ECLAIR: A High-Fidelity Aerial LiDAR Dataset for Semantic Segmentation

Iaroslav Melekhov* [1,2] Anand Umashankar* [1] Hyeong-Jin Kim[1] Vladislav Serkov[1] Dusty Argyle[1]

[1]Sharper Shape (https://sharpershape.com/) [2]Aalto University

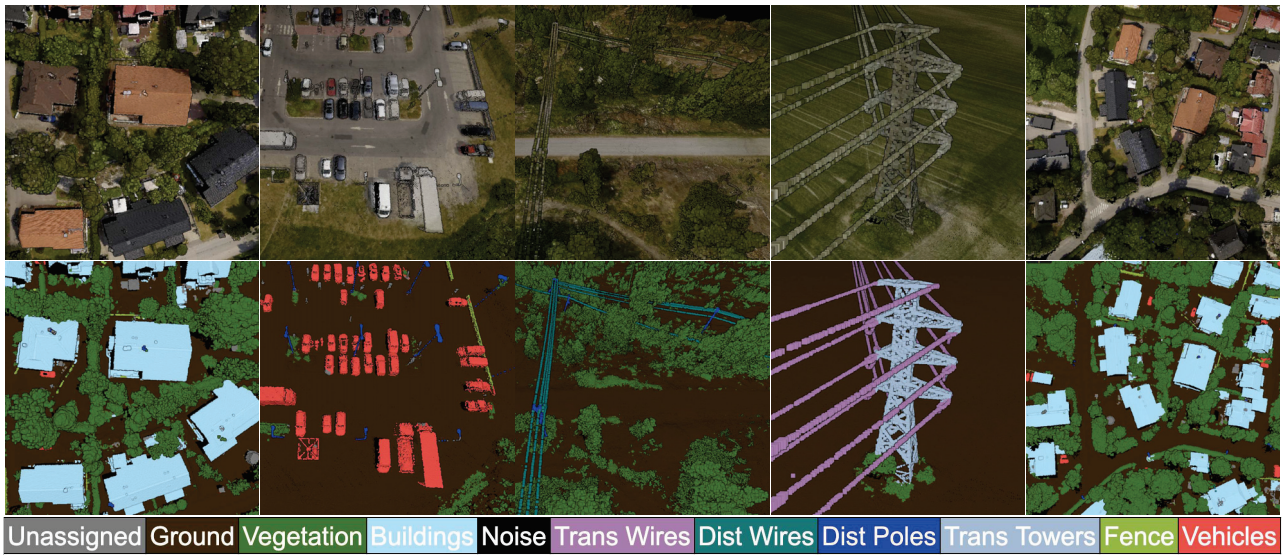

Figure 1. **Overview of the proposed ECLAIR dataset.** We introduce ECLAIR, a new outdoor large-scale aerial LiDAR dataset. It covers a total area of more than 10 square kilometers encompassing 11 semantic classes. The long-tail accurate annotations enable fine-grained semantic understanding. Different semantic classes are labeled by different colors.

Abstract

*We introduce ECLAIR (**E**xtended **C**lassification of **L**idar for **AI R**ecognition), a new outdoor large-scale aerial LiDAR dataset designed specifically for advancing research in point cloud semantic segmentation. As the most extensive and diverse collection of its kind to date, the dataset covers a total area of 10km^2 with close to 600 million points and features eleven distinct object categories. To guarantee the dataset's quality and utility, we have thoroughly curated the point labels through an internal team of experts, ensuring accuracy and consistency in semantic labeling. The dataset is engineered to move forward the fields of 3D urban modeling, scene understanding, and utility infrastructure management by presenting new challenges and potential applications. As a benchmark, we report qualitative and quantitative analysis of a voxel-based point cloud segmentation approach based on the Minkowski Engine. We release the dataset as open-source and it can be accessed at https://github.com/SharperShape/eclair-dataset*

*Indicates equal contribution. Correspondence by email: firstname.lastname@sharpershape.com

1. Introduction

Recent breakthroughs in the field of deep learning [18, 26, 36] are attributed to factors such as the availability of vast and extensive datasets and have enabled models to generalize effectively across diverse applications. However, such progress has not been mirrored in the domain of 3D LiDAR. For instance, the DALES dataset [37], comprising forty scenes, amounts to a few gigabytes. In contrast, the CommonCrawl dataset, one of the largest in Natural Language Processing (NLP) utilized by the LLaMA model [36], spans approximately 6 petabytes. Similarly in Computer Vision, the SegmentAnything[18] dataset occupies 11.3 terabytes of storage. The inherent nature of 3D datasets, with aspects such as the dimensionality and point density, also contributes to this disparity. These characteristics pose unique challenges for their collection, labeling, and management. This paper aims to bridge this gap with the aspiration of in-

creasing the availability of point cloud dataset with a more extensive and rich dataset comparable in size to DALES, and facilitate further research into the deep learning models and their quality.

Outdoor 3D scene understanding is fundamental to many applications in computer vision, including autonomous driving, robotics, Augmented and Virtual Reality (AR / VR) [31, 46]. The last several years, modern machine learning techniques have advanced state-of-the-art scene understanding algorithms for object detection, depth estimation, semantic and instance segmentation, 3D reconstruction, and more. Most of these approaches are enabled through a diverse set of real and synthetic RGB-(D) datasets [1, 13, 33].

The demand for diverse and accurately annotated datasets captured at a large scale is becoming more critical in point cloud semantic segmentation techniques. These machine learning-based techniques are instrumental across a multitude of applications ranging from autonomous driving to urban planning. At present, existing datasets for point cloud semantic segmentation exhibit a significant inclination towards scenarios predominantly related to autonomous vehicles [3, 5, 15, 24, 34] that use Mobile Laser Scanning (MLS) or Terrestrial Laser Scanning (TLS) systems to collect the data. While these datasets have advanced perception systems for self-driving cars, their scope, largely confined to vehicle-centric perspectives, introduces a notable gap in the diversity and coverage environments. This limitation particularly overlooks the potential of aerially captured data. Other datasets [1, 11, 30, 44] provide detailed scans of interiors for tasks such as object recognition, semantic segmentation, and novel view synthesis. While these datasets have been instrumental in advancing indoor mapping and navigation systems, they are inherently limited in their applicability to outdoor, large-scale environments due to their specific focus on indoor spaces.

