Turin3D: Evaluating adaptation strategies under label scarcity in urban LiDAR segmentation with semi-supervised techniques

# Turin3D: Evaluating Adaptation Strategies under Label Scarcity in Urban LiDAR Segmentation with Semi-Supervised Techniques

Luca Barco[1,2*] Giacomo Blanco[2*] Gaetano Chiriaco[2*] Alessia Intini[1]
Luigi La Riccia[1] Vittorio Scolamiero[3] Piero Boccardo[1] Paolo Garza[1] Fabrizio Dominici[2]

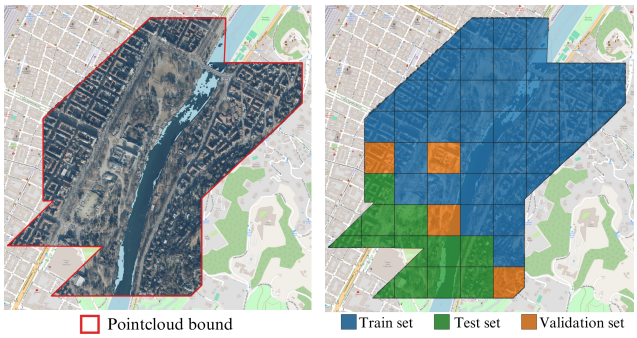
[1]Politecnico di Torino [2]LINKS Foundation [3]Sapienza Università di Roma

## Abstract

*3D semantic segmentation plays a critical role in urban modelling, enabling detailed understanding and mapping of city environments. In this paper, we introduce Turin3D: a new aerial LiDAR dataset for point cloud semantic segmentation covering an area of around* $1.43\ km^2$ *in the city centre of Turin with almost* $70M$ *points. We describe the data collection process and compare Turin3D with others previously proposed in the literature. We did not fully annotate the dataset due to the complexity and time-consuming nature of the process; however, a manual annotation process was performed on the validation and test sets, to enable a reliable evaluation of the proposed techniques. We first benchmark the performances of several point cloud semantic segmentation models, trained on the existing datasets, when tested on Turin3D, and then improve their performances by applying a semi-supervised learning technique leveraging the unlabelled training set. The dataset will be publicly available to support research in outdoor point cloud segmentation, with particular relevance for self-supervised and semi-supervised learning approaches given the absence of ground truth annotations for the training set.*

## 1. Introduction

Accurate 3D semantic segmentation is a fundamental task in urban mapping, enabling applications such as infrastructure monitoring, city planning, and environmental analysis. Aerial LiDAR (Light Detection and Ranging) technology has become an essential tool for acquiring large-scale, high-resolution 3D data in urban environments, offering detailed geometric representations of buildings, roads, vegetation, and other key elements of the urban landscape. However, despite the increasing availability of aerial LiDAR data, the number of publicly available labelled datasets designed specifically for semantic segmentation remains limited. In this work, we introduce Turin3D: a new aerial LiDAR dataset for 3D semantic segmentation, covering an urban area of approximately $1.43\ km^2$ in the city centre of Turin, Italy. The dataset was collected using an airborne LiDAR system capable of high-density point cloud acquisition, ensuring a detailed and precise representation of complex urban structures. Unlike datasets derived from terrestrial LiDAR, which are constrained by occlusions and perspective limitations, aerial LiDAR provides a complete top-down view of the urban scene, making it particularly well-suited for applications requiring large-scale mapping and monitoring.




(a) Acquired RGB point cloud and its perimeter (b) Dataset split in train, validation and test set

Figure 1. Turin3D point cloud whole extent and subdivision in blocks

The dataset is divided into three subsets: training, validation, and test. While the validation and test sets have been manually annotated to provide semantic labels and quantitative performance metrics, the training set remains unlabelled, given the prohibitive cost and time-intensive nature of the process.

To assess the usability of the dataset and establish benchmark results, we conduct experiments using popular deep learning models for 3D point cloud segmenta-

*Equal contribution

tion, such as Point Transformer [17] and RandLA-Net [7]. We first evaluate these models under both fully supervised conditions (leveraging existing annotated datasets) assessing their generalization capabilities on Turin3D. Then, we benchmark the best-performing architecture under semi-supervised conditions, where it is forced to learn from a training set where ground truth data is not available but instead artificially generated soft labels are used.

The main contributions of this work can be summarised as follows. *(i)* Introduction of a new publicly available aerial LiDAR dataset for 3D semantic segmentation in urban environments, with high-resolution point clouds collected over a dense city centre [1]. *(ii)* Benchmarking of popular segmentation models, evaluating their performance under different supervision settings, including scenarios without annotated training data.

By providing a new dataset and a thorough evaluation of deep learning models in different supervision regimes, this work aims to support future research in urban-scale 3D point cloud segmentation and promote the development of data-efficient learning approaches capable of leveraging partially annotated datasets.

The rest of the paper is organized as follows: Section 2 provides a review of related works in point cloud semantic segmentation, introducing existing datasets and popular methodologies. Section 3 provides an overview of how the dataset was collected, the taxonomy proposed and annotation process carried out. In Section 4, the applied methodology is explained in detail, covering the experimental setup of the different supervision settings. Section 5 presents the experimental results and performance evaluation of the proposed approach. Finally, Section 6 discusses the significance of the findings and potential avenues for future research.

