Urban3A: Active Annotation Assisted Semantic Segmentation of Large-scale Urban Scenes

Guoqing Yang
CSSE, Shenzhen University
Shenzhen, China
yangguoqing@szu.edu.cn

Haoyuan Lv
CSSE, Shenzhen University
Shenzhen, China
2310543017@email.szu.edu.cn

Mengke Li
CSSE, Shenzhen University
Shenzhen, China
csmengkeli@gmail.com

Ke Xie
CSSE, Shenzhen University
Shenzhen, China
ke.xie.siat@gmail.com

Hui Huang*
CSSE, Shenzhen University
Shenzhen, China
hhzhiyan@gmail.com

Abstract

Semantic segmentation of point clouds is critically dependent on large volumes of precisely annotated data, a significant barrier especially in extensive urban environments. Prior research has primarily focused on object-level or indoor scenes, contingent upon densely annotated datasets for satisfactory results, rendering these methods less feasible for large urban applications. Addressing this, our study introduces a novel active learning-based selection framework, termed *Urban3A*, which leverages semi-supervised learning to mitigate the challenges of labeling in urban landscapes. Urban3A integrates a unique selection metric that combines basic geometric attributes, uncertainty, and deep learning features, preserving essential contextual information and enhancing the semantic depth of the data. This metric facilitates the identification of informative blocks and crucial points, markedly boosting the segmentation model's performance. Our annotation approach allows users to selectively annotate the most informative points, optimizing the training process. We validate our methodology using the demanding *UrbanBIS* dataset and benchmark against leading semi-supervised and active learning techniques. Results indicate that Urban3A achieves competitive performance with less than 0.1% of the labeled data, representing a significant advancement in efficient and cost-effective semantic segmentation of large-scale urban point clouds.

***Keywords:** Semantic Segmentation, 3D Urban Scenes, Semi-supervised Learning, Active Learning*

1. Introduction

Point cloud is one of the most crucial representations of 3D scenes, providing a wealth of geometric details about the objects within the scene. Semantic segmentation in urban scenes, a practical research area in point cloud analysis, holds paramount significance across various domains [25], including autonomous driving, virtual tourism, automatic modeling, and urban planning, to name a few. Previous endeavors [29] employing traditional algorithms have proven inadequate, failing to capture the promising features of large-scale urban environments. The advent of deep learning-based approaches, started by the introduction of PointNet [24], has provided new insights into this field. A large number of works [25, 29, 24, 9, 5] with the ability to decipher the irregular and complex nature of point clouds have been proposed, enabling the effective capture of semantic information from point clouds.

Despite this commendable progress gained in recent years, previous methods primarily focus on fully-supervised learning, which necessitates a substantial volume of labeled data [25]. However, in large-scale urban scenes, the process of data annotation, particularly for irregular point clouds, is inherently time-consuming [35]. For instance, labeling a single scene in the ScanNet dataset takes 22.3 minutes [6], and considering that this dataset encompasses 1513 scenes, the cumulative effort becomes immeasurable. The difficulty in accurately annotating datasets and the greater complexity of urban environments, which in turn pose challenges for point cloud segmentation.

To alleviate the reliance on labeled data, semi-supervised methods that only require annotation of a small portion of the data have been introduced. These methods can be generally categorized into three groups: 1) contrastive train-

*Corresponding author

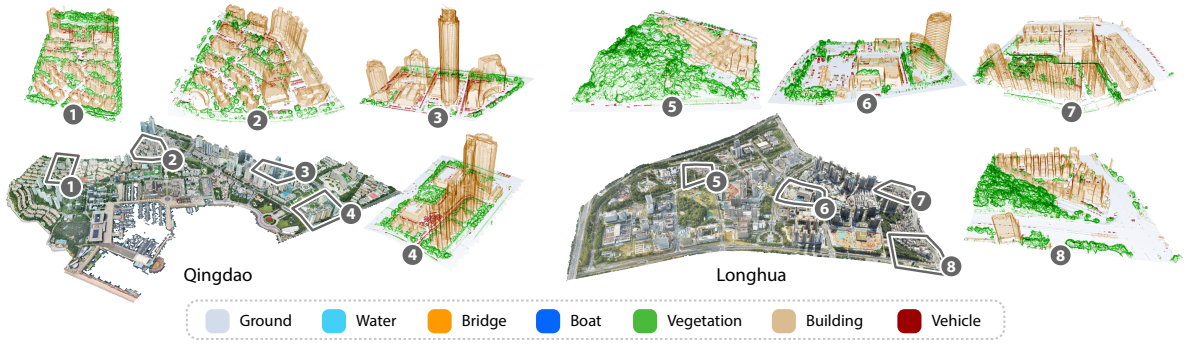

Figure 1. We introduce Urban3A, a semi-supervised approach for segmenting urban point clouds with minimal annotation. Employing an active selection strategy, Urban3A achieves effective segmentation results by annotating just 0.1% of points, making it ideal for large-scale urban environments.

ing [3, 32], 2) pseudo label propagation [17] and 3) co-training [37]. These methods, based on a small portion of annotated samples, adopt various strategies to explore latent information among unlabeled data [4], and design a series of similarity criteria to gradually improve the confidence of the obtained information. While these approaches demonstrate promising results in indoor scenes such as ScanNet [6] and S3DIS [1], they generally overlook the strategic selection of data requiring annotation. Previous research has shown that not all points in a scene contribute equally to model training [33]. Therefore, in large-scale urban scenes, the identification of high contribution points is crucial for reducing annotation burden and effectively capturing the inherently complex nature of urban scenes.

To facilitate the semantic segmentation for urban scenes using a minimal amount of annotated data, this paper proposes a novel active learning strategy, Urban3A, which is seamlessly integrated with semi-supervised paradigm. Building upon prior findings that different points exhibit varying degrees of importance [33], our active learning based annotation method selects samples with high importance to ensure optimal results within a limited annotation budget. To comprehensively utilize contextual information within the scene, instead of solely focusing on individual points, we propose a geometric block selection scheme that is capable of selecting informative blocks for initializing the model. Additionally, we introduce an innovative metric that integrates both deep feature diversity and uncertainty, identifying critical points for distinguishing semantics. This metric, guided by active learning principles, effectively directs the annotation process. A simple semantic segmentation framework for semi-supervision is exploited to validate the feasibility of this metric. We evaluate the proposed method on the UrbanBIS dataset [38], demonstrating that the points selected by Urban3A significantly diminish the demand for annotated data, yielding substantial improvement compared to prior methods and comparable results to supervised learning with a mere 0.1% of the data annotated. For example, we achieve an accuracy of 88.83% and a mIoU of 44.84% in Qingdao with an annotation rate close to 0.1%, surpassing the recently proposed HAPL [35] and CPCM [18] by clear margins, exceeding 10% in terms of both accuracy and mIoU. The main contributions are as follows:

- We propose Urban3A, an active learning assisted semi-supervised segmentation for point clouds in urban scenes. It consists of a block selection scheme utilizing contextual information for identifying informative blocks and a metric combining feature diversity and uncertainty for annotation.
- The proposed Urban3A, a simple yet effective strategy, can be seamlessly integrated into any fundamental model with a substantial reduction of the annotation budget.
- Extensive experiments illustrate that Urban3A, even with limited annotation, exhibits significant improvement and can attain performance levels approaching those of fully-supervised methods.

2. Related Work

2.1. Semantic segmentation of urban point cloud

Initial attempts often relied on manually crafted features and traditional classification methods [23], which are highly depend on the intricate design of features tailored to the diverse categories existed in urban environments. With the development of deep learning, existing works have gradually started to focus on the use of deep learning-based methods [10]. Liu et al. [16] aim for large-scale 3D point clouds and propose a reinforcement learning framework to solve that. Luo et al. [20] introduce MS-RRFSegNet, which initially segments extensive scenes into supervoxels. It subsequently considers both local and global relations to aggregate a more robust feature representation. Khan et al. [11]

incorporate spatial information into their method, employing a symmetric undirected graph model to represent features. Landrieu et al. [14] propose the SGP structure, a novel framework for organizing large-scale point clouds using defined corresponding operators. Zhou et al. [39] convert the urban point clouds into birds-eye-view images and utilize a conventional convolution network as the backbone for subsequent segmentation. Wei et al. [28] utilize point clouds as well as aerial images for the effective segmentation of urban point clouds.

2.2. Semi-supervised semantic segmentation methods

The challenge of semi-supervised segmentation is to assign labels to points with only a limited number of annotations [7, 36]. Xu et al. [34] introduce spatial smoothness constraints, demonstrating that training an incomplete supervision network with only 10% labeled data approximates the performance of fully-supervised training. Hu et al. [8] further show that segmentation performance experiences only a slight decline with as little as 1% of the data annotated. Liu et al. [19] propose the OTOC framework for semi-supervised segmentation, requiring annotation for only one point per object. Liu et al. [18] introduce CPCM, a semi-supervised point cloud semantic segmentation method utilizing the local geometric information. Yang et al. [36] propose a multimodal interlaced transformer that jointly considered 2D and 3D data for weakly segmenting indoor scenes.

Another important branch related to semi-supervised learning is active learning, which focuses on selecting meaningful samples for annotation within input data. Shi et al. [27] explore the significance of feature diversity and uncertainty in active learning, proposing a metric for point clouds segmentation. Meng et al. [21] annotate foreground objects and generate pseudo labels using a probabilistic model for detection tasks. Wu et al. [30] fuse color consistency, prediction result and shape structure to obtain selection metrics and perform experiments on the SemanticKITTI dataset [2]. Shao et al. [26] propose a hybrid strategy called SSDL-AL that utilizes farthest point sampling for feature diversity, then combines the prediction uncertainty with the diversity to choose the informative points. Xu et al. [35] develop a hierarchical point-based active learning strategy, measuring the uncertainty for each point by a hierarchical minimum margin uncertainty module. Despite the success of these works in specific scenarios, there remains a notable gap in the literature concerning the integration of active learning and semi-supervised segmentation tailored specifically for urban scenes. Kolle et al. [13] proposed a fully automated active learning framework for semantic segmentation of geospatial 3D point clouds, with a particular focus on crowdsourced annotation and human–machine collaboration. Their method combines uncertainty-based sampling with feature diversity to reduce labeling effort under a hybrid intelligence setting. In contrast, our work focuses on efficient annotation selection for large-scale urban point clouds without relying on crowdsourcing or human-in-the-loop modeling. We instead emphasize geometry-aware region selection and point-level uncertainty to better exploit spatial structure in dense urban environments.