Airborne LiDAR Scanning (ALS) systems generate point cloud data that significantly differ from those from self-driving and indoor datasets. The orientation of ALS sensors is typically close to a nadir view, leading to distinct occlusions compared to ground-based scanning systems. Aerial LiDAR data collection is often more costly than mobile LiDAR due to expenses associated with aerial flights [28, 37, 47]. To address the high costs associated with hardware, [6, 16, 20] propose generating point clouds from high-quality aerial images captured by a UAV-based mapping system using photogrammetry. Although cost-efficient, the quality of the reconstructed point cloud significantly depends on the discriminative performance of local image descriptors that can struggle to handle different lighting and weather conditions. The ALS systems provide advantages such as providing a more uniform density of point clouds and covering a broader area coverage during data collection. Moreover, it enables data collection in areas where terrestrial travel is challenging. These unique features make ALS ideal for urban planning and surveying applications.

The dataset we propose, named ECLAIR (**E**xtended **C**lassification of **L**idar for **AI** **R**ecognition), consists of a large-scale point cloud collected from a region in the city of Espoo, Finland. It covers a contiguous area of more than 10 square kilometers consisting of more than half a billion points captured by a long-range high-accuracy LiDAR. The focus of the data capture has been to cover the electrical transmission lines; consequently, the point clouds follow this network. A comparison of ECLAIR with some of the existing point cloud datasets is presented in Tab. 1. In addition to the raw data, we provide accurate ground truth and pseudo labels, and demonstrate their usability in a downstream supervised learning task: point cloud semantic segmentation. In contrast to DALES [37], ECLAIR further uses high-resolution nadir images to provide colorized point clouds. We describe the dataset capturing pipeline as well as the point cloud colorization process in Sec. 3. Along with color, the dataset also includes intensity, the return number, and the number of returns as features. Lastly, the proposed dataset not only shares similarities with existing datasets but also introduces unique distributions that, when combined with other datasets, facilitate large-scale generalized representational learning [41].

In summary, we make the following contributions: 1) We introduce *ECLAIR*, a new outdoor, large-scale aerial LiDAR dataset with point-wise semantic annotations; 2) The proposed dataset enables training and benchmarking point cloud segmentation approaches on large-scale, real-world scenes captured by a high-quality aerial LiDAR; 3) We thoroughly evaluate one of the existing voxel-based point cloud semantic segmentation approaches (*i.e.*, the Minkowski Engine [10]) on the proposed dataset and discuss quantitative results.

2. Related Work

Deep learning approaches for 3D semantic understanding require diverse, large-scale datasets in order to generalize to new scenes. Here we give a brief introduction to existing datasets, compare them with ECLAIR, and provide an overview of current methods for point cloud semantic segmentation.

2.1. Semantic Understanding of 3D Areas

Existing datasets for large-scale point cloud segmentation can be widely categorized into three groups: indoor scene-level 3D datasets, outdoor road-level point cloud datasets, and urban-level aerial LiDAR datasets.

Indoor scene-level 3D datasets. Early datasets in this category, such as SUN RGB-D [33], NYUv2 [32], and

Dataset	Category	Year	Spatial Size, m / Area, m^2	# Classes	# Points	# RGB	Sensor
S3DIS [1]	Indoor scene-level	2017	$6 \times 10^3\ m^2$	13	273M	✓	Matterport
ScanNet [11]		2017	$1.1 \times 10^5\ m^2$	20	242M	✓	RGB-D
ScanNet++ [44]		2023	$1.5 \times 10^4\ m^2$	1000	-	✓	RGB-D
Semantic3D [14]	Outdoor road-level	2017	-	8	4000M	✓	TLS
SemanticKITTI [3]		2019	$39.2 \times 10^3\ m^2$	25	4549M	✗	MLS
Toronto-3D [35]		2020	$1 \times 10^3\ m^2$	8	78.3M	✓	MLS
SemanticPOSS [24]		2020	-	14	216M	✗	MLS
ISPRS [28]	Aerial urban-level	2012	-	9	1.2M	✗	ALS
DublinCity [47]		2019	$2 \times 10^6\ m^2$	13	260M	✗	ALS
Campus3D [20]		2020	$1.58 \times 10^6\ m^2$	24	937.1M	✓	P
SensatUrban [16]		2020	$7.64 \times 10^6\ m^2$	13	2847M	✓	P
Swiss3DCities [6]		2020	$2.7 \times 10^6\ m^2$	5	226M	✓	P
DALES [37]		2020	$10 \times 10^6\ m^2$	8	505M	✗	ALS
ECLAIR (ours)		2024	$10.3 \times 10^6\ m^2$	11	582M	✓	ALS

Table 1. **Comparison of datasets.** We compare existing datasets in terms of area of coverage, point density, and the sensor type. While the coverage area in DALES is similar to ours, the proposed dataset has more semantic classes and additionally provides colorized point clouds. Similar to [16], we use the following notation: MLS - Mobile Laser Scanning system; TLS - Terrestrial Laser Scanning system; ALS - Aerial Laser Scanning system; P - photogrammetry.

S3DIS [1], represent RGB-D sequences captured by short-range depth scanners with low resolution and limited semantic annotations. Other datasets [7, 11, 30] provide annotation at scale, but the performance on long-tail classes is limited by the resolution of ground truth geometry from laser scans. ARKitScenes [2] and ScanNet++ [44] address this limitation by incorporating both RGB images and high-resolution 3D scene geometry captured by lasers. They provide sparse (bounding boxes) and dense semantic annotations respectively.

Outdoor road-level 3D data. This group of datasets is related to autonomous driving applications in which the data is captured by a LiDAR scanner together with RGB cameras mounted on a vehicle [3, 5, 8, 24, 25, 29, 34, 35]. The mobile LiDAR datasets, with their low-angle perspective and emphasis on driving-related segmentation tasks, often result in occlusions inside the point clouds, *e.g.*, missing roofs of buildings. While these datasets fulfill their primary purpose, they fall short for use in other domains, such as public utility asset management and urban planning.