## 2. Related Works

Accurate representation of the urban environment is critical for a variety of applications, including urban planning, environmental modelling, infrastructure monitoring, energy consumption estimation, and evaluation of the green energy potential. Many of these analyses cannot be effectively conducted using only two-dimensional data, such as images or cartographic information. Therefore, incorporating the vertical dimension through three-dimensional (3D) data enhances the structural and semantic characterization of urban areas. The combination of the availability of 3D data and the application of complex algorithms, i.e. Artificial Intelligence deep learning models, enables the development of advanced applications and analyses that can support decision-making in urban development and sustainability.

[1] https://huggingface.co/datasets/links-ads/Turin3D

### 2.1. Datasets for Urban Mapping

The development of accurate urban digital twins relies on high-quality 3D datasets for training and validation. These datasets are commonly represented as point clouds, consisting of sets of XYZ-points, typically acquired through LiDAR scanning or photogrammetric reconstruction. Alternative representations include meshes, which define surfaces through connected polygons, and volumetric models that partition space into regular grid cells (voxels). The majority of publicly available datasets in this domain provides point clouds.

The scanning modality is a key factor that affects the information encoded by the data [10, 11]. Ground-based (terrestrial) LiDAR provides highly detailed scans of building facades and street-level infrastructure but offers limited coverage of roof structures. Mobile LiDAR systems, mounted on vehicles, efficiently capture both road infrastructure and building facades along vehicle-accessible paths. Unmanned Aerial Vehicle (UAV) based systems enable flexible data collection from various heights and angles, particularly useful for capturing building roofs and areas inaccessible to ground vehicles. Aerial methods, conducted from aircraft, cover the largest areas most efficiently but provide lower detail of vertical surfaces. Photogrammetric reconstruction offers a cost-effective alternative to LiDAR by generating 3D point clouds from overlapping photographs, though potentially with lower geometric accuracy in some situations.

The datasets present in literature evidence considerable diversity in their acquisition methodologies, optical perspectives and colors, and class taxonomies.

**UAV-Based Datasets** SensatUrban [8] is an urban-scale photogrammetric point cloud dataset containing almost three billion points annotated with a taxonomy of 13 classes. It covers approximately 7.6 $km^2$ of urban landscape in three UK cities. The data were collected through aerial surveys, with automated flight paths pre-planned in a grid pattern. The collected data were then processed using commercial software, which applies Structure from Motion (SfM) and dense image matching techniques for point cloud reconstruction.
Similarly, Hessigheim 3D [9] is a dataset used for semantic segmentation of 3D point clouds and textured meshes, data were acquired from a LiDAR system and cameras integrated on the same Unmanned Aerial Vehicle (UAV) platform. The data were collected in the village of Hessigheim, Germany, and cover an area of about 0.19 $km^2$ with over $125M$ points. A distinctive feature of this dataset is its high spatial resolution; in fact, the point cloud features a density of about 800 $points/m^2$. The entire point cloud has been manually labeled following a taxonomy of 11 semantic classes.

**Aerial Datasets** Several datasets employ aerial acquisition methods. SUM [3] covers an area of $4\,km^2$ in Helsinki using airplane-mounted cameras, providing both meshes and point clouds with 6 semantic classes. The mesh was generated from oblique aerial images with a GSD (Ground Sampling Distance) of $7.5\,cm$, acquired in 2017 using a multi-camera system mounted on an aircraft, while reconstruction was performed with techniques of aerial triangulation, dense image matching and surface reconstruction.

The FRACTAL (FRench ALS Clouds from TArgeted Landscapes)[4] dataset is a large-scale LiDAR dataset designed for the semantic 3D segmentation of heterogeneous landscapes. It consists of $100,000$ point clouds acquired by Airborne LiDAR Scanning (ALS) and covers a total area of $250\,km^2$ in five regions of France. This dataset was constructed using open-source data from *the Institut national de l'information géographique et forestière (IGN)*. This dataset includes 9261 million points with an average density of 37 $points/m^2$, and a semantic annotation is provided for 11 classes.

Swiss3D [1] is a large-scale dataset designed for semantic segmentation of 3D point clouds acquired by drone photogrammetry. The dataset covers a total area of $2.7\,km^2$ spread over three Swiss cities and follows a five-class taxonomy. The data was collected by a multi-rotor drone following dual-grid flight paths, resulting in denser and more complete point clouds than those obtained with LiDAR sensors.

STPLS3D [2] is a large-scale dataset designed for semantic segmentation derived from aerial photogrammetry, covering more than $7\,km^2$ of landscapes and including 18 semantic categories. The novelty of this dataset compared to others is that it combines real-world data acquired from UAVs with three synthetic versions generated through a procedural pipeline. This approach addresses common challenges in real-world data collection and annotation, such as class imbalance and heterogeneous point quality. Synthetic data were obtained through a generation pipeline that mimics the real photogrammetric acquisition process by simulating UAV flights over virtual environments. These environments were built using open geospatial data and procedural modelling tools, enabling the creation of realistic 3D point clouds that remain compatible with real-world data while eliminating the need for manual annotations.

