) sensor and an Xsens MTi-630 9-axis Inertial Measurement Unit (IMU), both mounted on top of the robot. The LiDAR sensor provides full 360° horizontal coverage through 16 laser channels, with a measurement range of up to 100 m, while the IMU integrates gyroscope, accelerometer, and magnetometer measurements to deliver orientation, angular velocity, and linear acceleration data at high frequency.

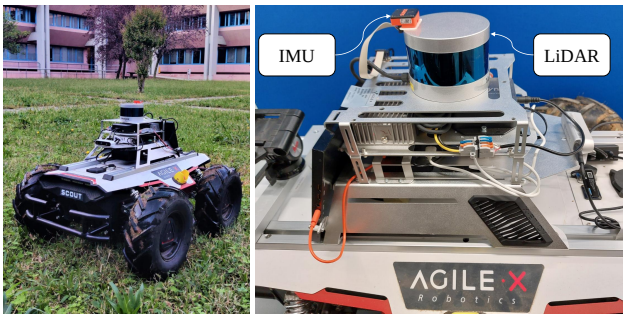
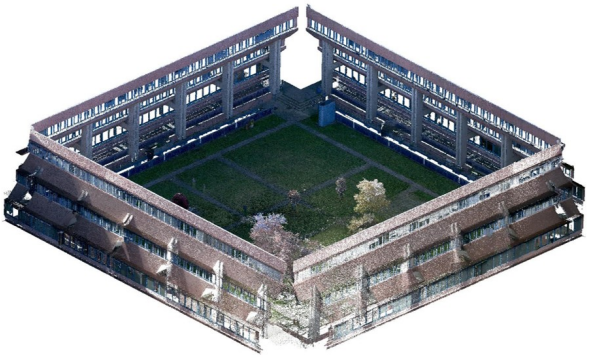


Figure 1. The robot employed to collect the data (left) and a close-up view of the onboard sensors (right).

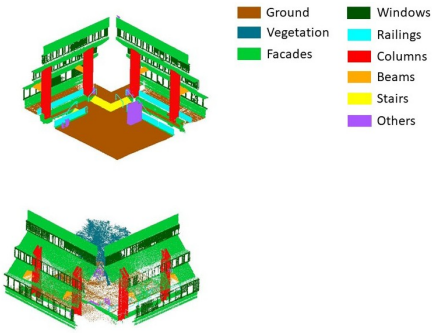
The experimental test was carried out in a 45 m × 45 m courtyard (measured along the centerline of the surrounding corridors), located within the main building of the Rizzi scientific campus of the University of Udine (Italy). This outdoor environment comprises covered walkways, structural elements, furniture, and vegetation (Fig. 2(a)).

3.2 SLAM Processing

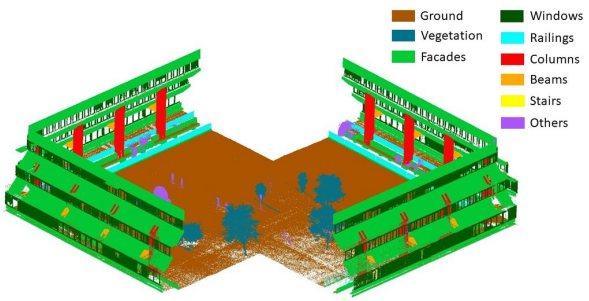
The raw LiDAR and IMU measurements were stored in ROS bag files and subsequently post-processed using four distinct SLAM algorithms: Real-Time Appearance-Based Mapping



(a) TLS point cloud.



(b) Manually labeled TLS point cloud, training set.



(c) Manually labeled TLS point cloud, test set

Figure 2. Reference TLS point cloud: (a) RGB view; (b) training set; (c) test set. Point clouds (b) and (c) are colored by point class.

(*RTAB-Map*), LiDAR Inertial Odometry via Smoothing and Mapping (*LIO-SAM*), Direct LiDAR Odometry (*DLO*), and Lightweight and Ground-Optimized LiDAR Odometry and Mapping (*LeGO-LOAM*).

More in detail, in the graph-based SLAM algorithm *RTAB-Map* (Labbé and Michaud, 2019), scan matching is performed on voxelized point clouds using an Iterative Closest Point (ICP)-based point-to-plane registration. A motion prediction, based on a constant-velocity model and optionally supported by IMU data, is used to initialize the alignment. The global map is incrementally updated after scan-to-map registration, while motion-induced scan distortion is neglected under the assumption of limited platform velocity. Differently from *RTAB-Map*, *LIO-SAM* (Shan et al., 2020) adopts a tightly coupled formulation that integrates LiDAR and IMU data within a factor graph framework. Edge and planar features are exploited for LiDAR scan matching, while IMU measurements provide both motion priors and distortion correction. The optimization incorporates LiDAR odometry, IMU integration, and loop closure constraints, and supports LiDAR–IMU extrinsic pre-calibration. A lighter processing strategy is followed by *DLO* (Chen et al., 2022), which performs scan matching directly on voxel-filtered point clouds without explicit feature extraction. Registration is carried out on a local submap using a Generalized ICP formulation, optionally supported by IMU gyroscope data for motion prediction. However, loop closure and explicit distortion compensation are not included. Finally, *LeGO-LOAM* (Shan and Englot, 2018) relies on a feature-based approach specifically designed for ground robots. Edge and planar features, including ground points, are extracted and used in a two-stage scan matching process to estimate the platform pose. IMU measurements are exploited to estimate sensor orientation and compensate for motion distortion, while a loop closure mechanism is adopted to mitigate drift.