Urban-level aerial datasets. These datasets are pivotal for advancing research and applications in the fields of remote sensing, environmental monitoring, and autonomous navigation. They have primarily been obtained by aerial LiDARs [28, 37, 43, 47] or by using photogrammetry [6, 16, 20]. In contrast to DALES [37], ECLAIR provides colorized, large-scale point clouds including high-resolution 3D geometry along with accurate semantic labels and the number of LiDAR returns for each point.

2.2. 3D Semantic Learning

In general, deep learning based, point cloud semantic segmentation methods fall into three main categories based on their approach to modeling point clouds: projection-based, voxel-based, and point-based methods. Projection-based strategies convert 3D points onto different image planes, leveraging 2D CNN architectures to extract features [9, 19, 21]. Voxel-based methods [10, 23, 39], on the other hand, turn point clouds into uniform voxel grids, making 3D convolution operations more manageable and improving their efficiency with sparse convolution techniques. In contrast, point-based approaches deal with point clouds directly with a notable recent trend towards adopting transformer-based architectures [27, 40, 42, 45]. In this work, we thoroughly evaluate one of the voxel-based approaches, using the Minkowski Engine [10], on the ECLAIR dataset and identify a number of key challenges revealed by our dataset.

3. ECLAIR: Dataset Creation

This dataset was created through a multi-step process. In Sec. 3.1, data capture details the sensors and parameters employed in acquisition. Subsequent data processing, described in Sec. 3.2, involved necessary transformations and preparation for analysis. Class specifications, crucial for semantic analysis, are defined in Sec. 3.3. Data curation included a manual examination step, outlined in Sec. 3.4. Finally, data visualization tools were utilized to illustrate the data and accelerate quality control, as discussed in Sec. 3.5.

3.1. Data Capture

The data presented are captured using our proprietary sensor system built in-house by Sharper Shape and mounted on a helicopter. It is a single lightweight multi-sensor system, capable of collecting the data required for utility inspection and analysis via helicopter. It is equipped with the following hardware: a) Long-range, survey grade, high accuracy LiDAR coupled with highly precise GNSS and FOG IMU sensors; b) High resolution RGB cameras capturing oblique and ortho imagery; c) Push broom hyperspectral cameras providing a broad spectrum of wavelength along the flight path; and, d) Ambient temperature and humidity sensors.

All sensors were calibrated according to manufacturer specifications before data processing. The system has the capability to attach more sensors such as 4-band, ultraviolet sensors, etc. It has also been tested under various conditions and has been a reliable system enabling data capture across multiple projects under Sharper Shape. The data was collected at a flight height of 100 meters and speed of 40 knots. With 600 PRR (Pulse Repetition Rate) and 234.5 lines per second, the LiDAR has a point density of 50 points/m^2 and a swath width of 328 meters. The LiDAR was calibrated using multiple overlap captures with a standard deviation error of less than 2 cm.

3.2. Data Processing

The data preparation process starts with the utilization of the RiProcess software to convert raw files into the LAZ file format. Subsequently, the LAZ files are fed into the tiling pipeline in which the point clouds are partitioned into smaller 100×100 m tiles. This is facilitated by a tool developed internally by Sharper Shape to enhance data manageability.

Following tiling, the point clouds undergo a colorization process. For each tile, relevant images are selected by identifying areas of overlap. Subsequently, utilizing the camera's parameters, each point from the point cloud is projected from 3D onto 2D space. Color information from the corresponding projected location within the selected image is then extracted and applied to the point. In instances where multiple images contribute to a specific point, color averaging is performed across these images. However, this averaging mechanism may introduce inaccuracies, particularly in the coloring of thin structures such as powerlines. It is also worth noting that the coverage of images is lower than that of the point cloud, and hence approximately 20% of the points lack color information.

Following colorization, the data undergoes point cloud segmentation utilizing a deep learning model. The employed proprietary model produces a total of 30 classes with a predominant focus on classes pertinent to electrical infrastructure. These classes are subsequently remapped to a set of 11 classes that are presented here. This remapping yields a notable reduction in classification errors, allowing the remapped version of the dataset to be effectively utilized for model training with only minimal manual intervention.

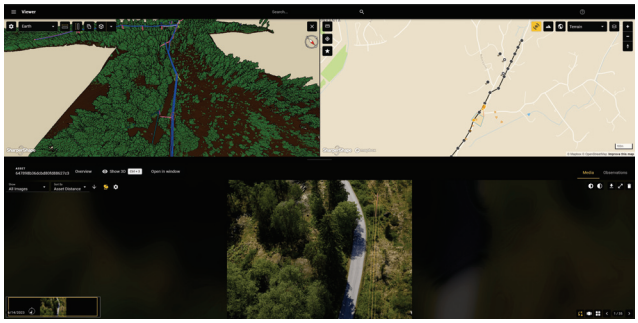

Figure 2. **Point cloud visualization**. CORE viewer combines a view of point clouds colored based on classifications, a map view, and image view. The point clouds and images also show the vector data of objects in 3D and 2D, respectively.

This comprehensive data preparation pipeline involves sequential processes aimed at converting, tiling, enhancing, and classifying point cloud data for various downstream applications including vectorization, vegetation encroachment analysis, etc. The coordinate system of the data is WGS 84 / UTM zone 35N (EPSG:32635).

3.3. Class Specifications

We identify the following list of semantic classes with each corresponding label ID provided in parentheses:
Ground (2) : All points representing the Earth's surface, including, soil, pavement, roads, and the bottom of water bodies.
Vegetation (3) : All points representing organic plant life, ranging from trees, low shrubs, and tall grass of all heights.
Buildings (4) : Man-made structures characterized by roofs and walls, encompassing houses, factories, and sheds.
Noise (5) : Sporadic points suspended in air or underground.
Transmission Wires (6) : High-voltage wires for long-distance transmission from power plants to substations. Either directly connected to transmission towers or poles. Also includes transmission ground wires.
Distribution Wires (7) : Lower-voltage overhead distribution wires distributing electricity from substation to end users. Includes span guy wires and communication wires.
Poles (8) : Utility poles used to support different types of wires or electroliers. These can include poles with either transmission or distribution wires. Down guy wires, crossarms and transformers are also included in this class.
Transmission Towers (9) : Large structures supporting transmission wires with the distinct characterisation of steel lattices and cross beams.
Fence (10) : Barriers, railing, or other upright structure, typically of wood or wire, enclosing an area of ground.