3. Methodology

3.1. Overview

Existing semi-supervised and active learning methods generally emphasize the significance of individual points, they tend to neglect crucial contextual information necessary for effective comprehension, particularly within complex urban scenes. Moreover, considering the uneven distribution of semantic categories observed in real urban scenes [38], not all categories receive adequate consideration especially when semi-supervised annotations typically account for less than 0.1% [17]. To address these issues, we employ a geometric metric to help the model learn meaningful context information and a specially designed metric for selecting individual point to achieve better segmentation results. The training data are divided into blocks of the same size first, with the geometric metric guiding the selection of the most informative block for annotation. This selected block, along with all other available unlabeled data, are then input into the network. The subsequent training stages encompass two sub-modules: the semantic segmentation module for metric calculation and the active learning module for annotating meaningful samples.

The input point clouds, represented as $P = \{p_1, p_2, p_3, \ldots, p_N\}$, are first fed into the feature extractor to obtain point-wise features, then passed to the segmentation head to generate the segmentation result. N is the total number of the points and $p_i \in \mathbb{R}^6$ is a single point with a 6-dimensional feature including coordinates and color information. We employ the 3D-UNet [5] as the backbone for feature extraction. After the 3D-UNet, the feature is represented as $F \in \mathbb{R}^{N \times D}$, where D is the dimension of the feature. The feature is fed into the segmentation head, yielding the segmentation result $C \in \mathbb{R}^{N \times S}$, in which S is the number of categories. The item c_{ij} in C represents the probability of the point p_i that belongs to the j-th category. The segmentation head is a fully connected dense layer that converts intermediate features into probabilities. To facilitate efficient processing, the raw input points are aggregated into supervoxels V, employing the method proposed in [15]. This operation is under the assumption that points within the same supervoxel share the same label. At the supervoxel level, uncertainty is computed in an online learning manner, namely based

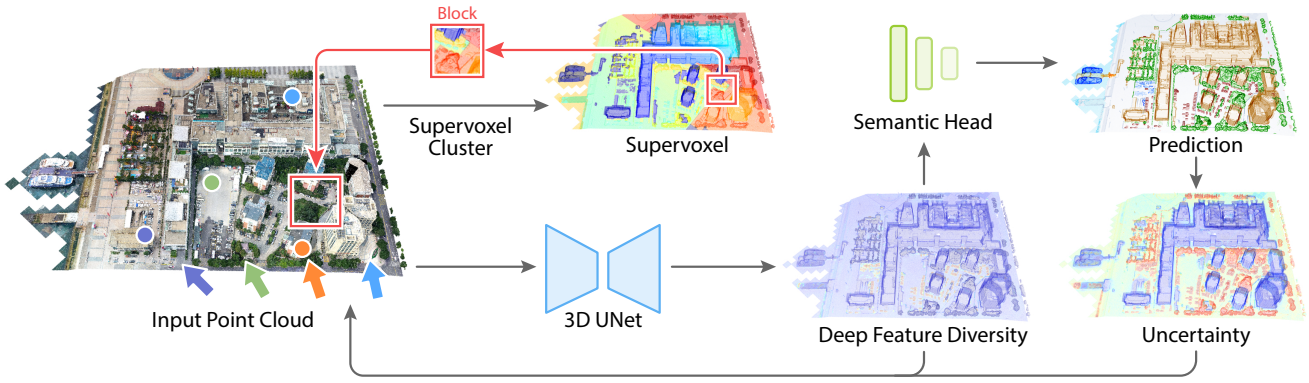

Figure 2. The pipeline overview of Urban3A. Initially, the input point clouds are clustered into supervoxels. Utilizing a geometric metric, we evaluate all blocks and select the most informative one to annotate. The raw point clouds are then fed into the backbone network, extracting point-wise deep features. The segmentation head regresses deep features into probabilities. We leverage the uncertainty derived from the predictions and the deep features to assess the value of individual point for training the model. The most valuable points are annotated for the subsequent training.

on the outcomes of the training process. The diversity of deep features is derived from the deep feature obtained from the segmentation backbone. A predetermined number of points with the highest uncertainty and diversity are selected for annotation after pre-defined epochs. The model is then retrained using these newly annotated data, and this iterative process continues through subsequent training stages. Fig. 2 illustrates the pipeline of Urban3A, which involves multiple training stages that alternate between annotation and forward propagation passing.

3.2. Geometric metric

Random point selection adopted in existing active learning methods [17, 22, 12] ignores contextual information within the whole scene and cannot ensure the richness of semantic selection. Although most methods will select points based on the amount of sample information in subsequent processing, they still overlook the crucial contextual information in urban scene analysis, resulting in inference errors. Therefore, such methods perform well in indoor or sparse outdoor scenes, they are not suitable for dense urban scenes. For a more detailed discussion, please refer to the supplementary material. To effectively initialize the model and ensure the acquisition of contextual information in the scene, we propose a geometry-based metric for the selection of a continuous region to initialize the model.

Conventional active learning schemes typically consider only the spatial location and deep features of individual points, which loses information from local neighborhoods. However, the geometric features of the points encompass semantic clues of point clouds and their neighbors. To harness contextual information within the scene effectively, we propose a basic geometric metric based on curvature to select a targeted small block for initializing the model. The curvature can provide valuable information about how the categories are distributed. We follow the method proposed in [29] to calculate the curvature C_i of point p_i through the eigenvalues derived within the neighborhood:

$$C_i = \frac{\lambda_{i3}}{\lambda_{i1} + \lambda_{i2} + \lambda_{i3}}, \tag{1}$$

where $\lambda_{i1} \geq \lambda_{i2} \geq \lambda_{i3} \geq 0$ are the eigenvalues of point p_i.

Incorporating as many semantic categories as possible within each block enables the model to capture sufficient geometric relationships between semantic categories. The sharp transitions in curvature reflect the location where semantic categories undergo a shift, as explicated in the supplementary material. Therefore, we choose the curvature variation of neighboring blocks as the evaluation indicator. Firstly, the input region is uniformly divided into blocks of equal size, as illustrated in Fig. 3. The average curvature within block B_i is represented by C_i. Subsequently, k neighboring blocks are selected for computation. The geometric score I_i^G for block i is calculated by:

$$I_i^G = \frac{1}{k} \sum_{j=0}^{k-1} ||C_i - C_j||_2 . \tag{2}$$

We then choose the block with the highest geometric score to initialize the model.

3.3. Annotated points selection

3.3.1 Uncertainty

Points within a scene vary in their significance for the learning process. Certain points, which exert a more profound impact on the model than others, need to be discerned. The points with the highest uncertainty in the prediction result serve as indicators of the model's adaptiveness to those samples, and their accurate labeling leads to more substantial

gradient updates. The output of our model represents the probability of each point belonging to different categories. The degree of uncertainty associated with each point can be mathematically described through information entropy, which serves as a metric for quantifying the information density in the prediction. Higher values for information entropy indicate that the point lacks confidence in each category. More details about the deviation can be found in the supplementary material. The uncertainty I_i^U of a point p_i can be calculated by:

$$I_i^U = \sum_{j=0}^{S-1} -c_{ij} \times \log(c_{ij}). \tag{3}$$

The uncertainty for supervoxel V_i can be calculated by averaging the uncertainty of all points in the corresponding supervoxel.

3.3.2 Feature diversity

Deep features provide rich information about the semantic category and distribution of a point, and points with the same label and distribution typically have similar deep features. The diversity of these training features can be a measure of the quality of training data, and is crucial for the performance of learning-based methods. A diverse distribution of data is a key characteristic of high-quality training data. Thus, based on the deep features of annotated data, we further filter unlabeled features that are farthest in distance, serving as an additional metric. Therefore, to achieve better performance with a limited annotation budget, our objective is to select points that maximize the diversity of training data.

$$I_i^D = \frac{1}{k} \sum_{j=0}^{k-1} \left|\left| F^{V_i} - F^{V_j} \right|\right|_2, \tag{4}$$

where the deep feature of the super voxel F^{V_i} is also calculated by averaging the deep feature F of all points in V_i. We set a fixed ratio between uncertainty and deep feature diversity to choose meaningful samples.

After getting the above metric, we combine them with two super parameters α_1 and α_2:

$$I_t = \alpha_1 \times I_t^U + \alpha_2 \times I_t^D. \tag{5}$$

Then we select k supervoxels with higher value to annotate. The points with true labels are used to calculate the cross-entropy loss:

$$Loss = -\frac{1}{|T|} \sum_{i \in T} \hat{c}_i \log c_i, \tag{6}$$

where $\hat{c}_i$ is the true label and c_i is the predicted label, T is the collection of points with true labels.

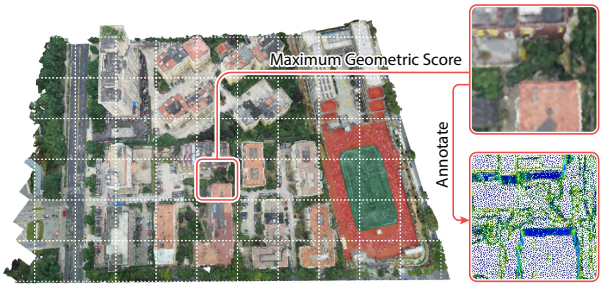

Figure 3. The illustration of geometric block partition. We split the input data into small, predefined blocks, calculate the geometric score for each block, and select the block with the highest score for annotation.

4. Experiments

To demonstrate the efficacy of the proposed Urban3A, we utilize two large scale urban scene datasets: UrbanBIS [38] and SemanticKITTI [2]. Additionally, we provide results on S3DIS [1] to show the performance of Urban3A in indoor scenes, as detailed in Section 3.4 of the supplementary material. Training details and critical parameters are also outlined in Section 3.1 of the supplementary material.

4.1. Experimental results

4.1.1 Dense urban scene

Mean accuracy results from different methods on UrbanBIS are shown in Table 1. Baseline comparisons include the fully-supervised method and random selection. In the cross-test scenario, the model trained on the Qingdao scene serves as the base model, and the test is conducted in another expansive scene, namely, Wuhu and Yingrenshi.