As a result of these methodological differences, the four SLAM algorithms produced point clouds with markedly different characteristics in terms of density, noise, and geometric quality (Tiozzo Fasiolo et al., 2022, Tiozzo Fasiolo et al., 2023). *RTAB-Map* generated the densest reconstructions, although with a slightly higher noise level; *DLO* and *LeGO-LOAM* resulted in significantly sparser point clouds, whereas *LIO-SAM* provided an intermediate density with limited outliers and a balanced noise level, representing a favorable trade-off between point cloud completeness and geometric quality. It is worth noting that *RTAB-Map*, *LeGO-LOAM*, and *DLO* can be operated either in LiDAR-only or LiDAR–IMU configurations. In this work, the IMU-assisted configurations were adopted for all three methods.

3.3 Reference Data and Annotation

In addition to the SLAM-derived reconstructions, a point cloud acquired using a Leica BLK360 G1 TLS was included as a reference dataset. In this work, the reference role is intended in terms of higher point cloud completeness and density, as well as lower noise with respect to the SLAM-based reconstructions, rather than as a benchmark for geometric accuracy.

Both the TLS (Fig. 2) and the SLAM-derived point clouds were manually annotated according to nine semantic classes: *ground*, *vegetation*, *facades*, *windows*, *railings*, *columns*, *beams*, *stairs*, and *others*. Manual labeling was performed using TerraScan by Terrasolid, a software suite specifically designed for point cloud management and processing. The availability of dedicated manual classification tools, together with advanced three-dimensional visualization capabilities, enabled a detailed and systematic inspection of the data. In particular, the use of cross-sections and the isolation of specific portions of the point cloud facilitated accurate and consistent label assignment across the different classes.

To support transparency and reproducibility, the manually classified point clouds generated in this study have been made publicly available through the open-access Zenodo repository at the following link: https://doi.org/10.5281/zenodo.18269477.

3.4 Classification Framework

Point cloud classification was performed using the well-established Random Forest (RF) algorithm, as implemented in the 3DMASC framework developed by a research group at the University of Rennes (Letard et al., 2024). The acronym MASC

stands for *Multiple Attributes, Scales and Clouds*, reflecting the multi-scale and multi-level nature of the approach. The framework can be employed either through a graphical user interface, available as a plugin for the open-source software CloudCompare, or via a command-line interface. In the present study, the latter option was adopted, and the workflow was executed within a Python environment by leveraging standard machine learning libraries, including scikit-learn.

Random Forest was selected as classification algorithm due to its robustness and proven effectiveness in point cloud semantic analysis, particularly in architectural and built heritage applications. RF models can efficiently handle high-dimensional feature spaces and heterogeneous data, and they exhibit good generalization capabilities in the presence of noise and class imbalance. Moreover, as a machine learning approach rather than a deep learning one, RF does not require large training datasets to achieve stable performance, making it suitable for scenarios with limited labeled data. Finally, the interpretability of feature importance make RF a reliable choice for comparative analyses involving point clouds with different characteristics, such as those produced by different SLAM algorithms.

The complete set of features employed by the RF classifier consists of 19 descriptors, namely the Z coordinate, roughness, mean curvature, surface density, first-order moments, and a set of features based on the covariance matrix. The latter include the sum of eigenvalues, omnivariance, eigenentropy, anisotropy, planarity, linearity, PCA1, PCA2, surface variation, sphericity, verticality, and the first, second, and third eigenvalues. With the exception of the Z coordinate, all features were computed at five neighborhood scales (0.05 m, 0.1 m, 0.25 m, 0.5 m, and 1 m), resulting in a total of 91 descriptors used as input to the classifier and enabling geometric properties to be captured at different spatial extents.

In the first experiment, which represents the main experimental setup of this study, the labeled point clouds were subdivided into training and test sets following an approximate 25–75% split, as reported in Tab. 1 and illustrated in Fig. 2(b) and Fig. 2(c). The training subsets were used to train the classification models, while the test subsets were employed for the quantitative evaluation of classification performance. For the four SLAM-derived point clouds, the training and test partitions were defined by adopting the same spatial subdivision established for the TLS point cloud, thus ensuring spatial consistency across datasets. As a consequence of differences in point density distribution, the actual percentages of points assigned to the training and test sets vary among the considered point clouds.