Vehicle (11) : All wheeled vehicles that can be driven.
Unassigned (1) : This category serves as a catch-all for non-subject points. Anything that is not on the class list is classified as Unassigned. These include wooden pallets, trash, structures not large or strong enough to put under buildings (tents, boulders, etc.), and house antennas.

3.4. Data Quality Control

The dataset classifications are derived from fully automated processes which may contain errors. To discern accurate classifications within the dataset, manual verification of tiles was conducted by Sharper Shape's internal data curation team of annotation experts in the power utility domain. The dataset is bifurcated into two primary categories: "Ground Truth" and "Pseudo-Labels". Within each tile, if misclassifications for a particular object class (excluding *Ground*, *Vegetation*, *Unassigned*, and *Noise*) exceed 10 points, the tile is categorized as a pseudo-label. Conversely, tiles devoid of misclassifications or with misclassifications totaling fewer than 10 points for object classes are allocated to the "Ground Truth" category. Overall, out of the 1246 tiles in the dataset, 624 are classified as ground truth, and 622 are categorized as pseudo labels.

3.5. Data Visualization

In order to facilitate quality control and to easily visualize the data, a software platform named Sharper CORE was used. This has been developed internally by Sharper Shape. It enables the use of geographical information with point clouds and images, allowing for sensor fusion. This enables comprehensive data analysis to make sure the quality aligns well with the expectation during quality control tasks.

Fig. 2 shows a screenshot from the CORE web software, showcasing various views available within the interface. The illustration highlights the integration of data from different modalities into a unified view, offering extensive contextual information during quality control tasks. In addition to visualizing point clouds and images, CORE also supports visualizing vector overlays which help with assessing quality of object classes such as power lines, poles, towers, etc. as illustrated in Fig. 3. This context is valuable since we no longer have to navigate every portion of point cloud data to evaluate our deep learning models and data.

CORE is built on top of specialized GIS databases which store information about the vector models that can be used to do further analysis and reporting. This allows for in-depth component inventory and environmental analysis in a structured and meaningful way so that data can be queried to build specific datasets. An example of inventory is shown in Fig. 4. In addition, CORE scales with large amounts of data seamlessly and allows custom rendering settings and cloud-based data serving. As a result, numerous individuals with limited technical proficiency in point cloud data were able to participate in the quality control process, collaborating simultaneously online with the aid of basic instructions.

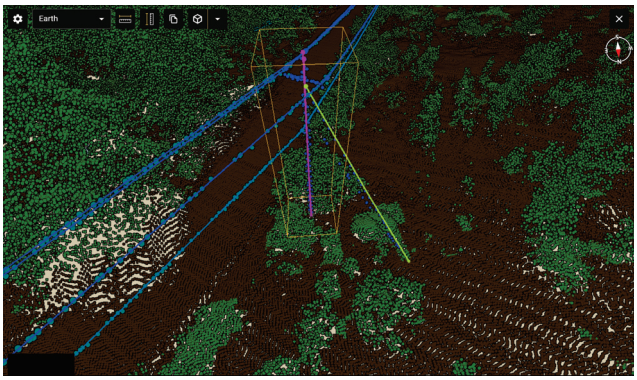

Figure 3. **Point cloud inspection.** 3D Point Cloud Navigation/Editing View provided by CORE.

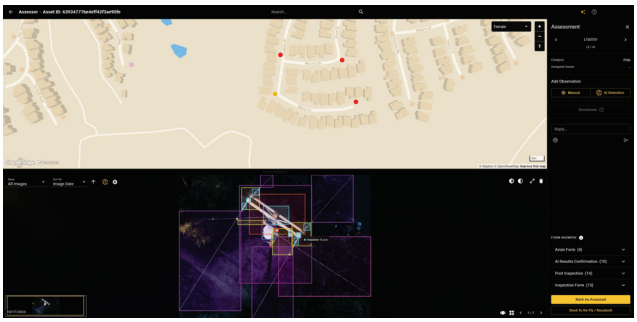

Figure 4. **Component inventory.** Detection and inventory of specific components in CORE based on multimodal data inputs.

4. Experimental Evaluation

We first present statistics of the proposed dataset in Sec. 4.1 and discuss evaluation metrics in Sec. 4.2. Next, we analyze different design choices and perform ablation studies of the Minkowski Engine-based baseline network with 4 different architectures and various sets of point features. We also provide details of loss functions that can efficiently handle class imbalance and investigate the influence of pseudo labels on semantic segmentation performance in Sec. 4.3. Finally, we provide quantitative and qualitative results of the baseline model on the proposed dataset.

4.1. Statistics of ECLAIR

The proposed dataset comprises 1246 tiles, each covering an area of 100×100 square meters. To ensure a robust evaluation framework, the dataset is divided into train, validation, and test splits following the standard proportions of 70%, 10%, and 20% respectively. The validation and test splits consist only of the ground truth tiles to ensure that the metrics generated are reliable and consistent.