The left of Table 1 demonstrates a notable improvement achieved by Urban3A compared to random selection under conditions of reduced annotation, resulting in a 40% increase in mean accuracy. In particular, the disparity between Urban3A and fully-supervised scenarios is merely 8% even with a minimal 0.1% annotations. Moreover, other methods demonstrate relatively comparable performance in terms of accuracy, nevertheless, our method consistently outperforms all alternatives in this evaluation. The right of Table 1 details the mean performance of Intersection over Union (mIoU) performance across all categories for each method evaluated in UrbanBIS. Notably, our proposed method demonstrates superior performance in the majority of categories, and shows particularly strong results on several small-scale classes, such as water and boat in Qingdao, as well as bridge and vehicle in the campus scene. In contrast, due to ignoring the uneven distribution present in real scenes, other methods struggle to make effective segmentation on these categories when only a small number of

Table 1. Performance comparison w.r.t. semantic segmentation across diverse scenes in UrbanBIS. Accuracy, mIoU, and IoU values for distinct categories under single and cross test settings are listed. Bold text highlights the superior performance achieved in the approaches besides supervised method. The column in which "-" is located represents that these categories are either non-existent or extremely rare on this validation data (over 100× less than other categories). The fully-supervised method employs a model consistent with Urban3A, but excludes the point selection operation, assuming all points are labeled by default. The results of the comparison methods are obtained through averaging across multiple experiments.

Scene	Method	Annotation (%)	Accuracy (%)	mIoU (%)	Ground	Vegetation	Water	Bridge	Vehicle	Boat	Building
Single Test											
Qingdao	Random	1	45.26	11.94	6.08	18.67	8.47	-	1.72	0.82	47.84
	Annotator [31]	0.1	47.82	14.48	25.97	25.41	0.99	-	1.99	0.29	46.74
	CPCM [18]	0.1	75.36	34.01	47.26	67.78	0.57	-	15.97	0	72.48
	HPAL [35]	0.1	53.41	15.74	29.42	32.56	0.03	-	0	0	48.20
	Fully-supervised	100	95.27	66.10	79.25	92.34	73.91	-	78.02	42.22	96.92
	Urban3A (Ours)	≈ 0.1	**88.83**	**44.84**	**62.68**	**81.28**	**13.60**	-	**51.40**	**12.75**	**92.14**
Longhua	Random	1	37.75	10.16	22.09	25.45	-	-	0.19	-	23.41
	Annotator [31]	0.1	61.69	18.49	37.72	33.28	-	-	0.99	-	57.45
	CPCM [18]	0.1	67.66	26.30	55.61	45.13	-	-	0.06	-	56.98
	HPAL [35]	0.1	46.05	12.92	40.19	22.47	-	-	0	-	27.75
	Fully-supervised	100	95.32	49.10	84.77	90.45	-	-	72.81	-	95.69
	Urban3A (Ours)	≈ 0.1	**90.28**	**40.76**	**72.19**	**82.92**	-	-	**38.95**	-	**91.23**
Campus	Random	1	34.91	9.54	17.13	9.69	**1.62**	0.88	1.10	-	36.37
	Annotator [31]	0.1	61.20	18.25	32.50	48.59	1.18	0.37	0.11	-	44.99
	CPCM [18]	0.1	71.04	24.39	50.69	64.99	0.15	0.02	0	-	54.86
	HPAL [35]	0.1	65.45	19.76	43.14	59.81	0	0	0	-	35.39
	Fully-supervised	100	94.40	55.39	81.59	93.25	16.76	42.16	59.14	-	94.82
	Urban3A (Ours)	≈ 0.1	**86.79**	**36.79**	**61.85**	**79.78**	0.02	**7.26**	**19.22**	-	**89.40**
Cross Test											
Wuhu	Random	1	47.06	11.44	6.23	20.62	**3.70**	-	0.92	-	48.57
	Annotator [31]	0.1	54.61	16.51	32.05	30.16	1.38	-	1.71	-	50.23
	CPCM [18]	0.1	40.13	12.35	39.25	31.46	0	-	3.97	-	11.73
	HPAL [35]	0.1	50.79	14.52	28.55	31.19	0.02	-	0	-	41.85
	Fully-supervised	100	91.35	43.95	68.48	78.13	0.34	-	65.41	-	95.61
	Urban3A (Ours)	≈ 0.1	**88.99**	**39.62**	**65.21**	**76.13**	0.16	-	**43.75**	-	**92.09**
Yingrenshi	Random	1	47.57	9.89	5.92	7.31	-	-	1.26	-	54.76
	Annotator [31]	0.1	35.78	8.30	13.36	6.99	-	-	2.10	-	35.55
	CPCM [18]	0.1	21.91	7.08	26.43	10.37	-	-	4.54	-	8.24
	HPAL [35]	0.1	30.94	7.83	15.29	9.94	-	-	0	-	29.63
	Fully-supervised	100	87.50	38.30	56.68	65.13	-	-	57.83	-	88.48
	Urban3A (Ours)	≈ 0.1	**86.12**	**33.94**	**59.43**	**46.90**	-	-	**41.78**	-	**89.47**

points can be selected. As a result, overall accuracy and IoU may either approach or even reduce to zero in these cases. The bottom of Table 1 shows the results of cross test. Although the overall performance of all methods has declined due to changes in the test data, our method still outperforms other methods. Compared to the performance change of fully-supervised method in Wuhu and Yingrenshi scenes, our method exhibits a comparatively lower decline, reflecting its proficiency in generalizing across diverse scenes.

We use the Qingdao scene to show the visual effect of different methods, which is shown in Fig. 4. It can be observed that our method exhibits a relatively small difference from supervised learning methods in overall visual performance. Although there may still exist instances of incorrect segmentation within specific categories, our method shows discriminative capabilities across most categories, particularly for the minority classes with smaller proportions. In contrast, other methods, namely, CPCM and HAPL overlook geometric properties of urban scenes encountering difficulty in discriminating within intricately distributed urban scenarios. They tend to prioritize prevalent categories while neglecting minority classes, impeding their practical applicability.

4.1.2 Sparse urban scene

SemanticKITTI [2], being a representative dataset of the LiDAR scanning dataset, has significant differences with those based on photogrammetry. Specific discussions can be found in the supplementary material. We also conducted validation specifically for urban scene data obtained via LiDAR. Since the SemanticKITTI does not has color information, the validation is different with the common setting. We only use sequences 0 to 9 which have true labels for training and testing. Sequence 10 is the test sequence. The corresponding results are presented in Table 2.

It can be seen from Table 2 that compared to the random method, Urban3A still significantly improves the model performance. However, the overall accuracy is slightly lower, with a certain gap compared to the level of dense urban scene results. This is because after discarding the important point color information, Urban3A experiences a

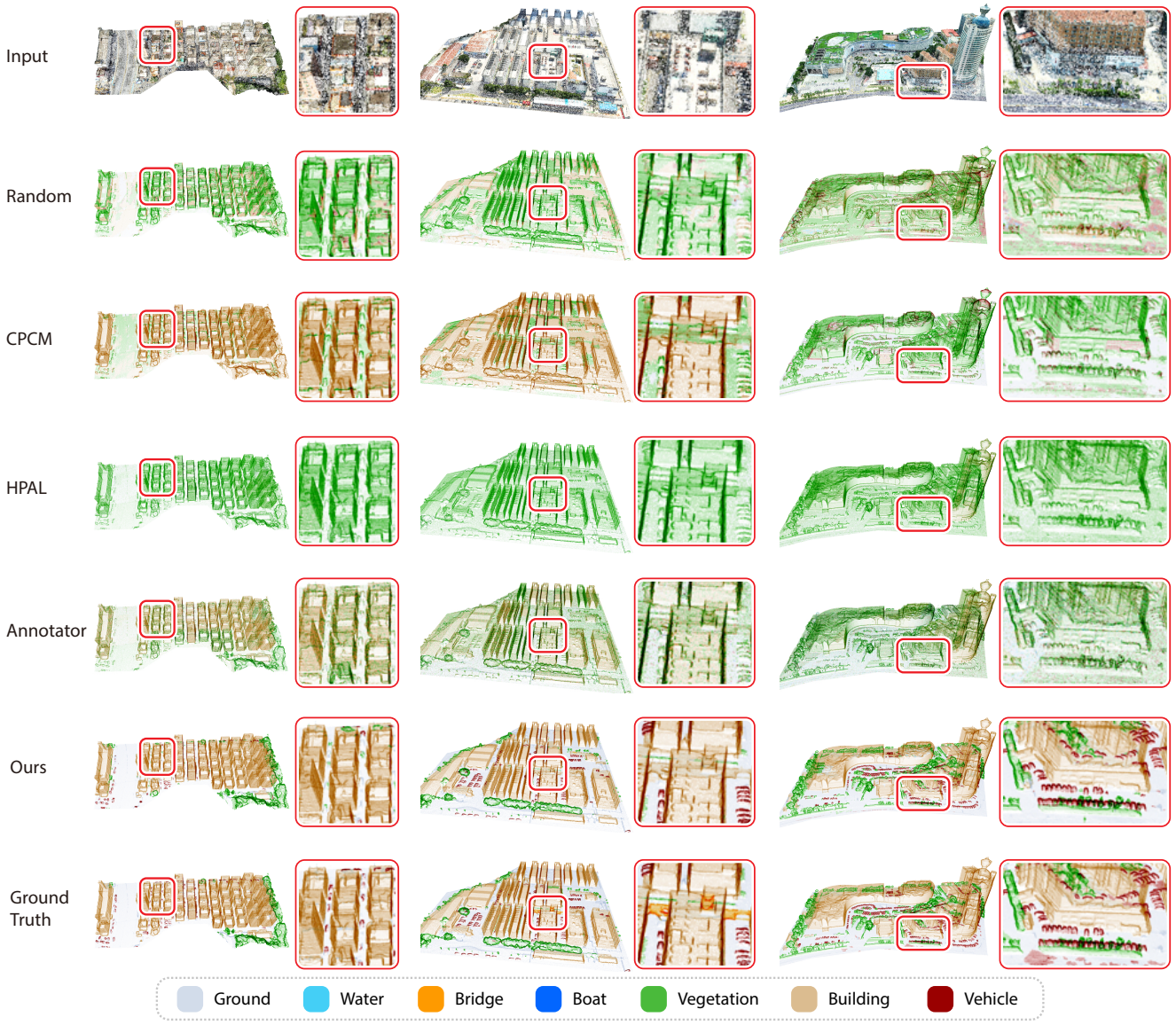

Figure 4. Qualitative results of Urban3A and comparison methods on four test blocks in Qingdao and Longhua scenes. The prediction of Urban3A shows more accurate segmentation results in this dense scene than other semi-supervised methods.

Table 2. Performance comparison across SemanticKITTI.

Method	Annotation (%)	Accuracy (%)	mIoU (%)
Random	0.1	21.66	7.97
Fully-supervised	100	84.08	65.30
Urban3A	0.1	81.68	54.56

decline in feature extraction capability. Meanwhile, due to the unclear distinctions between semantic categories in the dataset, the discriminative power of deep features decreases, which also affects its performance.