Point cloud	Surface density pts/m^2 ($r = 0.5$ m)	Training set n. points - %	Test set n. points %
TLS *BLK360*	24,416	9,641,000 - 23%	31,975,151 - 77%
SLAM[1] *RTAB-Map*	6,132	7,631,659 - 25%	22,773,002 - 75%
SLAM[2] *LIO-SAM*	226	265,284 - 25%	763,084 - 75%
SLAM[3] *DLO*	20	38,758 - 27%	105,007 - 73%
SLAM[4] *LeGO-LOAM*	9	17,230 - 27%	46,178 - 73%

Table 1. Key characteristics of the datasets and training–test splits used in the first (main) experiment.

Initially, a separate RF model was trained for each training set, including the TLS dataset and the four SLAM-derived point clouds, and evaluated on the corresponding test set. Subsequently, each trained model was cross-applied to the remaining test sets, resulting in a total of 25 classification runs.

To further assess the robustness of the classification results with respect to variations in the amount of training data, a second, complementary experiment was conducted. This experiment aimed at evaluating the impact of a reduced training set on model performance, while keeping the test set exactly the same as in the first experiment, in order to ensure full comparability of the results.

Starting from the TLS point cloud, a reduced training subset corresponding to approximately 11.3% of the available points was selected. By adopting the same spatial subdivision, reduced training sets were then defined for the SLAM-derived point clouds, corresponding to 15.3% for *RTAB-Map*, 15.8% for *LIO-SAM*, 16.3% for *DLO*, and 16.4% for *LeGO-LOAM*.

4. Results and Discussion

4.1 In-Domain Classification Performance

This section reports the classification results obtained for each point cloud by training and testing the Random Forest model on the corresponding training and test subsets, as defined in the main experiment. A qualitative overview of the classification outcomes is provided in Fig. 3, which visually highlights the spatial distribution and consistency of the predicted semantic classes across the different datasets. The figure also allows an immediate comparison between the TLS classified point cloud and the four SLAM-based reconstructions.

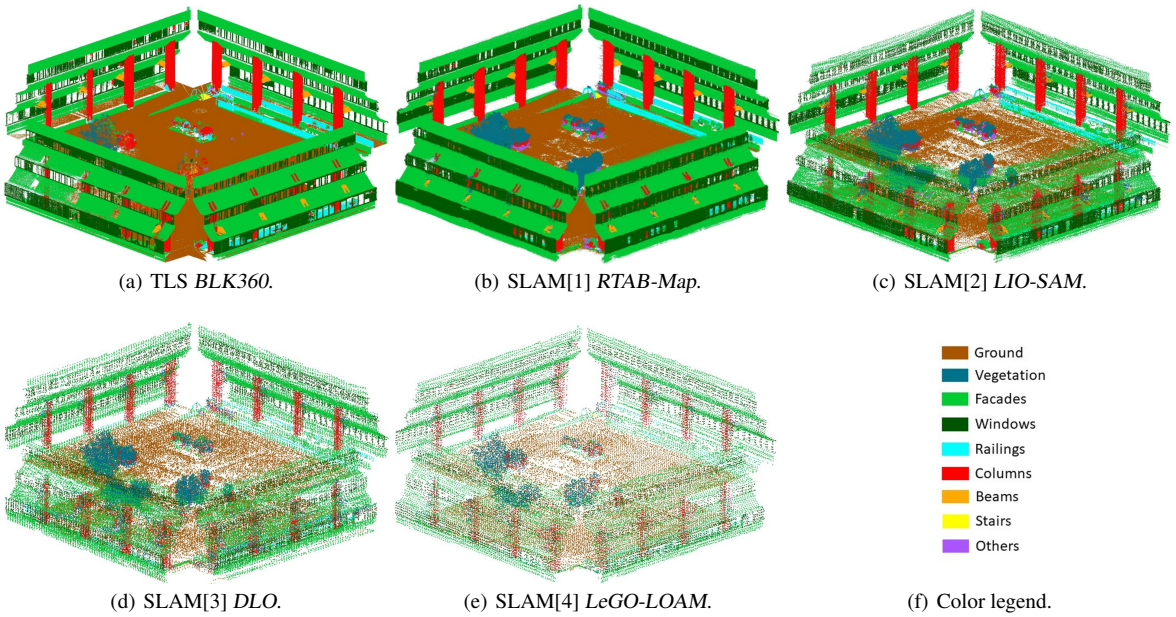
From a quantitative perspective, the evaluation was carried out using the metrics commonly adopted in point cloud classification, namely Overall Accuracy (OA), precision, recall, and F1-score. Precision, recall, and F1-score are reported in Tab. 2 both as class-wise mean values and as weighted averages, computed according to the number of points belonging to each class. The weighted formulation is particularly informative in this context, as simple class-wise averaging may be overly penalizing, assigning the same importance to poorly represented classes as to classes characterized by a substantially larger number of points. A detailed breakdown of the classification performance for each semantic class is provided in Tab. 3.