Features	macro F1	macro IoU (%)	per-class F1 / IoU (%) Ground	Vegetation	Buildings	Noise	Trans. wires	Dist. wires	Dist. poles	Trans. towers	Fence	Vehicles	Unassigned
f_{int}	0.841	76.72	**0.99** / 98.23	**0.99** / 98.32	0.92 / 85.87	0.81 / 68.09	**0.99** / 98.86	0.90 / 83.19	0.73 / 57.50	0.94 / 90.10	**0.85** / 73.88	0.86 / 76.35	0.23 / 13.58
f_{return}	0.842	76.63	**0.99** / 98.30	**0.99** / 98.40	0.93 / 86.82	**0.83** / **71.04**	**0.99** / 98.87	0.90 / 82.26	0.69 / 52.27	**0.96** / **91.62**	0.83 / 70.72	0.86 / 75.34	0.29 / 17.24
f_{color}	0.838	75.69	**0.99** / **98.33**	**0.99** / **98.43**	0.92 / 86.43	0.81 / 68.61	**0.99** / **99.20**	**0.92** / **85.65**	0.69 / 52.86	0.94 / 88.08	0.84 / 72.65	0.77 / 61.94	**0.34** / **20.46**
$f_{\text{int}}+f_{\text{color}}$	0.828	74.80	**0.99** / 98.32	**0.99** / **98.43**	0.92 / 85.43	0.81 / 67.75	**0.99** / 98.80	0.90 / 82.13	0.68 / 51.02	0.95 / 90.01	**0.85** / **74.50**	0.76 / 74.50	0.26 / 15.14
$f_{\text{return}}+f_{\text{color}}$	0.829	74.64	**0.99** / 98.18	**0.99** / 98.29	0.92 / 84.38	0.81 / 68.61	**0.99** / 98.68	0.89 / 79.80	0.70 / 53.23	0.93 / 87.63	0.81 / 68.66	0.81 / 68.03	0.27 / 15.52
$f_{\text{int}}+f_{\text{return}}$	**0.848**	**77.35**	**0.99** / 98.27	**0.99** / 98.36	**0.94** / **87.73**	0.82 / 69.02	**0.99** / 98.51	0.89 / 80.69	**0.74** / 58.65	0.95 / 90.87	0.83 / 70.39	**0.89** / **80.49**	0.30 / 17.89
$f_{\text{int}}+f_{\text{return}}+f_{\text{color}}$	0.843	76.48	**0.99** / 98.27	**0.99** / 98.41	0.88 / 78.76	**0.83** / 70.89	**0.99** / 98.95	**0.92** / 84.50	**0.74** / **59.06**	0.95 / 88.36	**0.85** / 73.62	0.85 / 73.87	0.28 / 16.51

(a) Point features. We compare the three different point features: intensity f_{int}; the return number and the number of returns f_{return}; the color f_{color} and their combinations (*cf*. Sec. 4.3) and report macro and per-class F1 and IoU metrics. We observe that the combination of intensity and return features achieves the best segmentation results.

Loss function	macro F1	macro IoU (%)	per-class F1 / IoU (%) Ground	Vegetation	Buildings	Noise	Trans. wires	Dist. wires	Dist. poles	Trans. towers	Fence	Vehicles	Unassigned
Minkowski+$\mathcal{L}_{\text{ce}}$	0.831	75.45	**0.99** / **98.36**	**0.99** / **98.46**	**0.93** / **87.78**	0.80 / 66.43	**0.99** / 98.73	0.90 / 80.00	0.62 / 45.08	0.94 / **89.06**	**0.85** / **73.72**	**0.88** / **77.99**	0.25 / 14.32
Minkowski+$\mathcal{L}_{\text{iwce}}$	0.726	63.70	**0.99** / 97.94	0.98 / 98.05	0.84 / 72.48	0.30 / 17.78	**0.99** / 98.85	0.92 / 85.84	0.50 / 32.89	0.91 / 82.61	0.60 / 43.30	0.75 / 60.27	0.19 / 10.72
Minkowski+$\mathcal{L}_{\text{fl}}$	**0.843**	**76.48**	**0.99** / 98.27	**0.99** / 98.41	0.88 / 78.76	**0.83** / **70.89**	**0.99** / **98.95**	**0.92** / **84.50**	**0.74** / **59.06**	**0.95** / 88.36	**0.85** / 73.62	0.85 / 73.87	**0.28** / **16.51**

(b) Loss functions. We train the Minkowski Engine [10] with different loss functions to handle class imbalance of the ECLAIR dataset. The following notation is used: $\mathcal{L}_{\text{fl}}$ - Focal loss; $\mathcal{L}_{\text{ce}}$ - Cross-Entropy loss; $\mathcal{L}_{\text{iwce}}$ - inverse weighted Cross-Entropy loss.

Training data	macro F1	macro IoU (%)	per-class F1 / IoU (%) Ground	Vegetation	Buildings	Noise	Trans. wires	Dist. wires	Dist. poles	Trans. towers	Fence	Vehicles	Unassigned
GT	0.561	46.45	0.99 / 97.10	0.99 / 97.10	0.62 / 44.60	0.68 / 51.09	0.96 / 91.86	0.58 / 40.84	0.11 / 05.92	0.35 / 21.10	0.61 / 43.50	0.30 / 17.66	0.0 / 0.0
Pseudo Labels	0.842	76.35	**0.99** / **98.27**	**0.99** / 98.36	**0.95** / **91.02**	0.80 / 66.55	**0.99** / 98.27	0.87 / 76.88	0.69 / 52.71	0.94 / **88.99**	**0.85** / 73.52	**0.87** / **76.62**	**0.31** / **18.59**
GT+Pseudo Labels	**0.843**	**76.48**	**0.99** / **98.27**	**0.99** / **98.41**	0.88 / 78.76	**0.83** / **70.89**	**0.99** / **98.95**	**0.92** / **84.50**	**0.74** / **59.06**	**0.95** / 88.36	**0.85** / **73.62**	0.85 / 73.87	0.28 / 16.51

(c) Training data. Comparing different annotation strategies (*cf*. Sec. 3.4), we observe that using only carefully curated tiles (GT) leads to poor semantic segmentation results due to a lack of data for rare classes (*cf*. Sec. 4.1). The pseudo labels combined with ground-truth significantly improve segmentation performance.