**Mobile Datasets** Toronto 3D [14] is a large-scale urban outdoor point cloud dataset for 3D semantic segmentation of urban environments; it is acquired through a Mobile Laser Scanning system (MLS) in Toronto, Canada and it covers $1\,km$ of urban streets comprising approximately 78.3 million points classified into 8 categories.

While these datasets could be used for urban mapping applications, aerial LiDAR datasets collected in dense urban environments remain relatively scarce. This limited availability poses challenges for developing and benchmarking algorithms specifically tailored for large-scale urban analysis.

## 2.2. 3D Semantic Segmentation

Deep learning methods for semantic segmentation of 3D point clouds in urban environments aim to extract hierarchical and spatially meaningful features, assigning each point to a specific semantic category. These approaches can be categorized into four main paradigms: *projection-based*, *discretization-based*, *point-based*, and *hybrid methods* [6].

Projection-based and discretization-based methods transform the point cloud into a structured representation, such as a 2D image or a voxel grid, where conventional deep learning techniques can be applied. The segmentation results are then reprojected onto the original point cloud. While these approaches leverage well-established CNN architectures, they introduce discretization artifacts and may lose fine geometric details. In contrast, point-based methods work directly on raw, unordered point clouds, preserving geometric precision but facing challenges due to the irregular distribution of points, which makes the application of standard convolutional operations less straightforward. Traditional convolutional networks struggle with sparse 3D data due to high computational costs and loss of sparsity, known as the Submanifold Dilation Problem. Submanifold Sparse Convolutional Networks (SSCN) address this by introducing Sparse Convolutions (SC) and Submanifold Sparse Convolutions (SSC) [5]. SC optimizes computation by assuming zero values for non-active sites, while SSC preserves the input sparsity structure, ensuring efficient feature extraction without unnecessary expansion. This approach improves point cloud processing efficiency, maintaining spatial continuity and minimizing resource waste. One of the first methods specifically designed for direct point cloud processing is PointNet [12], which applies shared Multi Layer Perceptrons (MLPs) to each point independently and aggregates global information through a symmetric pooling function. This architecture ensures permutation invariance and computational efficiency but has limitations in capturing fine-grained local geometric structures. To overcome this, PointNet++ [13] introduces a hierarchical feature extraction mechanism, recursively applying PointNet to spatially partitioned subsets of points. This allows the model to capture multi-scale local features while maintaining global context.

For large-scale point clouds, where computational efficiency is a key concern, RandLA-Net [7] has been proposed. It employs a random sampling strategy to reduce point density while integrating a local feature aggregation module, attentive pooling, and dilated residual blocks to

| Dataset | Year | # points | Classes | RGB | Intensity | Area | Sensor |
|---|---|---|---|---|---|---|---|
| SensatUrban [8] | 2020 | 2847M | 13 | ✓ | ✗ | 7.64 $Km^2$ | UAV Photogrammetry |
| Swiss3D [1] | 2020 | 226M | 5 | ✓ | ✗ | 2.7 $Km^2$ | UAV Photogrammetry |
| Toronto 3D [14] | 2020 | 78M | 8 | ✓ | ✗ | 1 $Km^2$ | MLS |
| Hessigheim [9] | 2021 | 74M | 11 | ✓ | ✓ | 0.19 $Km^2$ | ULS |
| SUM [3] | 2021 | 19M | 6 | ✓ | ✗ | 4 $Km^2$ | Aerial photogrammetry |
| STPLS3D (Real) [2] | 2022 | 150M | 6 | ✓ | ✗ | 1.27 $Km^2$ | Aerial Photogrammetry |
| FRACTAL [4] | 2024 | 9261M | 7 | ✓ | ✓ | 250 $Km^2$ | ALS |
| Turin3D (Ours) | 2025 | 69M | 6 | ✓ | ✓ | 1.43 $Km^2$ | ALS |

Table 1. Comparison of Turin3D dataset with the representative datasets for 3D semantic segmentation in urban scenarios MLS: Mobile Laser Scanning system, ULS: Unmanned Laser Scanning system, ALS: Airborne Laser Scanning system.

compensate for information loss due to downsampling. This approach enables efficient processing while preserving relevant spatial details, making it suitable for large-scale outdoor scenes.

Other methods integrate principles from convolutional neural networks or transformer-based architectures to enhance feature learning. KPConv (Kernel Point Convolutions) [15] replaces traditional MLP-based processing with learnable kernel points, allowing continuous convolutional operations that better capture local spatial relationships. However, this comes with a higher computational cost. An alternative approach is Point Transformer [17] and its variants, which utilize self-attention mechanisms to model long-range dependencies. By dynamically weighting interactions between points, these architectures improve the capture of both local and global contextual information, offering a flexible framework for point cloud segmentation.

Furthermore, frameworks have been developed and shared by authors to easily train and evaluate models using different datasets. In the context of this work, Open3D [18] has been used since it provides implementations of all the above-mentioned models.