Point cloud	OA	Mean value			Weighted mean value		
		Prec	Rec	F1	Prec	Rec	F1
TLS *BLK360*	0.968	0.867	0.764	0.770	0.965	0.968	0.961
SLAM[1] *RTAB-Map*	0.946	0.889	0.796	0.830	0.941	0.946	0.943
SLAM[2] *LIO-SAM*	0.910	0.869	0.696	0.735	0.905	0.910	0.903
SLAM[3] *DLO*	0.834	0.741	0.530	0.541	0.832	0.834	0.819
SLAM[4] *LeGO-LOAM*	0.851	0.769	0.540	0.558	0.848	0.851	0.838

Table 2. Classification results for the main experiment. Overall Accuracy (OA), Precision, Recall, and F1-score were computed on the test set from the same point cloud as the training data, averaged across the nine classes.

As shown in Tab. 2, the TLS dataset achieves the highest values in terms of OA and across all weighted metrics, confirming the suitability of dense and low-noise point clouds for supervised semantic classification tasks. Among the SLAM-based reconstructions, *RTAB-Map* yields the best performance (OA = 0.946, weighted F1-score = 0.943), followed by *LIO-SAM* (OA = 0.910, weighted F1-score = 0.903). In contrast, *DLO* and *LeGO-LOAM* exhibit noticeably lower accuracies, with OA values of 0.834 and 0.851, respectively.

As discussed in Sect. 3, *RTAB-Map* and *LIO-SAM* generate denser and more complete reconstructions, whereas *DLO* and *LeGO-LOAM* produce substantially sparser point distributions. The reduced point density negatively affects the stability of



(a) TLS *BLK360*. (b) SLAM[1] *RTAB-Map*. (c) SLAM[2] *LIO-SAM*.

(d) SLAM[3] *DLO*. (e) SLAM[4] *LeGO-LOAM*. (f) Color legend.

Figure 3. Qualitative classification results on the test point clouds. For visualization purposes, point size was increased for the *DLO* and *LeGO-LOAM* datasets to improve readability.

Classes	TLS *BLK360*			SLAM[1] *RTAB-Map*			SLAM[2] *LIO-SAM*			SLAM[3] *DLO*			SLAM[4] *LeGO-LOAM*		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Ground	0.995	0.999	0.997	0.978	0.991	0.984	0.971	0.982	0.976	0.957	0.964	0.961	0.983	0.967	0.975
Vegetation	0.984	0.235	0.380	0.898	0.897	0.898	0.837	0.871	0.854	0.670	0.892	0.765	0.676	0.813	0.738
Facades	0.956	0.994	0.974	0.947	0.975	0.961	0.883	0.951	0.916	0.794	0.873	0.832	0.806	0.919	0.859
Windows	0.991	0.920	0.954	0.969	0.971	0.970	0.962	0.938	0.950	0.892	0.739	0.808	0.896	0.779	0.834
Railings	0.822	0.954	0.883	0.914	0.810	0.859	0.873	0.662	0.753	0.756	0.450	0.564	0.751	0.432	0.549
Columns	0.846	0.990	0.912	0.937	0.976	0.956	0.897	0.957	0.926	0.781	0.766	0.774	0.750	0.784	0.767
Beams	0.728	0.942	0.822	0.943	0.791	0.860	0.907	0.494	0.640	0.922	0.030	0.058	0.765	0.026	0.050
Stairs	0.966	0.741	0.839	0.932	0.481	0.634	0.969	0.148	0.256	0.632	0.042	0.078	0.964	0.132	0.232
Others	0.510	0.101	0.168	0.480	0.270	0.345	0.526	0.257	0.345	0.267	0.014	0.027	0.333	0.008	0.016

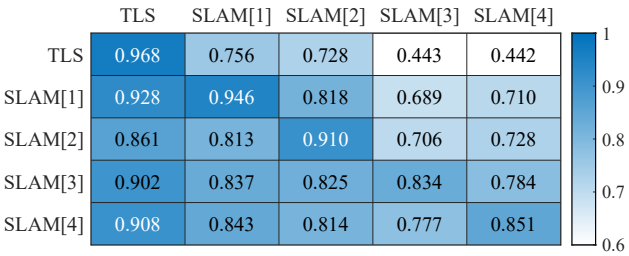
Table 3. Class-wise classification results for the nine semantic classes considered, reported in terms of Precision, Recall, and F1-score.

local geometric features and, consequently, the separability of semantic classes in the feature space used by the classifier. The gap between the best-performing SLAM solution (*RTAB-Map*) and the worst-performing one (*DLO*) exceeds 11 percentage points in terms of OA, suggesting that the choice of the SLAM algorithm alone can significantly influence not only the geometric characteristics of the reconstruction, but also the final quality of semantic labeling.