4.2. Limitation discussion

We further analyze the performance limitations of our method by comparing it with the supervised learning results, which are presented in Fig. 7. We can see that while our method gets ideal results on most of the points, the low vegetation and building surfaces are still misclassified as ground. The underlying reason is that, despite our method tries to include more diversified samples and rich context information of the scene, it remains constrained by the limited number of annotations. Consequently, it is challenging to cover all kinds of special cases in the real scene. Nevertheless, Urban3A outperforms other methods in these spe-

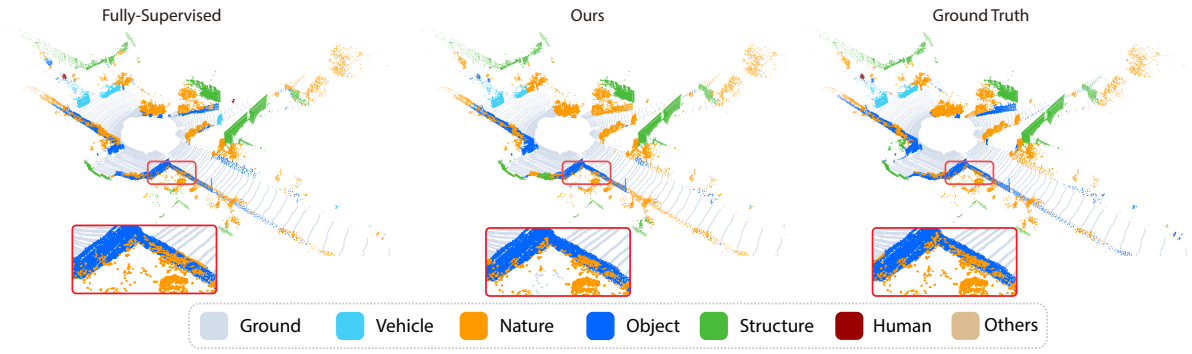

Figure 5. Qualitative results of Urban3A and supervised learning on KITTI Dataset.

cial areas, which is also due to the design of a continuous geometric patch, which optimizes the results not only by relying on the coordinates and color of individual points, but also by obtaining correlation information from the context within the scene. Despite the effectiveness of the proposed strategy, rare categories remain more challenging under extremely limited annotation budgets. This is partly due to their small spatial extent and lower likelihood of being selected by uncertainty- or diversity-based sampling in early AL iterations. Nevertheless, as the selection process proceeds, such regions are gradually incorporated through the iterative refinement of block selection. Addressing this limitation more explicitly remains an interesting direction for future work.

It also can be observed from Fig. 5 that the performance of Urban3A on the SemanticKITTI dataset is worse than the supervised method and cannot handle the situation when the points exhibit similar geometric patterns. This is because the dataset, obtained from sparse laser scans, differs significantly from densely distributed datasets. Not only are the geometric characteristics of semantic categories difficult to preserve, but the large spacing between data distributions also makes capturing local features more challenging. Urban3A and most current semi-supervised or active learning methods attempt to find data distribution patterns based on geometry and distribution. Therefore, they struggle to adapt to such sparse distribution data, resulting in subpar performance. Currently, the effectiveness of Urban3A remains limited to urban scene point clouds obtained through photogrammetry. Overall, Urban3A exhibits superior performance on dense point clouds. However, prediction errors may persist in sparsely distributed scenes that lack color information or in regions with less distinct geometric features. Although Urban3A demonstrates reasonable generalization across different urban scenes, its performance is primarily validated on dense photogrammetry-based datasets. Extending the evaluation to a broader range of urban sensing modalities is an interesting direction for future work.

Table 3. Ablation study of different components. The Qingdao scene from the UrbanBIS dataset is utilized.

Components			Accuracy (%)	mIoU (%)
Geometric patch	Uncertainty	Deep feature		
			45.26	11.94
	✓		85.57	36.40
✓	✓		88.75	41.69
✓	✓	✓	88.83	44.84

Table 4. Performance comparison of Urban3A and random method under different annotation samples. The Qingdao scene is utilized.

Method	Annotation (%)	Accuracy (%)	mIoU (%)
Urban3A (Ours)	≈0.1	88.83	44.84
	≈0.5	86.41	45.63
	≈1	89.12	44.75
Random	1	45.26	11.94
	10	80.71	39.74
Fully-supervised	100	95.27	66.10

4.3. Ablation study

4.3.1 Impact of the metrics

Since the geometric part is only selected for the initialization of the model, we neglect the experiment with only the geometric component. The results are summarized in Table 3, showing the influence of each component on the final outcome. Analysis of the table reveals a consistent improvement in performance with the incorporation of new components. In particular, the model achieves its optimal performance when utilizing all previously proposed indicators. The mIoU has improved by over 5% when we introduced geometric patches into the model. This also reflects the effectiveness of geometric patches in enhancing the model's utilization of contextual information. The discriminative ability has been strengthened across all categories.

4.3.2 Impact of the annotation quantity

We increased the annotation percentage according to our proposed active selection strategy to further evaluate the model performance. The results are presented in Table 4. It can be observed that with the increase in the number of

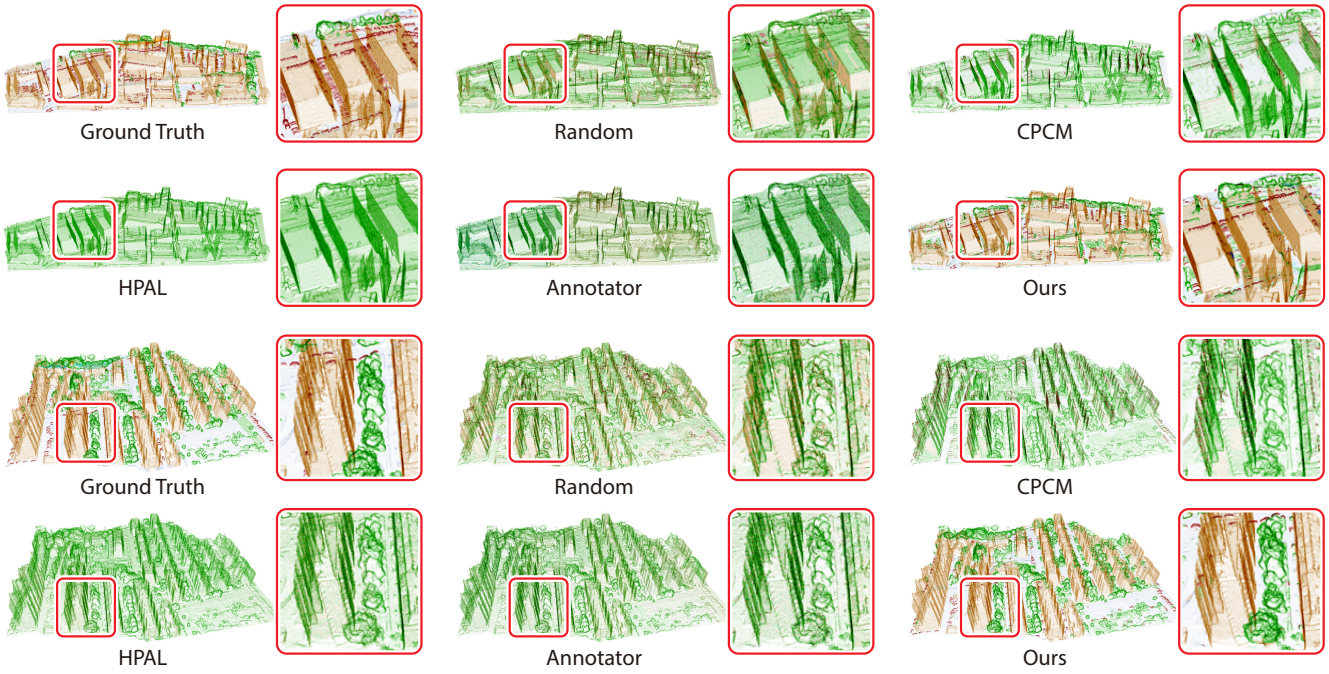

Figure 6. Qualitative results of Urban3A and comparison methods on cross-test of Yingrenshi and Wuhu scenes. Urban3A shows stronger generalization over different urban scenes.

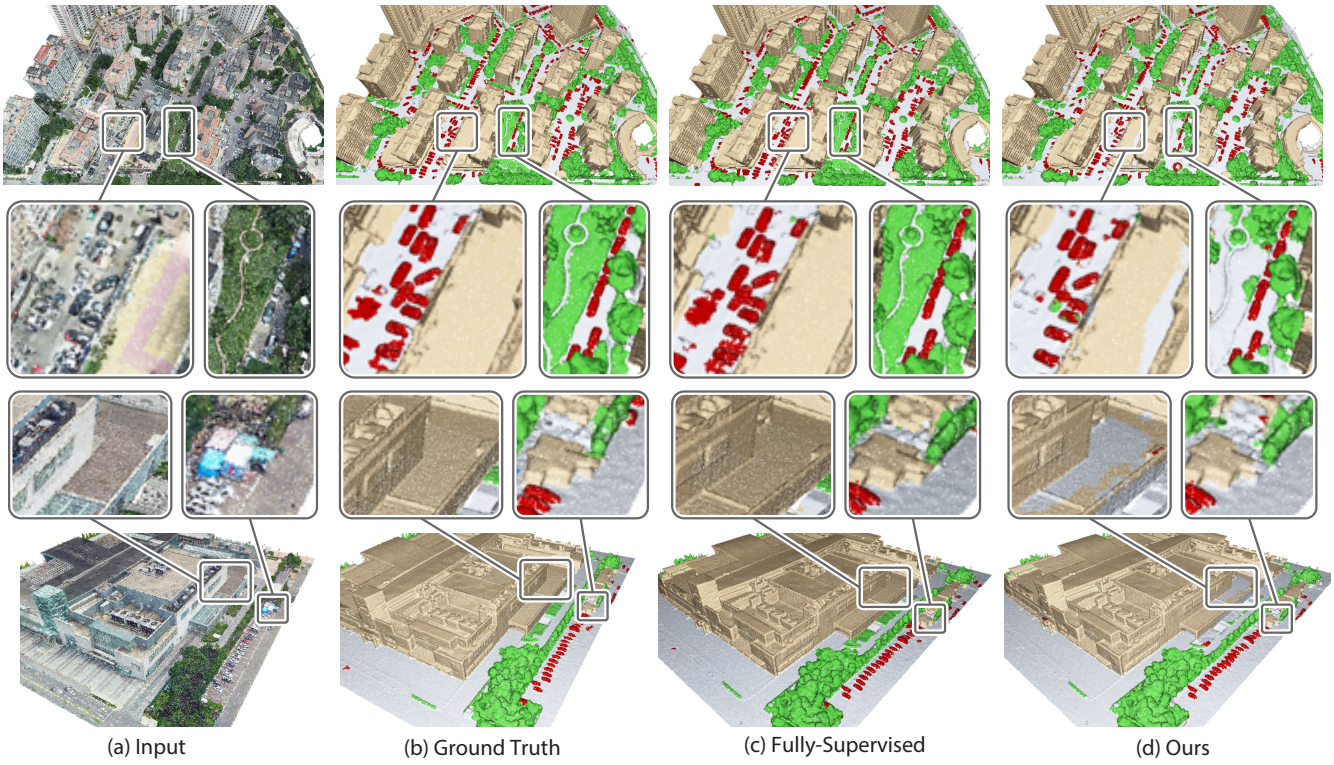

Figure 7. The misclassification of Urban3A in Qingdao is noted as well. Our method performs less effectively in some challenging areas, such as low-rise buildings and flat vegetation region, compared to fully-supervised methods. However, the overall results are nearly on par.