## 3. Turin3D Dataset

The following section details the steps taken to create the *Turin3D* dataset. It first covers data acquisition using the *Leica CityMapper-2* sensor, then explains the 3D reconstruction process combining LiDAR and aerial imagery. Additionally, this section details the chosen semantic taxonomy, annotation workflow, data partitioning strategy, and key statistics, providing an overview of the dataset's composition.

### 3.1. LiDAR Acquisition

Turin3D was acquired using the *Leica CityMapper-2*, an airborne hybrid sensor that combines optical imagery and LiDAR scanning. The aerial survey was conducted on 28-29 January 2022, over a large area in the metropolitan city of Turin, Italy. The LiDAR component of the system operated with a conical scanning pattern, enabling the capture of vertical surfaces from multiple directions. The LiDAR acquisition was performed at an altitude of approximately 1 $km$, with a scanning angle of 20°, resulting in a point density ranging from 30 to 40 $points/m^2$. This density ensures a detailed geometric reconstruction of both ground and above-ground structures, making the dataset well-suited for urban mapping applications. Simultaneously, optical imagery was acquired using a combination of nadir and oblique cameras. A total of $20,291$ images were collected, with each acquisition point capturing one nadir and four oblique images. The photogrammetric dataset features a Ground Sampling Distance (GSD) of 5 $cm$, with an $80\%$ longitudinal and $60\%$ lateral overlap, ensuring high-resolution coverage of the area. The system is equipped with two different cameras: *Camera NIR Lens 71*, used for nadir and multi-spectral acquisition, and *Camera RGB Lens 112/145*, used for oblique imagery.

### 3.2. 3D Point Cloud Processing

The raw LiDAR data and aerial images were processed using *Agisoft Metashape 2.1.0* and *nFrames SURE 5.2*. These tools enabled the derivation of dense point clouds, 3D meshes, orthophotos, and Digital Terrain and Surface Models. The fusion of LiDAR and photogrammetric data aimed to add RGB features, compensate for occlusions and improve vertical surface reconstruction, particularly in high-density urban environments. A sample of the final, colourized point cloud is illustrated in Figure 2. This integration enanches both the geometric accuracy of LiDAR and the radiometric consistency of the photogrammetric data, making it a valuable resource for urban mapping applications and benchmark studies in 3D semantic segmentation.

### 3.3. Dataset Description

The dataset was collected in the San Salvario district of Turin on 29 January 2022. The covered area shown in Figure 1a spans approximately 1.43 $km^2$ and is made up of $69,591,759$ points. The entire area was divided into 57 blocks, each roughly $25,000$ $m^2$ in size. The number of

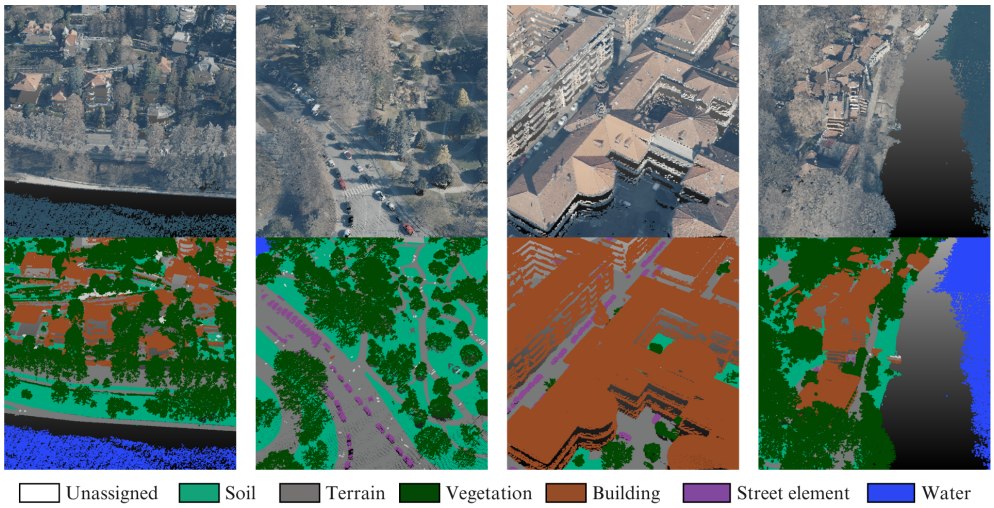



Figure 2. Close-in views of Turin3D. Top row displays the scenes in RGB coloring, bottom row shows the same areas with points colored according to their assigned class labels.

points per block varies significantly depending on the location of the point cloud. This variation is due to the diverse environments within the chosen area. The west side of the area shows a more urban and residential landscape. The central part is more vegetated, featuring a park, historic buildings, and a river. This area features an overall lower point density due to the scarcity of tall buildings and the limitations of LiDAR in accurately capturing water bodies. The east side is the most diverse, featuring a hilly terrain with a mix of vegetation and large houses. the heterogeneity of the landscape enhances the dataset's value, as it captures a wide range of urban environments despite being limited to a single city.

The point cloud data is stored in the standard LAS 1.4 format, which provides a structured framework for encoding each point with its attributes. Each point contains XYZ coordinates, intensity values, return number, number of returns, scan direction, scan angle, GPS time, and RGB color values.