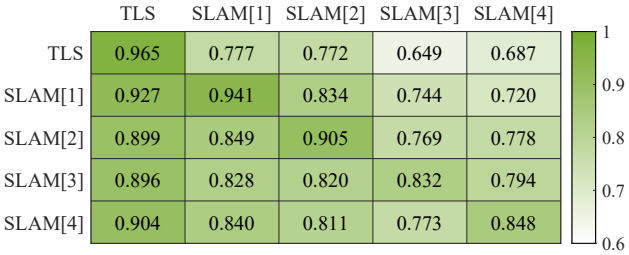
The class-specific metrics reported in Tab. 3, together with the visual inspection of the classified point clouds in Fig. 3, indicate that classes characterized by simple geometry and large spatial extent—most notably *ground* and *facades*—are reliably identified across all datasets. Very high performance is achieved for TLS and *RTAB-Map*, while satisfactory values are still observed even for the sparsest SLAM reconstructions. Slightly lower but relatively stable performance is also obtained for *windows* and *columns*. A distinctive behavior is observed for the *vegetation* class, which exhibits a low F1-score in the TLS dataset due to poor recall, whereas substantially higher values are achieved on SLAM-derived point clouds. This effect is mainly attributable to differences in point distribution and acquisition geometry, which influence the representation of vegetation elements in the point cloud. The most critical classes are *beams*, *stairs*, and *others*, for which decreasing point density leads to a severe degradation in classification performance. In particular, very low F1-scores are observed for the *DLO* and *LeGO-LOAM* datasets, reflecting the difficulty of identifying thin or weakly represented structural elements in sparse point clouds. Overall, these results confirm that the characteristics induced by the adopted SLAM algorithm substantially affect the reliability of classification at both global and class-specific levels.

4.2 Cross-Dataset Generalization

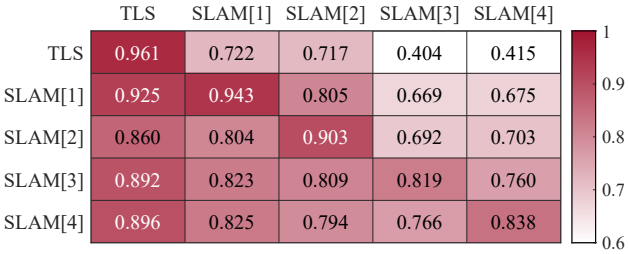
Figure 4 summarizes the classification results obtained by applying each trained model to all available test sets, thus enabling a direct assessment of cross-dataset generalization and model transferability. Each row corresponds to a classifier trained on a specific point cloud, while the columns report the performance achieved when the same model is applied to different target test sets. As previously reported, the model trained on the TLS dataset achieves the highest performance when evaluated on its own domain, consistently outperforming all SLAM-based models in the in-domain setting. However, when transferred to SLAM-derived point clouds, its performance progressively degrades. This decrease is relatively limited for the denser SLAM reconstructions, namely *RTAB-Map* (SLAM[1] in Fig. 4) and *LIO-SAM* (SLAM[2]), whereas it becomes more pronounced for the sparsest datasets, *DLO* (SLAM[3]) and *LeGO-LOAM* (SLAM[4]). This behavior highlights a strong dependency of the TLS-trained model on the geometric characteristics of the training data, particularly on point density and feature stability.



(a) Overall Accuracy.



(b) Precision (weighted).



(c) F1-score (weighted).

Figure 4. Classification metrics obtained in the first experiment for cross-dataset evaluation. Each row represents a trained model, while the columns report the classification performance achieved when applying the model to the different test sets.

An opposite trend is observed for models trained on SLAM-derived point clouds. These classifiers exhibit good performance on their original datasets and preserve, or even improve, their accuracy when applied to denser test sets, including the TLS point cloud. This effect is especially evident for models trained on the sparsest reconstructions, which benefit significantly from the increased point density of the target datasets. In these cases, the availability of richer geometric information in the test data compensates for the limited representativeness of the training samples. Overall, these results reveal a marked asymmetry in model transferability. While models trained on dense TLS data struggle to generalize to sparse SLAM-derived point clouds, classifiers trained on SLAM data—particularly on sparse reconstructions—tend to generalize more effectively across datasets with varying density levels. This suggests that training on geometrically simpler and noisier data encourages the learning of more robust and scale-tolerant decision boundaries, ultimately enhancing cross-dataset generalization.

4.3 Feature Relevance and Scale Dependency

The analysis of feature relevance provides additional insight into how differences in point cloud structure induced by the adopted SLAM algorithms affect the internal behavior of the Random Forest classifiers. In particular, it allows assessing which geometric descriptors and which spatial scales contribute most to the discrimination of semantic classes under varying density and noise conditions.