annotations, the model's performance gradually strengthens. This aligns with the general conclusion that more annotated data leads to performance enhancement. However, the performance improvement is marginal when the annotation quantity increases from 0.1% to 1%. This is because, with the adoption of the active selection strategy, high-value points have already been chosen. Subsequently, the increase in annotation quantity leads to redundancy and potential interference, merely further fine-tuning the final results without substantial enhancement. In contrast to random annotation, the advantages of active annotation are more pronounced. Active annotation achieves an accuracy of 8.12% and mIoU of 5.1%, respectively, utilizing only 0.1% of annotations, surpassing random selection using 10% of the data. Random selection overlooks the varying in the importance of different points. Increasing the number of randomly annotated points can introduce additional valuable points, thereby improving the model's performance. Furthermore, it can be observed from Table 4 that once the model is established, fully-supervised learning serves as the upper limit for semi-supervised learning. With an increase in the labeled data, the semi-supervised method consistently converges towards this limit. Based on the aforementioned analysis, users can select an annotation ratio commensurate with the input scale in tasks concerning the comprehension of extensive urban environments to strike a balance between annotation complexities and model efficacy.

5. Conclusion

In this paper, we have presented Urban3A, active learning based semantic segmentation approach for urban point clouds through an active annotation. Our method incorporates a novel geometric metric for selecting diverse object patches, thereby ensuring a well-initialized model. Further, we utilize information entropy-based uncertainty and deep feature diversity to composite the metric for assessing the value of each point and selecting valuable ones to annotate. This strategy is seamlessly integrated into a straightforward training network, and the method is rigorously tested, benchmarked against other semi-supervised and active learning techniques on the dataset UrbanBIS. The experimental results have demonstrated the superior performance of the proposed method in urban scenes, surpassing comparison counterparts by significant margins, highlighting the effectiveness of our active annotation strategy in efficiently utilizing minimal labeled data for robust semantic segmentation in complex urban environments.

Acknowledgement

This work was supported in parts by NSFC (62402322, 62402323), GD Basic and Applied Basic Research Foundation (2023A1515110090, 2023B1515120026), SZ Science and Technology Program (KJZD20240903100022028, RCBS20231211090659101), National Key Laboratory of Radar Signal Processing (JKW202403), and Scientific Development Funds from Shenzhen University.

References

[1] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 1534–1543, 2016. 2, 5

[2] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proc. Int. Conf. on Computer Vision*, pages 9296–9306, 2019. 3, 5, 6

[3] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Proc. Conf. on Neural Information Processing Systems*, pages 22243–22255, 2020. 2

[4] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proc. Int. Conf. on Computer Vision*, pages 9640–9649, 2021. 2

[5] C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 3075–3084, 2019. 1, 3

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 5828–5839, 2017. 1, 2

[7] J. Hou, B. Graham, M. Nießner, and S. Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 15587–15597, 2021. 3

[8] Q. Hu, B. Yang, G. Fang, Y. Guo, A. Leonardis, N. Trigoni, and A. Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *Proc. Euro. Conf. on Computer Vision*, pages 600–619, 2022. 3

[9] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 11108–11117, 2020. 1

[10] T. Jiang, Y. Wang, S. Liu, Q. Zhang, L. Zhao, and J. Sun. Instance recognition of street trees from urban point clouds using a three-stage neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:305–334, 2023. 2

[11] S. A. Khan, Y. Shi, M. Shahzad, and X. X. Zhu. Exploring deep 3d dpatial encodings for large-scale 3d scene understanding. *arXiv preprint arXiv:2011.14358*, 2020. 2

[12] M. Kölle, V. Walter, S. Schmohl, and U. Soergel. Learning on the edge: Benchmarking active learning for the semantic segmentation of als point clouds. *ISPRS Ann. Photogramm. Remote Sensing and Spatial Information Sciences*, X-1/W1-2023:945–952, 2023. 4

[13] M. Kölle, V. Walter, and U. Sörgel. Building a fully-automatized active learning framework for the semantic segmentation of geospatial 3d point clouds. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(2):131–161, 2024. 3

[14] L. Landrieu and M. Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 4558–4567, 2018. 3

[15] Y. Lin, C. Wang, D. Zhai, W. Li, and J. Li. Toward better boundary preserved supervoxel segmentation for 3d point clouds. *ISPRS J. Photogrammetry and Remote Sensing*, 143:39–47, 2018. 3

[16] F. Liu, S. Li, L. Zhang, C. Zhou, R. Ye, Y. Wang, and J. Lu. 3dcnn-dqn-rnn: A deep reinforcement learning framework for semantic parsing of large-scale 3d point clouds. In *Proc. Int. Conf. on Computer Vision*, pages 5678–5687, 2017. 2

[17] G. Liu, O. van Kaick, H. Huang, and R. Hu. Active self-training for weakly supervised 3d scene semantic segmentation. *J. Computational Visual Media*, 2023. 2, 3, 4

[18] L. Liu, Z. Zhuang, S. Huang, X. Xiao, T. Xiang, C. Chen, J. Wang, and M. Tan. Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 18413–18422, 2023. 2, 3, 6

[19] Z. Liu, X. Qi, and C.-W. Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 1726–1736, 2021. 3

[20] H. Luo, C. Chen, L. Fang, K. Khoshelham, and G. Shen. Ms-rrfsegnet: Multiscale regional relation feature segmentation network for semantic segmentation of urban scene point clouds. *IEEE Trans. on Geoscience and Remote Sensing*, 58(12):8301–8315, 2020. 2

[21] Q. Meng, W. Wang, T. Zhou, J. Shen, Y. Jia, and L. Van Gool. Towards a weakly supervised framework for 3d point cloud object detection and annotation. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 44(8):4454–4468, 2021. 3

[22] Z. Pan, N. Zhang, W. Gao, S. Liu, and G. Li. Less is more: Label recommendation for weakly supervised point cloud semantic segmentation. In *Proc. AAAI Conf. on Artificial Intelligence*, pages 4397–4405, 2024. 4

[23] J.-J. Ponciano, M. Roetner, A. Reiterer, and F. Boochs. Object semantic segmentation in point clouds—comparison of a deep learning and a knowledge-based method. *ISPRS Int. J. Geo-Information*, 10(4):256, 2021. 2

[24] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 652–660, 2017. 1

[25] D. Robert, B. Vallet, and L. Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 5575–5584, 2022. 1

[26] F. Shao, Y. Luo, P. Liu, J. Chen, Y. Yang, Y. Lu, and J. Xiao. Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning. In *Proc. ACM Int. Conf. on Multimedia*, pages 2575—-2585, 2022. 3

[27] X. Shi, X. Xu, K. Chen, L. Cai, C. S. Foo, and K. Jia. Label-efficient point cloud semantic segmentation: An active learning approach. *arXiv preprint arXiv:2101.06931*, 2021. 3

[28] W. Wei, M. R. Oswald, F. K. Nejadasl, and T. Gevers. Apnet: Urban-level scene segmentation of aerial images and point clouds. In *Proc. Int. Conf. on Computer Vision*, pages 1755–1764, 2023. 3

[29] M. Weinmann, B. Jutzi, and C. Mallet. Semantic 3d scene interpretation: A framework combining optimal neighborhood size selection with relevant features. *ISPRS Ann. Photogramm. Remote Sensing and Spatial Information Sciences*, 2(3):181–188, 2014. 1, 4

[30] T.-H. Wu, Y.-C. Liu, Y.-K. Huang, H.-Y. Lee, H.-T. Su, P.-C. Huang, and W. H. Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proc. Int. Conf. on Computer Vision*, pages 15510–15519, 2021. 3

[31] B. Xie, S. Li, Q. Guo, C. Liu, and X. Cheng. Annotator: A generic active learning baseline for lidar semantic segmentation. In *Proc. Conf. on Neural Information Processing Systems*, pages 48444–48458, 2023. 6

[32] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Proc. Euro. Conf. on Computer Vision*, pages 574–591, 2020. 2

[33] M. Xu, Z. Zhou, J. Zhang, and Y. Qiao. Investigate indistinguishable points in semantic segmentation of 3d point cloud. In *Proc. AAAI Conf. on Artificial Intelligence*, pages 3047–3055, 2021. 2

[34] X. Xu and G. H. Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 13706–13715, 2020. 3

[35] Z. Xu, B. Yuan, S. Zhao, Q. Zhang, and X. Gao. Hierarchical point-based active learning for semi-supervised point cloud semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 18098–18108, 2023. 1, 2, 3, 6

[36] C.-K. Yang, M.-H. Chen, Y.-Y. Chuang, and Y.-Y. Lin. 2d-3d interlaced transformer for point cloud segmentation with scene-level supervision. In *Proc. Int. Conf. on Computer Vision*, pages 977–987, 2023. 3

[37] C.-K. Yang, Y.-Y. Chuang, and Y.-Y. Lin. Unsupervised point cloud object co-segmentation by co-contrastive learning and mutual attention sampling. In *Proc. Int. Conf. on Computer Vision*, pages 7335–7344, 2021. 2

[38] G. Yang, F. Xue, Q. Zhang, K. Xie, C.-W. Fu, and H. Huang. Urbanbis: A large-scale benchmark for fine-grained urban building instance segmentation. In *Proc. SIGGRAPH*, pages 16:1–16:11, 2023. 2, 3, 5

[39] Z. Zou and Y. Li. Efficient urban-scale point clouds segmentation with bev projection. *arXiv preprint arXiv:2109.09074*, 2021. 3

Urban3A: Active Annotation Assisted Semantic Segmentation of Large-scale Urban Scenes (Supplementary Material)

Guoqing Yang
CSSE, Shenzhen University
Shenzhen, China
yangguoqing@szu.edu.cn

Haoyuan Lv
CSSE, Shenzhen University
Shenzhen, China
2310543017@email.szu.edu.cn

Mengke Li
CSSE, Shenzhen University
Shenzhen, China
csmengkeli@gmail.com

Ke Xie
CSSE, Shenzhen University
Shenzhen, China
ke.xie.siat@gmail.com

Hui Huang*
CSSE, Shenzhen University
Shenzhen, China
hhzhiyan@gmail.com

1. Discussion of Metrics

1.1. Uncertainty

Cross-entropy is typically utilized to compute the loss for segmentation tasks, we can rewrite the cross-entropy function as the equivalent Lagrangian function:

$$L(c_{ij}, \lambda) = \sum_{j=0}^{S-1} -c_{ij} \times \log(\hat{c}_{ij}) - \lambda \times (\sum_{j=0}^{S-1} \hat{c}_{ij} - 1), \quad (1)$$

given that the sum of the prediction probabilities must equal 1.