Training data	macro F1	macro IoU (%)	per-class F1 / IoU (%) Ground	Vegetation	Buildings	Noise	Trans. wires	Dist. wires	Dist. poles	Trans. towers	Fence	Vehicles	Unassigned
Res16UNet34A	0.793	69.66	**0.99** / 97.60	**0.99** / 98.01	0.62 / 45.11	0.80 / 66.55	0.99 / 97.81	0.86 / 75.22	0.63 / 45.59	0.86 / 74.84	0.83 / 70.23	**0.88** / **78.95**	**0.28** / **16.37**
Res16UNet14	0.797	70.61	**0.99** / 97.78	**0.99** / 98.16	0.62 / 44.93	0.82 / 69.97	**0.99** / 97.77	0.86 / 74.78	0.68 / 51.56	0.94 / 88.63	0.81 / 68.54	0.83 / 71.44	0.23 / 13.10
Res16UNet34C	0.843	76.48	**0.99** / **98.27**	**0.99** / **98.41**	0.88 / 78.76	0.83 / 70.89	**0.99** / 98.95	**0.92** / 84.50	0.74 / 59.06	0.95 / 88.36	**0.85** / **73.62**	**0.85** / **73.87**	**0.28** / **16.51**
Res16UNet14C	**0.845**	**77.29**	**0.99** / 98.18	**0.99** / 98.29	**0.92** / **84.51**	**0.84** / **72.33**	**0.99** / **99.05**	**0.93** / **86.34**	**0.75** / **60.23**	**0.96** / **91.34**	**0.85** / 73.54	0.84 / 72.79	0.24 / 13.54

(d) Network architectures. We explore different architectures of the Minkowski Engine and find that the *Res16UNet14C* backbone leads to the best semantic segmentation results in terms of macro F1 (IoU) and per-class metrics.

Table 2. **Ablation studies.** We provide ablation studies of the proposed dataset for different sets of point features in Tab. 2a, various loss functions to handle the class imbalance (*cf*. Tab. 2b); different point-wise annotation strategies in Tab. 2c; 4 different network architectures in Tab. 2d and report the F1 / IoU metrics. The best results for each category of experiments are marked in **bold**.

Fig. 5 provides statistics over all points of the proposed dataset revealing a significant imbalance across the semantic labels. Predominant categories such as Ground, Vegetation, and Buildings are overrepresented forming the majority of the dataset. In contrast, critical but less frequent categories (*e.g.*, Transmission Towers, Distribution Wires, Poles, and Vehicles) account for less than 1% of the total number of points, underscoring a challenge in achieving balanced representation. This imbalance reflects real-world conditions, presenting an opportunity to test the robustness and generalizability of point cloud classification models under skewed distribution scenarios. In addition to the carefully curated tiles (`GT` in Fig. 5), we also release a subset with pseudo labels generated by our proprietary point cloud classification model (`pseudo labels` in Fig. 5). Although having imperfect semantic labels, this dataset improves the model's generalization performance leading to better segmentation results (*cf*. Sec. 4.3).

4.2. Metrics

The F1 score and Intersection over Union (IoU) are both metrics used to evaluate the performance of semantic segmentation models, including those applied to point clouds. The F1 score is the harmonic mean of precision and recall. Precision measures the correctness of the positive predictions made by the model, while recall measures the model's ability to detect all actual positives. The F1 score is beneficial when the balance between precision and recall is required, especially when there is an uneven class distribution. It ensures that a model is not simply predicting the majority class. In contrast, IoU may be less informative in scenarios in which class imbalance affects the model's performance as the score primarily focuses on the spatial accuracy of the segmentation and not on the model's ability to detect rare classes. Therefore, in scenarios when it is essential to both identify every instance of a given class (recall) and ensure the accuracy of these detected instances (precision), particularly in context of significant class imbalance (*cf*. Sec. 4.1), we employ the F1 score as a key metric for evaluating the performance of semantic segmentation models. To be consistent with other works [20, 24, 37], we also report the IoU.

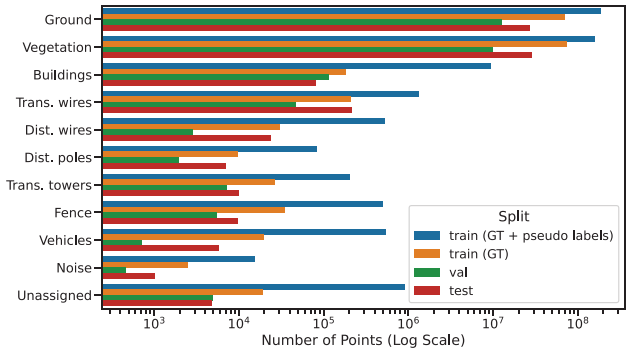

Figure 5. **The distribution of semantic classes.** We report the total number of points for each semantic category showing a high imbalance of the proposed dataset (note the logarithmic scale for the horizontal axis).

4.3. Ablation Studies

In this section, we conduct several ablation studies to examine the impact of various factors on segmentation performance. These factors include different architectural design choices using the Minkowski Engine [10], the effect of diverse point features and objective functions, and the quality of semantic ground-truth labels.

Point features. We consider the following features: a) Intensity f_{int}; b) Return data f_{return}: it includes the LiDAR return number and the number of returns; c) Color f_{color}; d) a combination of Intensity and Return data $f_{\text{int}} + f_{\text{return}}$; e) a combination of Intensity and Color $f_{\text{int}} + f_{\text{color}}$; f) a combination of Return data and Color $f_{\text{return}} + f_{\text{color}}$; g) all the features combined, *i.e.* $f_{\text{int}} + f_{\text{return}} + f_{\text{color}}$. The results presented in Tab. 2a show that combining intensity and return data achieves the best performance. The color feature seems quite powerful and can also improve semantic segmentation results for important classes, such as *Fence*.

Loss functions. Fundamentally, urban areas typically exhibit a highly skewed distribution of categories with a few dominant classes such as vegetation and ground occupying the majority of points, while smaller, yet critical, categories such as wires constitute a tiny fraction of points. The highly imbalanced distribution presents a major challenge from the ECLAIR dataset for accurate semantic segmentation (*cf*. Sec. 4.1). To address this issue, adopting more advanced loss functions is a common strategy [4, 22, 38]. We assess the efficacy of three off-the-shelf available loss functions, using the Minkowski Engine [10] as a baseline model. The evaluated objective functions include: cross-entropy $\mathcal{L}_{\text{ce}}$, weighted cross-entropy based on inverse frequency $\mathcal{L}_{\text{iwce}}$, and the focal loss $\mathcal{L}_{\text{fl}}$ [22]. The quantitative comparison of the baseline model with different loss functions is presented in Tab. 2b. We observe that using Focal loss leads to the best results, indicating that it can efficiently handle rare classes, *e.g.*, *Distribution Poles*.