Figure 5 reports the cumulative importance of features grouped by spatial scale, while Fig. 6 details the contribution of individual features computed at specific neighborhood sizes for each training dataset. The results reveal a clear relationship between point cloud density and the spatial scale at which features become most discriminative.

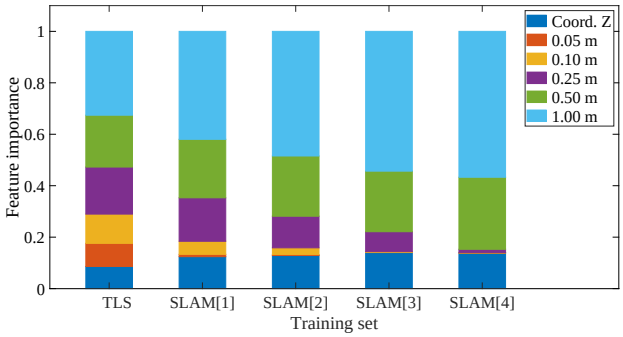


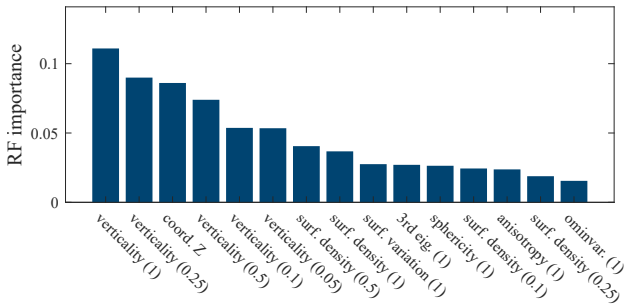
Figure 5. Cumulative feature importance by scale for each RF model.

For the TLS dataset and the denser SLAM reconstructions, i.e., *RTAB-Map* (SLAM[1]) and *LIO-SAM* (SLAM[2]), most of the informative content is concentrated at medium to small neighborhood sizes (0.10–0.50 m). This indicates that local geometric variations can be reliably captured and exploited by the classifier when point density is sufficiently high. Conversely, for the sparse reconstructions produced by *DLO* (SLAM[3]) and *LeGO-LOAM* (SLAM[4]), feature importance progressively shifts toward larger neighborhood sizes (0.50–1.00 m), reflecting the need to aggregate geometric information over wider spatial extents to obtain stable descriptors. Across all training datasets, the Z coordinate and verticality consistently emerge as the most influential variables, particularly at larger neighborhoods. These are followed by covariance-based descriptors such as sphericity, anisotropy, surface variation, and eigenentropy, whose relevance increases as point density decreases. This trend confirms that, in sparse point clouds, robust geometric characterization relies primarily on coarse-scale features that integrate information from larger neighborhoods.

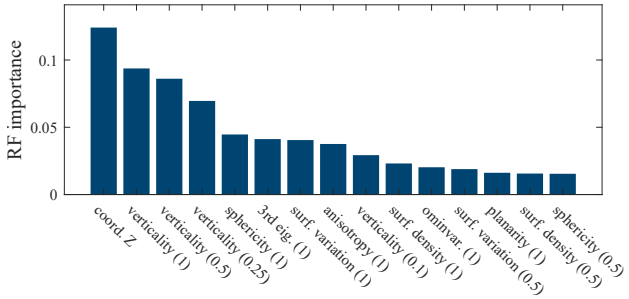
Overall, these findings explain the observed asymmetry in model generalization: models trained on sparse SLAM data rely on coarse and robust features that transfer well to denser datasets, whereas models trained on dense TLS data exploit fine-scale features that are not consistently available in sparse reconstructions, leading to reduced transferability.

4.4 Robustness to Reduced Training Data

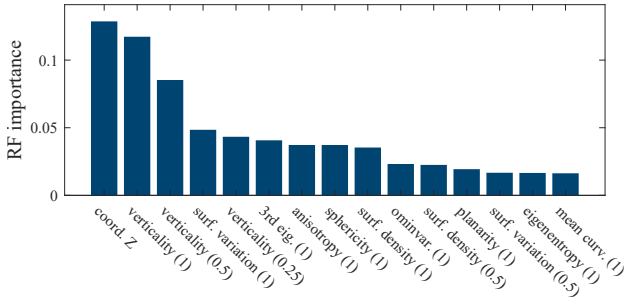
The second experiment was designed to evaluate the robustness of the classification framework under reduced availability of labeled training data. The quantitative results reported in Tab. 4 show that halving the size of the training sets does not lead to a substantial degradation of classification performance. For the TLS dataset, the overall accuracy decreases from 0.968 to 0.960, while for *RTAB-Map* it drops from 0.946 to 0.927. Similar trends are observed for *LIO-SAM*, *DLO*, and *LeGO-LOAM*, with only moderate reductions in both OA and weighted F1-score. These results indicate that the proposed classification workflow remains stable even when trained on a limited number of labeled samples.