By taking partial derivatives of each variable in Eq (1), we can obtain the conditions under which this equation achieves extremum as:

$$\log(\hat{c}_{i1}) + 1 = \log(\hat{c}_{i2}) + 1 = ... = \log(\hat{c}_{iS-1}) + 1. \quad (2)$$

The analysis elucidates that when information entropy attains an extremum, it necessitates all predicted outcomes to be normalized to $\frac{1}{S}$. When the cross-entropy reaches an extremum, it can accelerate the model training process, at the same time, the information entropy reaches its maximum. Therefore using information entropy to measure the uncertainty for the selection metric is reasonable.

1.2. Feature diversity

To verify the effectiveness of the deep feature metric in Urban3A, we conducted statistical analyses on the deep features predicted by the network for different semantic categories. Using the given ground truth labels, we clustered the deep features according to semantic categories. Subsequently, the average deep features for each category is obtained by calculating the average Euclidean distance between semantic categories. For intra-category, we adopted a pairwise traversal method to compute the average value of each pair of features within the category. The results are presented in Table 1.

Table 1. Average feature distance between different semantic categories. We selected a test area in Qingdao for calculation. Due to the absence of the 'Bridge' and 'Boat' categories in the test areas, the distances related to these two categories are replaced with '-'. The distance within the same semantic category is calculated by the average distance between all features within that category.

	Terrain	Vegetation	Water	Bridge	Vehicle	Boat	Building
Terrain	**3.01**	5.20	3.69	-	4.73	-	6.54
Vegetation	5.20	**2.46**	7.29	-	5.01	-	5.30
Water	**3.69**	7.28	6.69	-	7.55	-	8.03
Bridge	-	-	-	-	-	-	-
Vehicle	4.73	5.01	7.55	-	**4.72**	-	5.76
Boat	-	-	-	-	-	-	-
Building	6.54	5.30	8.03	-	5.76	-	**2.21**

It can be observed that a pronounced correlation exists between deep features and semantic categories, as significant distances separate the average deep features across various categories. Furthermore, certain semantic categories manifest proximate deep feature distances, for instance, 'Water' and 'Terrain' within urban scenes, which could potentially result in misclassification errors. According to publicly available results, the UrbanBIS dataset has undertaken denoise procedure, with all retained points annotated with semantic categories. Therefore, our method currently does not face the issue of annotating noisy points. However, noise can indeed affect training methods with limited annotations and is unavoidable. We will further analyze the noise patterns in urban scenes in the future to effectively

*Corresponding author

filter out noise during the annotation process, thereby reducing its impact on the training.

1.3. Curvatures

This section analyzes why curvature is chosen as a geometric metric. Firstly, the curvature within the scene is clustered according to semantic categories. Table 2 shows the mean and variance of curvature values for various semantic categories. The correlation between curvature and semantic categories can be observed from the mean values. The variance demonstrates that the differences within classes are relatively small. Additionally, we visualize the curvature in Figure 1, which demonstrates that curvature possesses discriminative characteristics with respect to semantic types, thereby validating the rationale of our proposed approach.

Table 2. Mean and variance of curvature in the selected scene. Both the mean and variance values are calculated based on ground truth labels.

	Mean (x1000)	Variance
Ground	25.38	0.002
Vegetation	90.48	0.005
Water	20.47	0.002
Bridge	-	-
Vehicle	84.60	0.004
Boat	-	-
Building	56.79	0.005

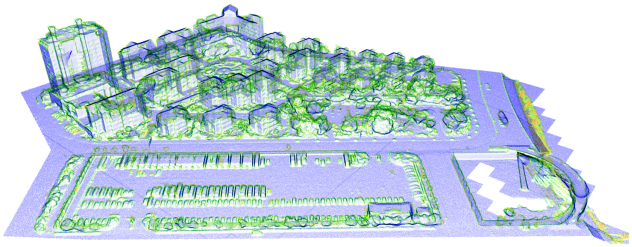

Figure 1. Illustration of curvature magnitudes within the scene. The abrupt changes in color reflect significant changes in curvature magnitudes, with the majority of changes occurring at semantic boundary.

Based on the above analysis, the geometric metrics adopted in this work, including uncertainty, deep feature diversity, and curvature, are designed to serve as lightweight and robust indicators for guiding the annotation process rather than as precise geometric descriptors. In particular, the curvature metric is computed from local point statistics and does not rely on explicit surface reconstruction or strong planarity assumptions. It is only used at the initialization stage to characterize geometric diversity at a coarse level, rather than to model fine-grained surface details. For complex urban scenes containing irregular or non-surface structures, such geometric variations are naturally reflected in the curvature distribution and do not lead to systematic information loss. Moreover, the influence of the geometric metric is limited to early-stage block selection, while subsequent annotation and learning are driven by deep feature representations and uncertainty estimation. This design achieves a balance between computational efficiency, robustness, and practical applicability in large-scale urban scenarios, which aligns with the overall objective of minimizing annotation cost without introducing heavy model dependencies.

2. Discussion of Supervoxel

2.1. Supervoxel description

Supervoxel is a technique for clustering point clouds, aiming to represent discrete point clouds in a more compact way. It aggregates nearby similar points into a region based on their geometric and color features, changing the basic processing unit of point cloud objects from discrete points to clustered regions. Since it clusters points based on similarity, and the number of clusters is much larger than the semantic categories present in the scene, it is generally assumed that points within the same region have similar properties, especially consistent semantic labels. Therefore, supervoxels can not only be used for point clouds feature analysis but also serve as an effective method to reduce the scale of point clouds processing. Many point cloud processing methods utilize supervoxel clustering as a preprocessing step, especially for semi-supervised approaches. Based on the premise assumption that points within the same region have the same label, these methods can conveniently propagate labels within the same supervoxel through diffusion, significantly reducing the difficulty of handling limited annotations.

We adopted the point clouds supervoxel cluster method proposed by Lin et al. [6]. This method transforms the supervoxel segmentation problem into a subset selection and employs a heuristic algorithm based on local information to solve such problems. The main parameter is the cluster resolution. Depending on the resolution setting and the original input point cloud quantity, the final number of supervoxels obtained varies. This method achieves higher precision in segmenting boundary point clouds compared to previous methods. In order to further improve the precision of subsequent processing in large-scale urban scenes where boundary situations are more complex, we chose this method for preprocessing the point cloud.

Although multi-scale or hierarchical supervoxel strategies could further adapt to objects of different sizes, they also introduce additional computational cost and system complexity. In this work, we adopt a single-scale but geometry-aware supervoxel configuration, which achieves a good balance between efficiency and representational adequacy. Since the supervoxels are only used for guiding

annotation selection rather than final prediction, the impact of scale mismatch is largely mitigated by the subsequent point-level learning process.

2.2. Cluster performance

Lin et al. [6] employ a heuristic approach for clustering, inevitably introducing cluster errors. Most semantic segmentation methods that use supervoxels for preprocessing assume that this process does not introduce any errors, thus overlooking this aspect. In order to further evaluate the impact of different operations on scene segmentation, in this section, we discuss the influence of supervoxels on clustering urban point clouds.

As supervoxel operations are performed independently on segmented datasets, we statistically analyze the segmentation performance of each test scene. The semantic category with the highest frequency within the same supervoxel is considered as the main semantic category of that supervoxel. We calculate the number of points in the input point cloud that differ from this main semantic category as the classification error. We record the maximum and minimum classification errors in each scene and aggregate all misclassified points in the entire scene. The overall results are summarized in Table 3.

Table 3. Error rate of supervoxel cluster in different scenes. Minimum and maximum mean minimum and maximum error rate among all test areas in this scene separately. Mean error is calculated by counting all misclassified points in the scene.

Scene	Minimum (%)	Maximum (%)	Mean Error (%)
Qingdao	1.41	4.18	2.08
Longhua	**0.41**	2.79	1.46
Wuhu	**1.45**	**4.01**	2.27
Yuehai	1.02	**1.60**	1.90
Lihu	0.85	1.92	1.41
Total	-	-	1.74

It can be seen from the data in the table that on average, there are classification errors ranging from 1% to 4% across different scenes. These errors are difficult to correct under the basic assumptions and thus accumulate into the final error. However, considering the possibility of mispredictions specifically targeting the main categories, this classification error is not directly added to the final result in a linear manner.

We selected certain scenes from Qingdao for visualization, as shown in Fig. 2. Although this method attempts to minimize processing errors at the edges, it still cannot achieve perfect segmentation results in the boundary regions. The overall average error is relatively small, thus the impact on subsequent segmentation performance is also minor. Therefore, we choose to overlook the classification errors introduced by supervoxels.

Figure 2. Illustration of misclassification in supervoxel clustering process. Misclassified positions are highlighted in red, indicating that these errors mostly occur near semantic boundaries and are relatively minor. Reasons for misclassification at semantic boundaries include the sparsity of point cloud data, reconstruction errors in boundaries, and subjective errors during annotation.