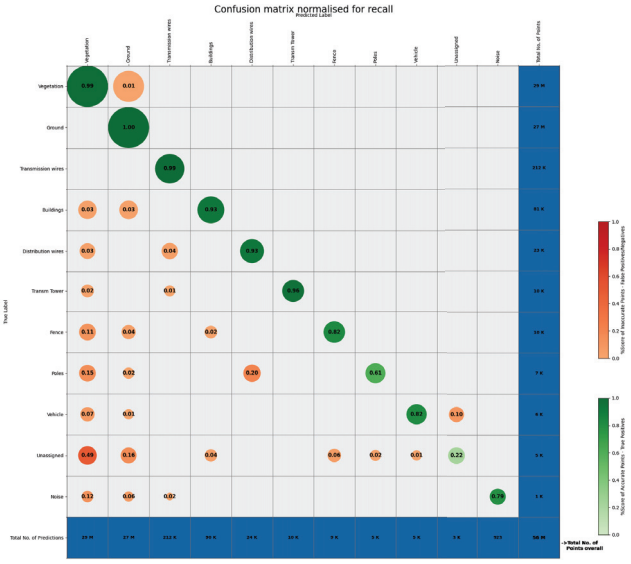

Figure 6. **Confusion matrix.** We report semantic segmentation results for our best model using a confusion matrix. Here, the size of circles corresponds to the total amount of points of each semantic class. The per-class F1 score is reported inside each circle. We find that the model performs well overall but falls short at segmenting rare, important classes, *e.g.*, *Poles*. Please zoom in to see the details.

Ground truth vs. pseudo labels. We ran experiments with the two point cloud annotation strategies discussed in Sec. 3.4 and Sec. 4.1. As shown in Tab. 2c, adding pseudo labels obtained by our proprietary point cloud segmentation approach leads to consistent improvements across all the classes compared to the manually verified ground truth data. Inaccurate labels introduce a form of noise into the training process, which can help the model learn to generalize better by forcing it to learn from a broader range of examples than it might from a smaller, perfectly labeled dataset (*cf*. Fig. 5: train (GT) vs. train (GT + pseudo labels)).

Network architecture. In order to evaluate the performance with different model capacities, we consider the Minkowski Engine [10] with different number of layers. We list the architecture configurations and report corresponding semantic segmentation performance in Tab. 2d. Our experiments show that using the Res16UNet14C architecture improves overall segmentation quality achieving the best per-class performance.

4.4. 3D Semantic Understanding

Technical details. The model architecture utilized in this study is ResUNet [12] implemented with the Minkowski Engine [10]. Conventional convolution layers are replaced with spatial convolution layers to accommodate point cloud data. Training of the model is conducted on AWS g5.12xlarge instances, each equipped with 4 NVIDIA

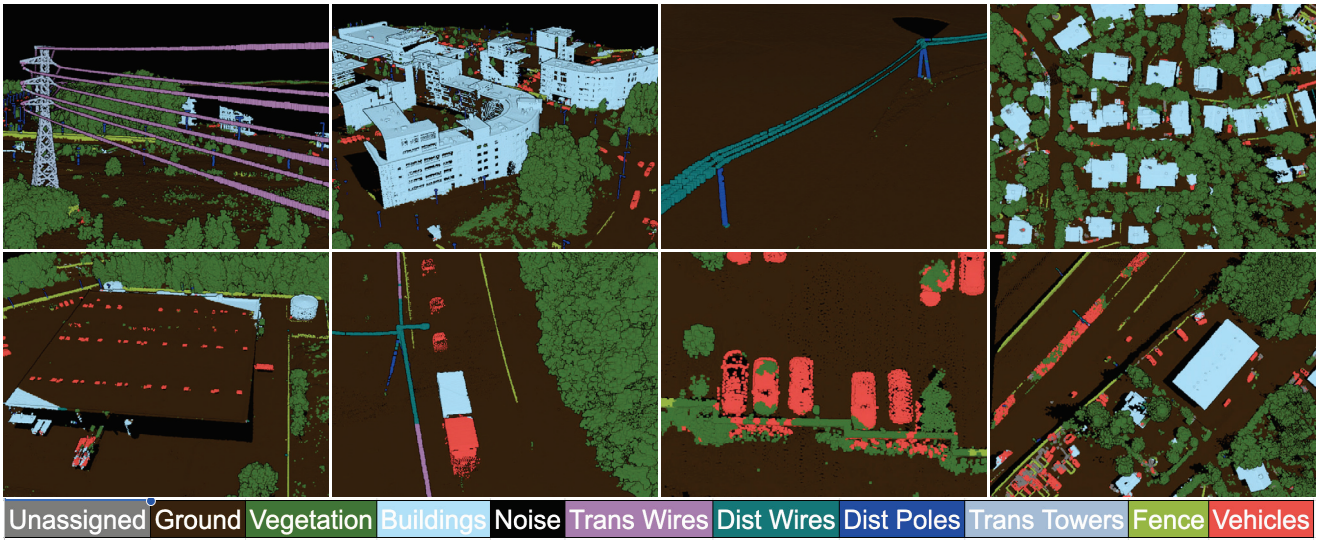

Figure 7. **Qualitative results of the best model.** We assess segmentation performance of the strongest baseline model on the test split of the proposed ECLAIR dataset. The top row demonstrates successful cases where the model performs well. In contrast, the bottom row highlights failure scenarios where the model produces noisy, inaccurate predictions.