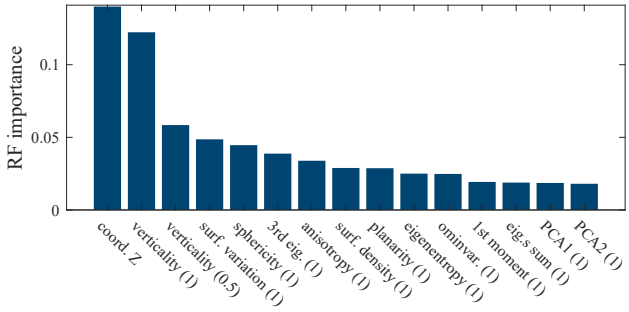
(a) TLS model.



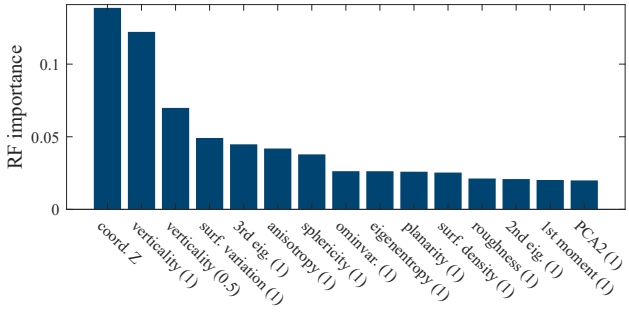
(b) SLAM[1] *RTAB-Map* model.



(c) SLAM[2] *LIO-SAM* model.



(d) SLAM[3] *DLO* model.



(e) SLAM[4] *LeGO-LOAM* model.

Figure 6. RF feature importance for each training dataset.

Point cloud	OA	Mean value			Weighted mean value		
		Prec	Rec	F1	Prec	Rec	F1
TLS *BLK360*	0.960	0.862	0.729	0.755	0.956	0.960	0.952
SLAM[1] *RTAB-Map*	0.927	0.843	0.741	0.774	0.920	0.927	0.919
SLAM[2] *LIO-SAM*	0.892	0.831	0.648	0.689	0.882	0.892	0.878
SLAM[3] *DLO*	0.827	0.682	0.518	0.531	0.820	0.827	0.812
SLAM[4] *LeGO-LOAM*	0.838	0.747	0.527	0.547	0.835	0.838	0.826

Table 4. Classification results of the second experiment (reduced training set). Performance metrics were computed on the test set from the same point cloud as the training data, averaged across the nine classes.

Further insights are provided by the cross-dataset evaluation summarized in Fig. 7. The general transferability patterns observed in the first experiment are largely preserved. Notably, in several cross-domain cases—when models are applied to point clouds different from those used for training—the classification performance is comparable to, or even slightly higher than that obtained in the first experiment with larger training sets. This behavior is particularly evident for models trained on SLAM-derived point clouds, which benefit from being applied to denser target datasets, where the increased availability of geometric information enhances class separability.

Overall, these findings suggest that classification performance does not critically depend on the sheer size of the training set, but rather on the representativeness of the training data. This result further supports the suitability of RF classifiers for semantic labeling tasks in scenarios characterized by limited annotated data, as is often the case for SLAM-derived point clouds.

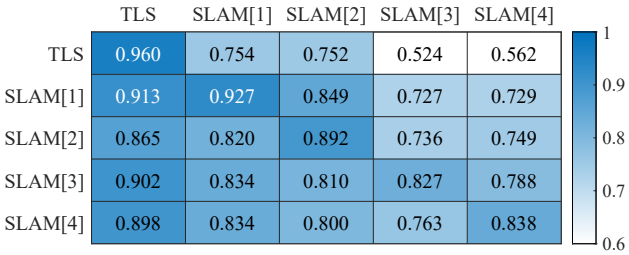


Figure 7. Overall accuracy obtained in the second experiment for cross-dataset evaluation, using the same layout as in Fig. 4.

5. Conclusion

This study investigated how different SLAM algorithms, used for three-dimensional reconstruction, influenced the classification of the resulting point clouds. The outcomes demonstrated that point cloud characteristics induced by the SLAM approach, particularly point density, have a strong influence on classification accuracy and class reliability. An important result concerns model generalization: classifiers trained on dense TLS data showed limited robustness when applied to sparse SLAM point clouds, whereas models trained on SLAM-derived data—especially sparse ones—exhibited better transferability across datasets with varying density. Feature relevance analysis confirmed that these effects are linked to changes in the spatial scales and geometric descriptors exploited by the classifier. The experiments further showed that stable classification performance can be achieved even with reduced training data, supporting the use of Random Forest models in scenarios suited for SLAM surveying.

Future work will extend the analysis to alternative classification strategies, including deep learning–based point-wise seg-

mentation models, to assess whether similar trends hold for approaches with higher representational capacity. In this context, the public release of the ground-truth annotated datasets is intended to facilitate reproducibility and to encourage further experimentation and comparative studies by the research community.