### 3.4. Semantic Labels Taxonomy

The definition of semantic labels followed these principles: *(i)* each class must be distinguishable from the others, with high heterogeneity between classes and high homogeneity inside a class, *(ii)* each label class should add value for following downstream tasks and analysis, particularly for urban area planning and green applications. We decided to adopt a taxonomy composed of six distinct semantic labels. Compared to other datasets, we opted for a lower number of classes to avoid labels that are too similar and difficult for human annotators to distinguish reliably. Additionally, we included the *'Unassigned'* category for points that result from noise in the acquisition and reconstruction process, as well as masses of points that are too small to classify. All points belonging to this class were not taken into account in the experiments described in Section 4. The following is a list of the proposed taxonomy: *Unassigned*, all unidentified points; *Soil*, points that make up all kinds of natural surfaces, like meadows, soil; *Terrain*, points that make up artificial grounds, such as streets, sidewalks, cemented trails; *Vegetation*, all points belonging to trees, shrubs, bushes, and any other kind of low and high vegetation; *Building*, all points from walls, fences, barriers, residential and historic buildings; *Street elements*, cars, trucks, poles, benches; *Water*, points that make up all kinds of water elements, like river, water canals and pools.

The proposed taxonomy constitutes an initial attempt to systematically differentiate and categorize the primary components typically found in urban environments.

### 3.5. Annotation Process

The dataset was split into training, validation, and test sets, aiming for a point distribution as close as possible to a $70\%/10\%/20\%$ split. Only the validation and test sets were manually annotated, while the training set is used with soft labels only. This partition, shown in Figure 1b, was designed to ensure a proportional representation of the different urban settings and similar distribution of the six semantic classes, illustrated in Figure 3. The annotation process was conducted by a team of four annotators. The 17 test and validation blocks, resulting in a total amount of almost $19M$ points, were evenly distributed among them, and each annotator worked independently in the initial phase. Each point was assigned to one of the taxonomy classes defined in Section 3.4.

The annotation workflow began with the application of

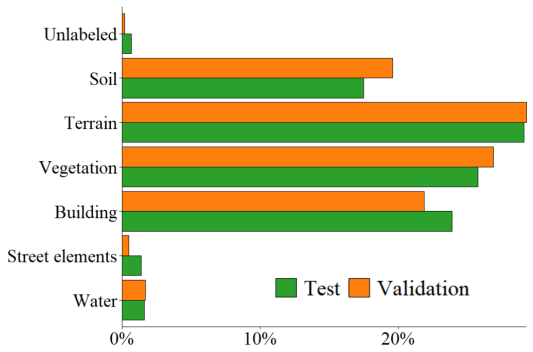



Figure 3. Distribution of classes across test and validation sets. The percentage indicates the proportion of each of the six classes within their respective sets, not relative to the entire dataset.

the CSF filter algorithm [16] to distinguish ground points from non-ground points. Ground points were then divided into natural and artificial ground, utilizing aerial imagery, RGB, and intensity features to facilitate the distinction. Non-ground points were manually classified into their respective categories, by leveraging point-wise features and available aerial images. Once all points were assigned, a final review was performed to verify coherence within local neighbourhoods and consistency with associated characteristics. To ensure consistency and reduce individual biases, a second round of review was conducted, during which the annotators collectively examined the tagged point cloud, addressing discrepancies and standardizing classification decisions across all blocks with round-table discussions.

A comparison of Turin3D with other representative datasets for 3D semantic segmentation is provided in Table 1. With a balanced class taxonomy, manually annotated points and a diverse urban landscape acquired with high-precision aerial LiDAR sensors, Turin3D offers a valuable benchmark for evaluating point cloud segmentation models under real-world conditions

# 4. Methodology

## 4.1. Problem Formulation

This research addressed the challenge of semantic segmentation of urban 3D point clouds across different urban environments, focusing on two approaches: transfer learning and semi-supervised learning with pseudo-labelling.

Let $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$ represent the source domain composed of $N_S$ point clouds from literature datasets, where $x_i^S \in \mathbb{R}^{P_i \times F}$ denoted a point cloud with $P_i$ points of $F$ features, and $y_i^S \in \mathcal{C}^{P_i}$ denoted point-wise labels from class set $\mathcal{C} = \{$*Unassigned, Soil, Terrain, Vegetation, Building, Street Element, Water*$\}$, according to the taxonomy proposed in Section 3.4. The literature datasets included SensatUrban [8], DELFT SUM [3], Toronto3D [14], FRACTAL [4], STPLS3D (Real) [2], Swiss3D [1], and Hessigheim [9], with each dataset's original classes mapped to exactly one class of $\mathcal{C}$.

Let $\mathcal{D}_T = \{(x_j^T)\}_{j=1}^{N_T^{train}} \cup \{(x_k^T, y_k^T)\}_{k=1}^{N_T^{val}} \cup \{(x_l^T, y_l^T)\}_{l=1}^{N_T^{test}}$ represent the Turin3D dataset with unlabeled training and labeled validation and test sets, consisting of $N_T^{train}$, $N_T^{val}$ and $N_T^{test}$ point clouds, respectively.