3. Experimental supplement

3.1. Experiment settings

Ten Quadro P6000 GPUs are utilized for model training, employing a batch size of 4 during the training process. The Adam optimizer [5] is applied, with an initial learning rate set to 0.001. The training epoch is set to 400. To ensure a fair comparison, we follow the original experimental settings of each method, making modifications solely to the hyperparameters to align with our dataset. For the 3D sparse convolution, we set the voxel size as $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} m^3$. During the training process, to balance the GPU memory limit and data block size, we set the maximum number of points as 500,000 and randomly adjust the input size by cropping a block if its size exceeds the maximum, similar to [4]. Annotations are conducted every 10 epochs, with the annotation count dynamically adjusted based on the total training instances across diverse scenes. Mean accuracy is employed as a metric to assess the performance of semantic segmentation. Mean Intersection-over-Union (mIoU) is utilized to evaluate the segmentation performance across various categories. Because our method uses an iterative training method, the annotated quantity remains consistent throughout each annotation epoch. For a fair comparison, the total annotation is set to be approximately the same across the different methods. While the number of labeled points may vary across different scenes, the percentages of annotation in different scenes maintain on the same level. We follow the official dataset partitioning in UrbanBIS [12].

The validation of all methods is conducted through the single-scene test, encompassing two large scenes, Qingdao and Longhua, along with two small scenes, Yuehai and Lihu. These scenes Collectively constitute the campus scene. We select supervised learning method which annotate every point in the scene and random annotation method which adopts the same setting with Urban3A except for the annotation strategy as the baseline methods. Be-

sides, we also choose the state-of-the-art semi-supervised method like CPCM [7] and point active learning methods like HPAL [11] and Annotator [10]. Annotator adopts voxelization as the pre-processing step, while HPAL consider single point learning to avoid region error in their original code. Both of the active learning methods neglect the supervoxel cluster in their pipeline. We carry out the comparison tests following the instructions of their original codes.

3.2. Neighbor quantity selection

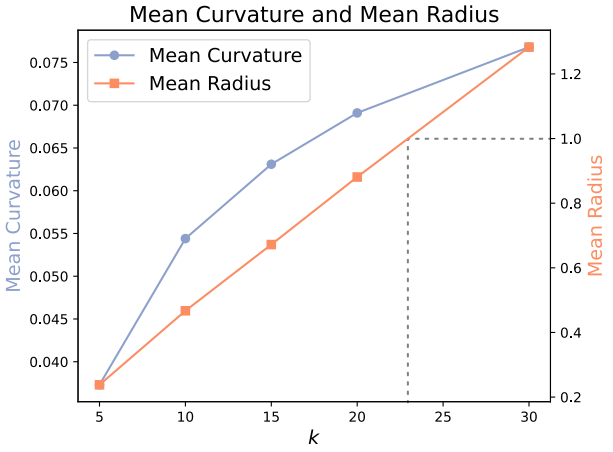

Figure 3. The illustration of the average bounding box and average curvature change with k. We selected the Qingdao scene for analysis and set the base scale to 1 m.

The selection of neighborhood quantity needs to fully consider the characteristics of the dataset. A small value can result in significant calculation bias in curvature, while an excessive quantity can lead to overly large neighborhood sizes, including the computation of many irrelevant points. For the UrbanBIS dataset, we statistically analyze the average curvature values and average bounding radius under different neighborhood quantities in Qingdao test areas, as shown in Fig. 3. With the increase in neighborhood quantity, both the calculated average curvature and average bounding radius continue to increase. Here, we consider that semantic objects within typical urban scenes may have scales smaller than 2 m, such as vehicles, trees, etc. Therefore, we choose a 1 m average bounding radius as a standard. It is observed that at this point, the neighborhood quantity should be less than 25. For computational convenience, we choose 20 as the result.

3.3. Comparison among different scenes

Urban3A is developed for urban scene point clouds established based on photogrammetry, such as UrbanBIS [12], SensatUrban [3], etc. These datasets are notably distinct from indoor scene data or urban data obtained through laser scanning. We compare the characteristics of different datasets in this section. This comparison aims to clarify the differences between various scenes and the suitability of our method for dense urban point clouds scenes similar to UrbanBIS. We selected representative outdoor semantic parsing datasets SemanticKITTI [2] and indoor point cloud datasets S3DIS [1], along with our experimental dataset UrbanBIS [12], for analysis. SemanticKITTI, as an outdoor scene dataset, primarily focuses on driving roads and is used for environmental parsing in autonomous driving tasks, while S3DIS is obtained through indoor environment scans. We conducted an analysis of the geometric statistics by calculating the average curvature within the scenes. The relevant curvature statistical values are shown in Table 4

The mean curvature reflects the richness of the geometry in the scene, while the variance reflects the differences among semantic categories. It can be inferred from Table 4 that UrbanBIS and S3DIS exhibit a more obvious structure than SemanticKITTI, and urban datasets have larger intra-class differences. We can conclude that urban scene datasets typically exhibit dense point cloud representations, rich semantic categories, and large scales, all of which need to be considered when designing relevant processing algorithms. In the design process of Urban3A, we utilized sparse convolutional networks as the backbone network to handle such dense large-scale data. Additionally, we incorporated diversity in semantic categories into the point selection strategy for deep feature extraction. The above reasons make Urban3A currently only suitable for the application of dense urban point clouds, and may perform poorly in other datasets.

Table 4. Comparison of statistics in different dataset. As for UrbanBIS, we choose Qingdao as the representative scene.

Dataset	Curvature Variance ($\times$1000)	Mean Curvature
SemanticKITTI	5.04	0.06
S3DIS	0.79	8.07
UrbanBIS	5.33	69.09

3.4. S3DIS results

S3DIS contains 3D scans from Matterport scanners in 6 areas including 271 rooms. Each point in the scan is annotated with one of the semantic labels from 13 categories [1]. We employed a common training paradigm, using Area 5 as the test area and Areas 1 to 6 excluding Area 5 as the training areas to train the fully-supervised, random annotation methods, and our Urban3A. Since other methods have already been tested in the indoor scenes, we directly adopted their results from the original paper here. The relevant statistical data is listed in Table 5.

As mentioned above, Urban3A is a method tailored for the urban point clouds, showing promising performance under scenes which exhibits large geometric variances. As for

Table 5. Performance compassion across S3DIS. Since most methods only compare based on the mIoU metric, we also solely used the IOU metric to evaluate each method to ensure the completeness of the comparison.

Method	Annotation (%)	mIoU (%)
Random	1	42.1
CPCM [7]	0.1	66.3
HPAL [11]	0.1	59.3
SSDR-AL [9]	11.7	58.3
Urban3A	0.1	44.1

indoor scenes like S3DIS, the difference in geometric characteristics among semantic categories is relatively small, thus a performance gap exists between Urban3A and other semi-supervised or active learning methods.

3.5. Discussion on efficiency and performance trade-off

A key objective of Urban3A is to reduce annotation cost while maintaining competitive segmentation performance. Unlike fully supervised approaches that require dense annotations for all training samples, Urban3A is designed to operate under extremely limited annotation budgets and focuses on maximizing annotation efficiency through active selection. To clarify this design goal, we provide an explicit discussion on the performance gap between Urban3A and fully supervised training, and explain how this gap should be interpreted in the context of efficiency-oriented urban scene understanding. Fully supervised models trained with complete annotations naturally achieve higher upper-bound performance. In contrast, Urban3A operates with only a small fraction of labeled data (e.g., 0.1%), and therefore prioritizes annotation efficiency over absolute accuracy. Table 6 conceptually summarizes the comparison between Urban3A and fully supervised training.

Table 6. Conceptual comparison between Urban3A and fully supervised training. The annotation cost for the fully supervised setting is estimated based on the average labeling speed reported in the original work, together with the total number of annotated points in the Longhua dataset. The training time of Urban3A is obtained from our implementation running on a server with 8 NVIDIA RTX 4090 GPUs. These values are reported to provide an indicative comparison of annotation and training efficiency rather than an exact runtime equivalence.

Method	Annotation Ratio	mIoU	Annotation Cost (h)
Fully Supervised	100%	49.10	360
Urban3A (Ours)	0.1%	40.76	12.5

Although a performance gap exists, Urban3A achieves a favorable trade-off by maintaining competitive accuracy with orders-of-magnitude fewer annotations. This comparison highlights the efficiency-oriented nature of the proposed framework. The performance efficiency trade-off of Urban3A is a deliberate design choice rather than a limitation. Specifically: Annotation efficiency: Urban3A significantly reduces labeling cost by selecting informative samples through uncertainty and diversity criteria, instead of relying on exhaustive annotation. Computational efficiency: The framework avoids dependence on heavy pretraining or complex geometric learning modules, enabling efficient deployment on large-scale urban scenes. Scalability: The method is designed to scale to city-level point clouds, where full supervision is often impractical. As a result, Urban3A should be evaluated primarily in terms of performance per annotation budget, rather than absolute performance under full supervision. We acknowledge that a performance gap remains between Urban3A and fully supervised models, particularly for rare categories. This is an expected outcome given the extremely limited annotation budget. Bridging this gap further would likely require additional supervision, stronger priors, or domain-specific pretraining, which is beyond the scope of this work. Nevertheless, the experimental results demonstrate that Urban3A achieves a strong balance between efficiency and accuracy, making it a practical solution for large-scale urban annotation scenarios.

3.6. Additional results

We show the rest of the representative results in Fig. 4, Fig. 5 and Fig. 6.

4. Miscellaneous

4.1. Ethic concerns

Our work is based on a publicly available dataset, UrbanBIS. As confirmed by querying the dataset's official website[1], UrbanBIS does not involve any ethical issues related to point clouds. Therefore, our work does not pose any ethical concerns either.

4.2. Discussion of efficiency

The aim of this work is to provide a possibility for reducing annotation time. By employing Urban3A, users can achieve satisfactory semantic perception of urban scenes without annotating the entire point cloud. Current active learning methods all require iterative training, thus the specific time required for this process is inevitable and highly dependent on hardware. Therefore, such methods often demonstrate the effectiveness of their approach in reducing annotation time through the quantity of annotations [8, 10, 11]. In the experiments, we compared Urban3A with other active learning methods under the same number of annotations and achieved better performance in urban scenes. From the experimental data, it can also be observed that the accuracy is positively correlated with the number of annotations, which directly affects the time required.

[1] https://vcc.tech/UrbanBIS

4.3. Imperfect annotation

Imprecise annotation. Although the pipeline of Urban3A involves manual labeling of selected points, we validate utilizing pre-annotated datasets in our experiments. Therefore, there are no label errors in our experiments. Regarding biases in the collected data, even if the annotation of the selected dataset is not precious, we cannot correct labeling errors without accessing the original data and must assume both the annotated labels and input data are ideal. However, to mitigate overfitting, we introduce operations during data loading such as random jittering, rotation, and color normalization for point cloud positions and colors. Thus, we believe minor errors in input data will not significantly affect the final results. Nonetheless, this imperious label issue remains a valuable research direction. In the future, we will delve deeper into the critical factor of labeling errors, particularly exploring biases introduced by human annotation tendencies.