References

Ceccarelli, L., Bevilacqua, M. G., Caroti, G., Castiglia, R. B., Croce, V., 2023. Semantic segmentation through Artificial Intelligence from raw point clouds to H-BIM representation. *Disegnarecon*, 16(30), 17–1.

Chen, K., Lopez, B., Agha-mohammadi, A.-a., Mehta, A., 2022. Direct LiDAR Odometry: Fast localization with dense point clouds. *IEEE Robotics and Automation Letters*.

Chiabrando, F., Sammartano, G., Spanò, A., Spreafico, A., 2019. Hybrid 3D models: When geomatics innovations meet extensive built heritage complexes. *ISPRS International Journal of Geo-Information*, 8(3), 124.

Conti, A., Pagliaricci, G., Bonora, V., Tucci, G., 2024. A comparison between terrestrial laser scanning and hand-held mobile mapping for the documentation of built heritage. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 141–147.

Croce, V., Caroti, G., De Luca, L., Jacquot, K., Piemonte, A., Véron, P., 2021. From the semantic point cloud to heritage-building information modeling: A semiautomatic approach exploiting machine learning. *Remote Sensing*, 13(3), 461.

De Geyter, S., Vermandere, J., De Winter, H., Bassier, M., Vergauwen, M., 2022. Point cloud validation: On the impact of laser scanning technologies on the semantic segmentation for BIM modeling and evaluation. *Remote sensing*, 14(3), 582.

Franzini, M., Casella, V., Niglio, O., 2023. Leica BLK2GO point cloud classification with machine learning algorithms: The case study of Sant'Eusebio's crypt in Pavia (Italy). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 593–600.

Grilli, E., Farella, E. M., Torresani, A., Remondino, F., 2019. Geometric features analysis for the classification of cultural heritage point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 541–548.

Grilli, E., Menna, F., Remondino, F., 2017. A review of point clouds segmentation and classification algorithms. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 339–344.

Grilli, E., Remondino, F., 2020. Machine learning generalisation across different 3D architectural heritage. *ISPRS International Journal of Geo-Information*, 9(6), 379.

Labbé, M., Michaud, F., 2019. RTAB-Map as an open-source LiDAR and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics*, 36(2), 416–446.

Letard, M., Lague, D., Le Guennec, A., Lefèvre, S., Feldmann, B., Leroy, P., Girardeau-Montaut, D., Corpetti, T., 2024. 3DMASC: Accessible, explainable 3D point clouds classification. Application to Bi-spectral Topo-bathymetric lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 207, 175–197.

Maset, E., Cucchiaro, S., Cazorzi, F., Crosilla, F., Fusiello, A., Beinat, A., 2021. Investigating the performance of a handheld mobile mapping system in different outdoor scenarios. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 103–109.

Maset, E., Valente, R., Iamoni, M., Haider, M., Fusiello, A., 2022. Integration of photogrammetry and portable mobile mapping technology for 3D modeling of cultural heritage sites: The case study of the Bziza temple. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43(B2-2022), 831–837.

Matellon, A., Maset, E., Beinat, A., Visintini, D., 2024. Surface reconstruction from SLAM-based point clouds: Results from the datasets of the 2023 SIFET benchmark. *Remote Sensing*, 16(18), 3439.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R., Remondino, F., 2020. Comparing machine and deep learning methods for large 3D heritage semantic segmentation. *ISPRS International Journal of Geo-Information*, 9(9), 535.

Moyano, J., León, J., Nieto-Julián, J. E., Bruno, S., 2021. Semantic interpretation of architectural and archaeological geometries: Point cloud segmentation for HBIM parameterisation. *Automation in Construction*, 130, 103856.

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S., Frontoni, E., Lingua, A. M., 2020. Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sensing*, 12(6), 1005.

Shan, T., Englot, B., 2018. LeGO-LOAM: Lightweight and ground-optimized LiDAR odometry and mapping on variable terrain. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 4758–4765.

Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D., 2020. LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 5135–5142.

Tiozzo Fasiolo, D., Scalera, L., Maset, E., 2023. Comparing LiDAR and IMU-based SLAM approaches for 3D robotic mapping. *Robotica*, 41(9), 2588–2604.

Tiozzo Fasiolo, D., Scalera, L., Maset, E., Gasparetto, A., 2022. Experimental evaluation and comparison of LiDAR SLAM algorithms for mobile robotics. V. Niola, A. Gasparetto, G. Quaglia, G. Carbone (eds), *Advances in Italian Mechanism Science*, Springer International Publishing, Cham, 795–803.

Yang, S., Hou, M., Li, S., 2023. Three-dimensional point cloud semantic segmentation for cultural heritage: A comprehensive review. *Remote Sensing*, 15(3), 548.