Three semantic segmentation architectures $\mathcal{A} = \{$RandLA-Net[7], PointTransformer[17], SparseConv[5]$\}$ were evaluated using two experimental approaches as described in the following sections.

## 4.2. Data Augmentation

Throughout all experiments, consistent data augmentation was applied to improve model generalization. Geometric augmentations included linear normalization to scale coordinates, point recentering along all axes, and rotation up to 30° to handle varied terrain elevations. For color, ChromaticAutoContrast, ChromaticJitter, and HueSaturationTranslation were applied to address lighting and appearance variations across datasets. Additionally, RandomHorizontalFlip was applied to x and y axes only, preserving the natural orientation of ground surfaces in urban environments.

## 4.3. Transfer Learning

The first approach addressed the fundamental challenge of generalizing to previously unseen urban environments. Each architecture $a \in \mathcal{A}$ was trained on the literature datasets to obtain $\theta_a^* = \arg\min_\theta \mathcal{L}(f_\theta^a, \mathcal{D}_S)$

These models were then evaluated on the Turin3D test set $\mathcal{D}_T^{\text{test}}$. For these experiments, the feature set $F = \{x, y, z, R, G, B\}$ was used, excluding intensity since it was not available across all literature datasets.

The transfer learning experiments evaluated whether models trained on existing literature datasets could effectively generalize to the unseen urban environment of Turin without any domain-specific adaptation.

## 4.4. Semi-Supervised Learning with Iterative Pseudo-Labeling

The second approach leveraged the large amount of unlabeled data in Turin3D through an iterative pseudo-labeling strategy. Based on the transfer learning results, the best-performing architecture on the Turin3D validation set was selected, i.e., $a^* = \arg\max_{a \in \mathcal{A}} \text{mIoU}(f_{\theta_a^*}, \mathcal{D}_T^{\text{val}})$.

The selected model was used to generate predictions on the unlabeled Turin3D training set $\mathcal{D}_T^{train}$. Each point was assigned the class label with the highest confidence score, but only if that score exceeded a class-specific confidence threshold $\tau_c$. This filtering resulted in a set of high-confidence pseudo-labels $\hat{y}_j^T = f_{\theta_{a^*}^*}(x_j^T)$ for a subset of points in the training set.

The confidence threshold for each class was calculated as a weighted average of confidence scores from predictions on $\mathcal{D}_T^{\text{train}}$:

$$\tau_c = \sum_{u \in \mathcal{U}_c} u \cdot \frac{n_{u,c}}{\sum_{v \in \mathcal{U}_c} n_{v,c}}$$

where $\tau_c$ was the confidence threshold for class $c$, $\mathcal{U}_c$ was the set of unique confidence values observed for points predicted as class $c$, $u$ was a specific unique confidence value, $n_{u,c}$ was the count of points predicted as class $c$ with confidence value $u$, and $\sum_{v \in \mathcal{U}_c} n_{v,c}$ was the total number of points predicted as class $c$.

A new instance of the selected architecture was trained from scratch on $\mathcal{D}_T^{\text{train}}$ using the pseudo-labelled points: $\theta^{**} = \arg\min_\theta \mathcal{L}(f_\theta^{a^*}, \{(x_j^T, \hat{y}_j^T)\}_{j \in \text{confident}})$. For these trainings, an expanded feature set $F = \{x, y, z, R, G, B, Intensity\}$ was used, since intensity values were available in the Turin3D dataset.

Two iterative refinement approaches were implemented to progressively improve pseudo-label quality. The first approach, Fixed Confidence Thresholds, maintained the same class-specific confidence thresholds across all iterations. This strategy allowed assessment of whether iterative training alone could enhance performance without threshold adjustment. The second approach, Adaptive Confidence Thresholds, recalculated confidence thresholds after each iteration based on the model's evolving performance and confidence distributions. This acknowledged that as the model adapted to the target domain, the optimal confidence thresholds might shift. Thresholds were systematically adjusted based on performance metrics from the previous iteration, creating a bootstrapping mechanism that progressively refined both the model and its pseudo-labelling criteria.

Each iteration consisted of a full training cycle on $\mathcal{D}_T^{\text{train}}$ using a new instance of the selected architecture. At the end of the iteration, the best checkpoint on $\mathcal{D}_T^{\text{val}}$ was used to generate predictions and compute new thresholds to obtain pseudo-labels for the next iteration.

For the first iteration only, thresholds $\tau_c$ were manually adjusted for *Soil* and *Water* classes by $\pm 0.3$: reducing soil to $0.1$ (from $0.4$) to include more points, and increasing water to $0.9$ (from $0.6$) to retain only high-confidence points. These adjustments preserved under-represented classes in the pseudo-labelled data while managing class imbalance.

# 5. Results

## 5.1. Experimental Settings

Models were trained using NVIDIA A100 GPUs with Multi-Instance GPU (MIG) partitioning, specifically utilizing 20GB and 40GB MIG slices. Training ran for 200 epochs with a batch size of 4 and a maximum of $65,536$ points per batch element, balancing accuracy and memory constraints. An initial learning rate of $0.001$ was applied for model optimization. All the experiments were globally evaluated using mean Intersection over Union (mIoU) and F1-score. For Turin3D, IoU per class was also considered to provide more granular performance analysis.