Redundant Annotation. Excessive redundant annotations would waste the labeling resources, our method endeavors to minimize such a situation. In cases of redundancy, the majority resemble the already annotated samples, and since the model performs well on these annotated samples, such redundant points typically do not exhibit high uncertainty. Additionally, Urban3A tends to prioritize annotating features that differ significantly from current indicators, thus mitigating redundancy in deep features as much as possible.

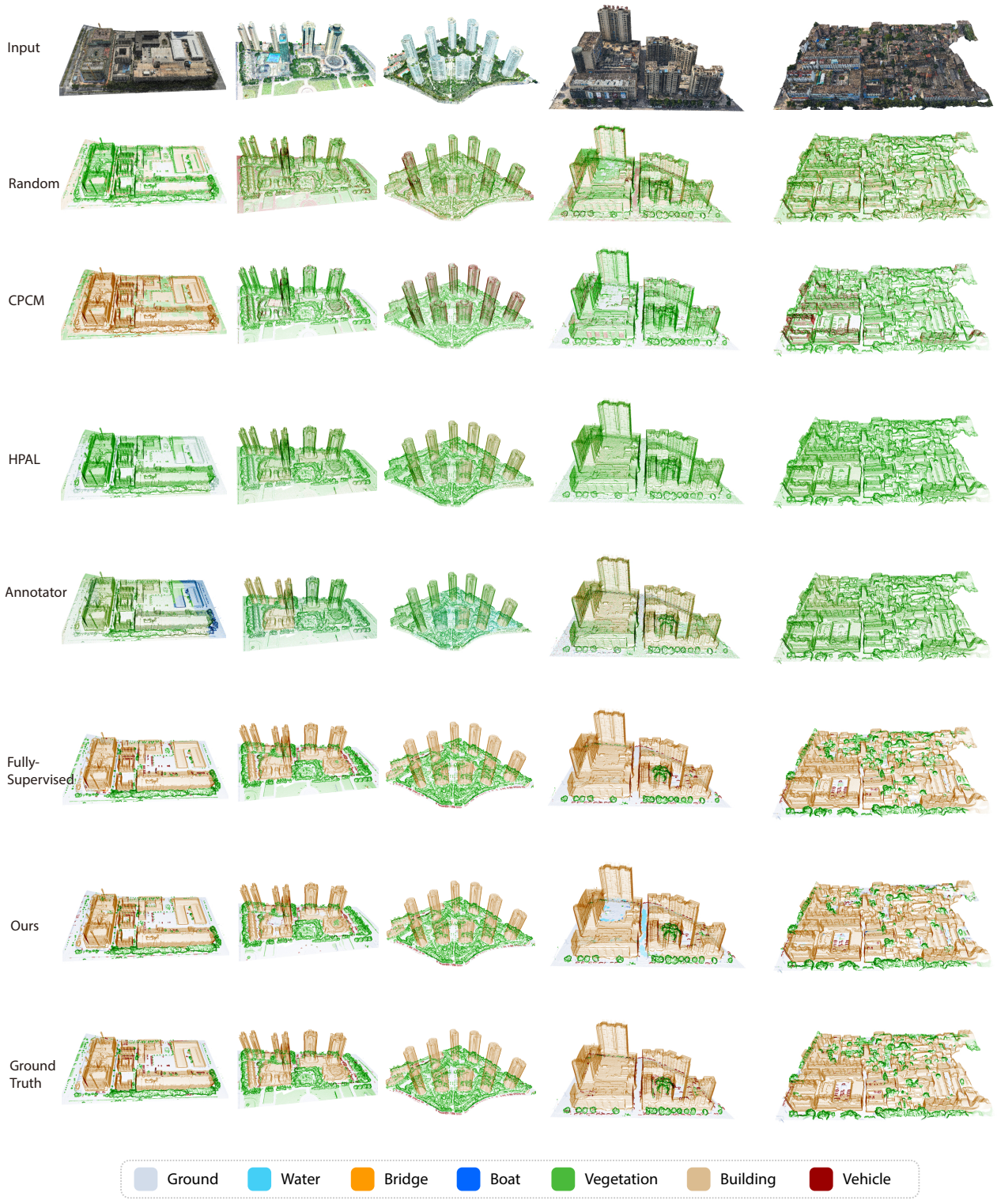

Figure 4. Additional qualitative comparisons on UrbanBIS (1/2).

Figure 5. Additional qualitative comparisons on UrbanBIS (2/2).

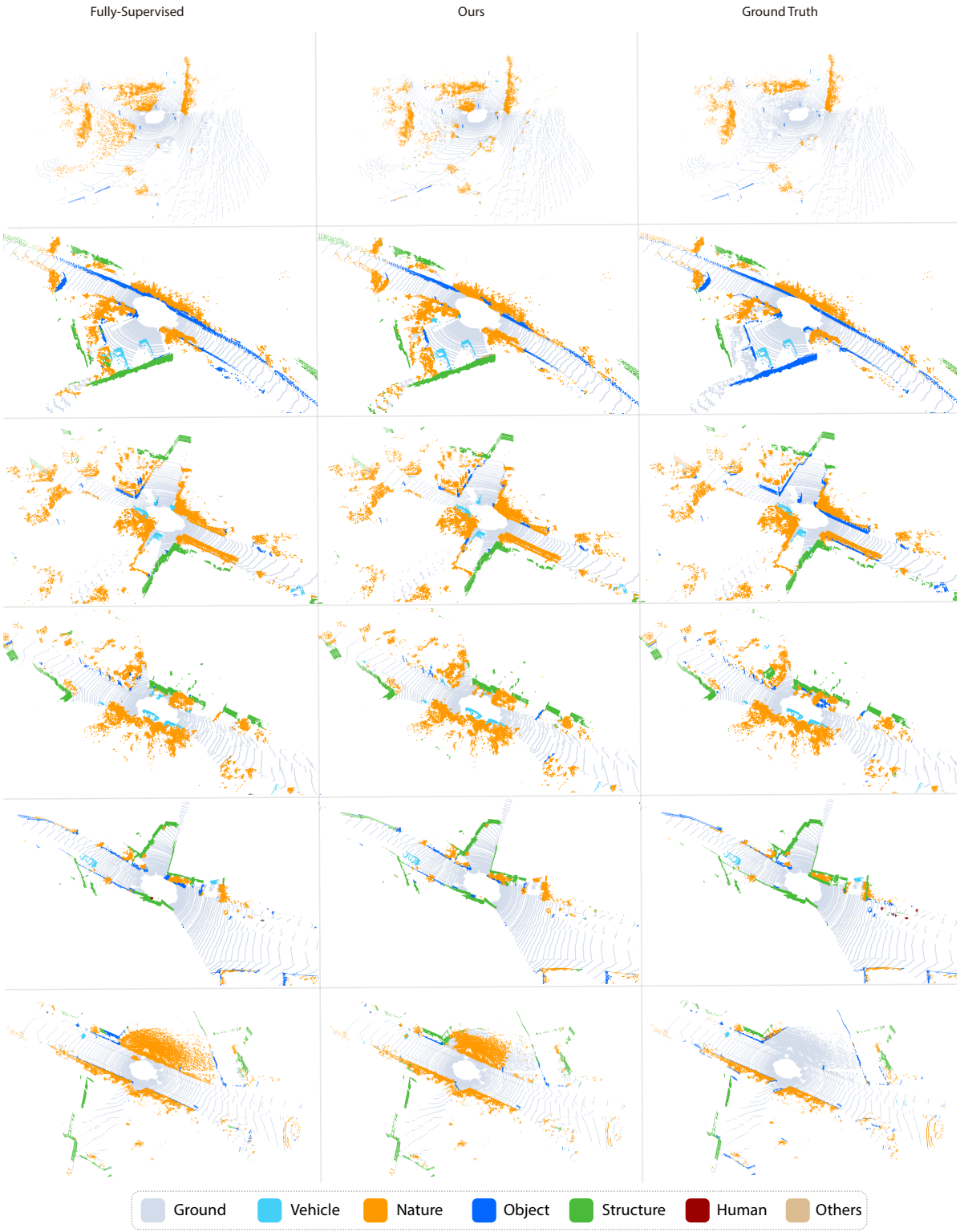

Figure 6. Additional qualitative comparisons on SemanticKITTI (1/1).

References

[1] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 1534–1543, 2016. 4

[2] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proc. Int. Conf. on Computer Vision*, pages 9296–9306, 2019. 4

[3] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *Int. J. Computer Vision*, 130(2):316–343, 2022. 4

[4] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 4867–4876, 2020. 3

[5] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. In *Proc. Int. Conf. on Learning Representations*, pages 1–15, 2015. 3

[6] Y. Lin, C. Wang, D. Zhai, W. Li, and J. Li. Toward better boundary preserved supervoxel segmentation for 3d point clouds. *ISPRS J. Photogrammetry and Remote Sensing*, 143:39–47, 2018. 2, 3

[7] L. Liu, Z. Zhuang, S. Huang, X. Xiao, T. Xiang, C. Chen, J. Wang, and M. Tan. Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 18413–18422, 2023. 4, 5

[8] Z. Liu, X. Qi, and C.-W. Fu. You only need one thing one click: Self-training for weakly supervised 3D scene understanding. *arXiv preprints arXiv:2303.14727*, 2023. 5

[9] F. Shao, Y. Luo, P. Liu, J. Chen, Y. Yang, Y. Lu, and J. Xiao. Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning. In *Proc. ACM Int. Conf. on Multimedia*, pages 2575—-2585, 2022. 5

[10] B. Xie, S. Li, Q. Guo, C. Liu, and X. Cheng. Annotator: A generic active learning baseline for lidar semantic segmentation. In *Proc. Conf. on Neural Information Processing Systems*, pages 48444–48458, 2023. 4, 5

[11] Z. Xu, B. Yuan, S. Zhao, Q. Zhang, and X. Gao. Hierarchical point-based active learning for semi-supervised point cloud semantic segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 18098–18108, 2023. 4, 5

[12] G. Yang, F. Xue, Q. Zhang, K. Xie, C.-W. Fu, and H. Huang. Urbanbis: A large-scale benchmark for fine-grained urban building instance segmentation. In *Proc. SIGGRAPH*, pages 16:1–16:11, 2023. 3, 4