A10G GPUs with 24 GB of vRAM per GPU. Given the memory-intensive nature of these networks, we maintain a batch size of 2 per GPU, resulting in an effective batch size of 8. Tiles are cropped to ensure a maximum size of 100m $\times$ 100m. Data augmentation techniques such as random coordinate scaling, random jitter, and random flip are applied, followed by normalization of coordinates before quantization. For features, RGB and intensity are scaled from 0 to 1 as floating-point values, and return number and number of returns are one-hot encoded before being fed into the network. The Adam optimizer [17] is employed with a learning rate set to 0.001, augmented by a step scheduler to adjust the learning rate every 10 steps. The voxel size is set to 0.05 with a normalisation factor of 10.0 applied to the coordinates before voxelisation.

Analysis. According to the ablation study performed in Sec. 4.3 and reported in Tab. 2, we chose the Res16Unet14C architecture as a backbone of the Minkowski Engine network. The model was trained using both point features, *i.e.* f_{int}+f_{return} with Focal Loss [22] achieving macro F1 of 0.848 and macro IoU of 77.35%. To further investigate segmentation performance of our baseline model, we provide a confusion matrix illustrated in Fig. 6. As can be seen, the model performs well successfully classifying the major categories, such as *Ground*, *Vegetation*, *Transmission Wires*, and *Buildings*, while classes such as *Fence* and *Poles* have very poor generalization scores. Similar to [24, 37], we believe that the imbalanced distribution of semantic classes significantly affects the model's ability to generalize, as it mainly aligns with dominant classes while struggling to effectively capture the distinct characteristics of less representative but important classes. Qualitative results are illustrated in Fig. 7.

5. Conclusion

We present ECLAIR, a high-fidelity aerial LiDAR dataset and demonstrate how it can be used as a challenging benchmark for 3D semantic segmentation. The high-quality ground-truth labels along with pseudo labels allow benchmarking of existing point cloud semantic segmentation approaches at scale. Additionally, long-tail annotations of point clouds facilitate fine-grained semantic understanding while accommodating the uncertainty of labels. We hope that the ECLAIR dataset will introduce new challenges and stimulate the development of innovative point cloud semantic segmentation approaches that better generalize to real-world scenarios. We also aim to expand ECLAIR's capabilities to include instance segmentation annotations as future work. Furthermore, we would like to expand the ECLAIR dataset to cover larger areas.

Acknowledgments The authors would like to express sincere gratitude to Glenn Colvin for authorizing this research within Sharper Shape. We extend our thanks to Jaro Uljanovs for his intellectual contributions and the valuable code used in this work. We also appreciate the technical insights provided by Khurram Gulzar, Joonas Heikkilä, Rami Piiroinen, and Jussi Sainio. Finally, we recognize Polina Novikova for managing the data curation team, and the team members themselves – Jere Isokääntä, Kia Liljegren-Forss, Svetlana Kuznetcova, and Walter Dewald – for their essential work. Their careful review of the data significantly enhanced the quality of this research.

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1534–1543, 2016. 2, 3

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2021. 3

[3] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9297–9307, 2019. 2, 3

[4] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4413–4421, 2018. 7

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11621–11631, 2020. 2, 3

[6] 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. 2, 3

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In International Conference on 3D Vision (3DV), pages 667–676, 2017. 3

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8748–8757, 2019. 3

[9] Hui-Xian Cheng, Xian-Feng Han, and Guo-Qiang Xiao. Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving, 2022. 3

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3075–3084, 2019. 2, 3, 6, 7

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5828–5839, 2017. 2, 3

[12] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 7

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[14] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D. Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d.net: A new large-scale point cloud classification benchmark, 2017. 3

[15] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset, 2020. 2

[16] 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, 3

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), 2015. 8

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1

[19] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12697–12705, 2019. 3

[20] Xinke Li, Chongshou Li, Zekun Tong, Andrew Lim, Junsong Yuan, Yuwei Wu, Jing Tang, and Raymond Huang. Campus3d: A photogrammetry point cloud benchmark for hierarchical understanding of outdoor scene. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 238–246. Association for Computing Machinery, 2020. 2, 3, 6

[21] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In Advances in Neural Information Processing Systems (NeurIPS), pages 820–830. Curran Associates, Inc., 2018. 3

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2999–3007, 2017. 7, 8

[23] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 922–928, 2015. 3

[24] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 687–693, 2020. 2, 3, 6, 8

[25] Quang-Hieu Pham, Ramanpreet Singh Pahwa Pierre Sevestre, Chun Ho Pang Huijing Zhan, Vijay Chandrasekhar Yuda Chen, Armin Mustafa, and Jie Lin. A*3d dataset: Towards autonomous driving in challenging environments. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2020. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[27] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 17195–17204, 2023. 3

[28] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3:293–298, 2012. 2, 3

[29] Xavier Roynard, Jean-Emmanuel Deschaud, and François Goulette. Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research*, 37(6):545–557, 2018. 3

[30] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), pages 125–141. Springer-Verlag, 2022. 2, 3

[31] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality, 2022. 2

[32] Nathan Silberman and Rob Fergus. Indoor scene segmentation using a structured light sensor. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pages 601–608, 2011. 2

[33] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 567–576, 2015. 2

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2446–2454, 2020. 2, 3

[35] 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 Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 202–203, 2020. 3

[36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1

[37] Nina Varney, Vijayan K Asari, and Quinn Graehling. Dales: A large-scale aerial lidar data set for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 186–187, 2020. 1, 2, 3, 6, 8

[38] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9690–9699, 2021. 7

[39] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: octree-based convolutional neural networks for 3d shape analysis. *ACM Trans. Graph.*, 36(4), 2017. 3

[40] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In Advances in Neural Information Processing Systems (NeurIPS), pages 33330–33342. Curran Associates, Inc., 2022. 3

[41] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training, 2023. 2

[42] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 3

[43] 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. 3

[44] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 12–22, 2023. 2, 3

[45] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 16259–16268, 2021. 3

[46] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J. Guibas. Gimo: Gaze-informed human motion prediction in context, 2022. 2

[47] S. M. Iman Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Rogério E. da Silva, Morteza Rahbar, and Aljoscha Smolic. Dublincity: Annotated lidar point cloud and its applications. In Proceedings of the British Machine Vision Conference (BMVC), 2019. 2, 3