## 5.2. Transfer Learning

Transfer learning experiments, reported in Table 2, revealed significant performance variations across architectures when generalizing to Turin3D.

RandLA-Net performed best ($38.73$ mIoU with augmentation) but still showed a substantial drop from its performance on literature datasets ($67.39$ mIoU), highlighting cross-city generalization challenges. Data augmentation improved overall performance, especially for *Vegetation* and *Buildings*, though *Soil* classification degraded.

Other architectures struggled significantly: Point Transformer ($7.15$ mIoU) performed worse with augmentation than without, while SparseConv achieved only $6.48$ mIoU. These models particularly struggle with street elements and water classes, often failing completely. *Vegetation* appears to be the most transferable class across all models, likely due to its more consistent appearance across different urban environments. Indeed, the complete failures in class transferability across different architectures, particularly evident for *Water* and *Street Elements*, likely stem from significant variations in how these elements appear in different urban environments, combined with potential annotation inconsistencies between datasets. For instance, *Water* features in Turin3D may have distinct geometric or reflectance properties compared to those in the training datasets, rendering them unrecognizable to models without domain-specific adaptation. Nevertheless, RandLA-Net with augmentation achieves $8.12$ IoU on Water, suggesting that data augmentation can partially mitigate this situation.

These findings established RandLA-Net as the best architecture for subsequent semi-supervised learning experiments.

## 5.3. Semi-Supervised Learning

Semi-supervised learning experiment, reported in Table 3, were conducted using RandLA-Net with augmentation, comparing fixed and adaptive confidence thresholding strategies across multiple iterations. The initial results of both approaches represent a $+6.50$ mIoU improvement over the transfer learning baseline. The adaptive thresholding approach showed consistent improvement, with mIoU peaking at $48.49$ in the second iteration and F1 score reaching $74.45$ in the third iteration. *Water* segmentation demonstrated marked improvement with adaptive thresholding, increasing from $17.88$ to $30.51$ IoU ($+12.63$) across iterations, while *Vegetation* maintained consistently high performance around $87$ IoU and *Soil* steadily improved to $32.89$ IoU. In contrast, the fixed thresholding strategy ex-

| Model | Augmentation | $\mathcal{D}_T^{test}$ (Turin3D) | | | | | | | | $\mathcal{D}_S$ Test | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Soil* | *Terrain* | *Vegetation* | *Building* | *Street Elements* | *Water* | **mIoU** | **F1** | **mIoU** | **F1** |
| RandLA-Net [7] | ✗ | 22.44 | 43.43 | 53.98 | 55.08 | 11.04 | 0.0 | 30.99 | 43.04 | 31.24 | 43.28 |
| | ✓ | 10.05 | 43.75 | 81.42 | 72.36 | 16.70 | 8.12 | **38.73** | **49.42** | **67.39** | **78.59** |
| Point Transformer [17] | ✗ | 9.40 | 16.19 | 22.65 | 10.45 | 0.0 | 0.0 | 16.97 | 9.86 | 14.76 | 19.72 |
| | ✓ | 1.70 | 2.56 | 28.87 | 9.59 | 0.0 | 0.0 | 7.15 | 11.88 | 13.87 | 18.05 |
| SparseConv [5] | ✗ | 8.84 | 0.0 | 29.65 | 0.0 | 0.0 | 0.0 | 6.41 | 20.67 | 12.39 | 30.77 |
| | ✓ | 7.73 | 0.0 | 31.16 | 0.0 | 0.0 | 0.0 | 6.48 | 30.93 | 12.32 | 31.30 |

Table 2. Results for Transfer learning experiments, with and without augmentations, evaluated on both test sets of literature selected datasets ($\mathcal{D}_S$) and labeled test set of Turin3D ($\mathcal{D}_T^{test}$), considering mIou and F1 score. For Turin3D also IoU per class is reported.

| Pseudo-Label Thresholding | Iteration | *Soil* | *Terrain* | *Vegetation* | *Building* | *Street Elements* | *Water* | mIoU | F1 |
|---|---|---|---|---|---|---|---|---|---|
| Fixed Confidence per iteration | 1 | 26.17 | 50.26 | 85.38 | 73.94 | 17.77 | 17.88 | 45.23 | 57.67 |
| | 2 | 32.26 | 34.38 | 86.52 | 66.50 | 27.55 | 0.0 | 41.20 | 63.16 |
| | 3 | 26.29 | 32.64 | 68.60 | 55.74 | 7.32 | 0.0 | 31.76 | 51.49 |
| Adaptive Confidence per iteration | 1 | 26.17 | 50.26 | 85.38 | 73.94 | 17.77 | 17.88 | 45.23 | 57.67 |
| | 2 | 30.28 | 52.29 | 87.80 | 77.68 | 19.69 | 23.27 | **48.49** | 61.12 |
| | 3 | 32.89 | 50.88 | 87.62 | 69.92 | 18.01 | 30.51 | 48.30 | **74.45** |

Table 3. Results for experiments with Semi-supervised learning with fixed and adaptive confidence per iteration, using RandLA-Net with Augmentations, evaluated on test set of Turin3D ($\mathcal{D}_T^{test}$ ) considering IoU per class, mIoU and F1 score.

hibited progressive performance deterioration, declining to 31.76 mIoU by the third iteration. Notably, *Water* classification completely disappeared after the first iteration with fixed thresholds, highlighting a critical limitation of this approach: low-confidence classes become progressively excluded from pseudo-labels, creating a self-reinforcing cycle of degradation.*Building* and *Terrain* classes showed performance fluctuations with both approaches, indicating challenges in generating consistent pseudo-labels for these classes with complex and diverse urban appearances. Results on *Water* highlight the efficacy of the adaptive approach: the adaptive method successfully maintained and improved water classification, unlike the fixed approach where this class disappeared entirely.

In conclusion, the semi-supervised learning methodology with adaptive confidence thresholding yielded a 9.76 absolute mIoU improvement over the transfer learning baseline. This demonstrates the effectiveness of leveraging unlabeled target domain data through iterative pseudo-labeling for cross-city point cloud segmentation.

## 6. Conclusions and Future Works

In this work, we introduced *Turin3D*, a new aerial LiDAR dataset for urban semantic segmentation, and evaluated different learning strategies to address the challenge of label scarcity. Our experiments compared transfer learning and semi-supervised learning techniques, demonstrating how the latter methods yielded superior segmentation performance by effectively leveraging the unlabeled training set. These results highlight the potential of data-efficient learning strategies in large-scale urban point cloud analysis, where full annotation is often impractical.

Despite these promising results, several aspects remain open for future research. First, an extended annotation effort on the training set would allow for a fully supervised benchmark, providing a more precise evaluation of different learning strategies. Additionally, future work could explore the application of more recent deep learning architectures for point cloud segmentation, potentially improving performance over the baseline models used in this study. Another important direction is the investigation of domain adaptation techniques specifically designed for point cloud segmentation, which could enhance model generalization to unseen urban environments, further reducing the reliance on extensive manual labelling. Lastly, a further step for future research could consist of expanding the proposed taxonomy to incorporate a more detailed classification of urban elements, thereby improving its descriptive power and applicability across a broader range of urban scenarios. By making Turin3D publicly available, we aim to support further research in semi-supervised and transfer learning strategies for urban LiDAR segmentation, providing a challenging yet realistic benchmark for the community.

## Acknowledgements

This work was carried out in the context of Horizon Europe project UP2030 (G.A. n.101096405), Space IT Up project funded by the Italian Space Agency (ASI) and the Ministry of University and Research (MUR) under contract n. 2024-5-E.0 - CUP n. C53C24000530005 and funded by the European Union - NextGenerationEU, Mission 4 Component 2 - ECS00000036 - CUP B13D21011790006

## References


[1] Gülcan Can, Dario Mantegazza, Gabriele Abbate, Sébastien Chappuis, and Alessandro Giusti. Semantic segmentation on swiss3dcities: A benchmark study on aerial photogrammetric 3d pointcloud dataset. *Pattern Recognition Letters*, 150: 108–114, 2021. 3, 4, 6

[2] Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 429. BMVA Press, 2022. 3, 4, 6

[3] Weixiao Gao, Liangliang Nan, Bas Boom, and Hugo Ledoux. Sum: A benchmark dataset of semantic urban meshes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179:108–120, 2021. 3, 4, 6

[4] Charles Gaydon, Michel Daab, and Floryne Roche. Fractal: An ultra-large-scale aerial lidar dataset for 3d semantic segmentation of diverse landscapes. *ArXiv*, abs/2405.04634, 2024. 3, 4, 6

[5] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9224–9232, 2018. 3, 6, 8

[6] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *CoRR*, abs/1912.12033, 2019. 3

[7] 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. 2, 3, 6, 8

[8] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4977–4987, 2021. 2, 4, 6

[9] 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, 2021. 2, 4, 6

[10] Franz Leberl, A. Irschara, T. Pock, Philipp Meixner, Michael Gruber, Susanne Scholz, and Alexander Wiechert. Point clouds: Lidar versus 3d vision. *Photogrammetric Engineering and Remote Sensing*, 76:1123–1134, 2010. 2

[11] Francesco Nex and Fabio Remondino. Uav for 3d mapping applications: a review. *Applied Geomatics*, 6(1):1–15, 2014. 2

[12] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. 3

[13] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, 2017. 3

[14] Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. Toronto-3D: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 202–203, 2020. 3, 4, 6

[15] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6410–6419. IEEE, 2019. 4

[16] Wuming Zhang, Jianbo Qi, Peng Wan, Hongtao Wang, Donghui Xie, Xiaoyan Wang, and Guangjian Yan. An easy-to-use airborne lidar data filtering method based on cloth simulation. *Remote Sensing*, 8(6), 2016. 6

[17] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16239–16248. IEEE, 2021. 2, 4, 6, 8

[18] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 4