International Cartographic Association
Association Cartographique Internationale

Urban-Scale Semantic Segmentation Using PointMamba and Mobile Laser Scanning Point Clouds

Jiafeng Wu [a], Lingfei Ma [a,*], Hongxin Yang [a], Jonathan Li [b]

[a] *School of Geospatial Artificial Intelligence, East China Normal University, Jiafeng Wu - 51273901114@stu.ecnu.edu.cn, Lingfei Ma - mlfcufe@163.com, Hongxin Yang - hxyang@geoai.ecnu.edu.cn*
[b] *Department of Geography and Environmental Management, University of Waterloo, Jonathan Li - junli@uwaterloo.ca*

* Corresponding Author

Abstract: Point cloud semantic segmentation is a critical task in autonomous driving and digital twin applications. This study introduces a novel semantic segmentation approach leveraging the PointMamba network, specifically designed to address the challenges of complex urban scene point cloud data. The PointMamba network integrates a state space model (SSM) with point cloud serialization and advanced feature extraction techniques, yielding significant performance improvements in semantic segmentation tasks. PointMamba was rigorously evaluated on the Toronto3D urban scene point cloud dataset, achieving an Overall Accuracy (OA) of 93.94% and a mean Intersection over Union (mIoU) of 66.03%. Comparative studies demonstrated that PointMamba outperformed existing point-based methods, including PointNet++ and PointNet, in handling intricate urban environments, delivering superior semantic segmentation outcomes on complex urban road environments.

Keywords: PointMamba, urban scenes, point clouds, semantic segmentation

1. Introduction

LiDAR scanning technology has become widely adopted in autonomous driving, remote sensing, and mobile mapping due to its ability to efficiently acquire high-precision 3D point cloud data (Long et al., 2021). Compared to traditional methods, LiDAR-based 3D data acquisition demonstrates superior efficiency, accuracy, and adaptability across diverse environments (Ma et al., 2018). In point cloud processing, semantic segmentation involves clustering the input data into homogeneous regions, where points within the same region share identical attributes (Nguyen and Le, 2013). As a high-level task, semantic segmentation is critical for scene understanding and has become an essential component in complex scene analysis and digital twin construction (Yang et al., 2022).

Traditional rule-based approaches for point cloud classification and segmentation typically rely on manually extracted features combined with machine learning algorithms to build discriminative models (Li et al., 2024). However, in complex urban environments, LiDAR point clouds are often sparse, unordered, and heavily affected by noise and outliers, significantly limiting the effectiveness of such traditional methods. In contrast, deep neural networks have emerged as the dominant approach for point cloud semantic segmentation, owing to their superior feature learning capabilities (Yang et al., 2022).

More recently, Mamba model, a new network backbone using the State Space Model (SSM) (Gu et al., 2021), has achieved superior context learning capability in the processing of sequence data. Accordingly, PointMamba model was designed for the classification and part segmentation tasks using point clouds collected in indoor environments (Liang et al., 2024), which outperformed most transformer-based nerural networks.

Thus, this paper extends PointMamba for the first time to semantic segmentation of large-scale outdoor point clouds, significantly enhancing its capability to process urban-scale data. Feature extraction in PointMamba consists of two branches. The first extracts features through N stacked PointMamba Blocks, selecting three representative layers for global feature concatenation. The second maps features to individual points using PointNet++'s feature propagation. Global and per-point features are then concatenated and passed to the segmentation head for final predictions.

The main contributions of this study are three-fold: (1) The PointMamba model was extended to point cloud semantic segmentation for the first time and evaluated in complex urban environments, demonstrating its effectiveness in large-scale segmentation tasks. (2) The efficiency and applicability of PointMamba were validated through comparative experiments with the PointNet series, particularly in handling complex and fine-grained structures. (3) To further enhance segmentation performance, the Block Division strategy was integrated into the PointMamba framework, optimizing point cloud data organization, and improving the model's ability to learn inherent local features.

2. Related Works

Point cloud semantic segmentation represents a fundamental task in 3D data processing, facilitating a deeper un-

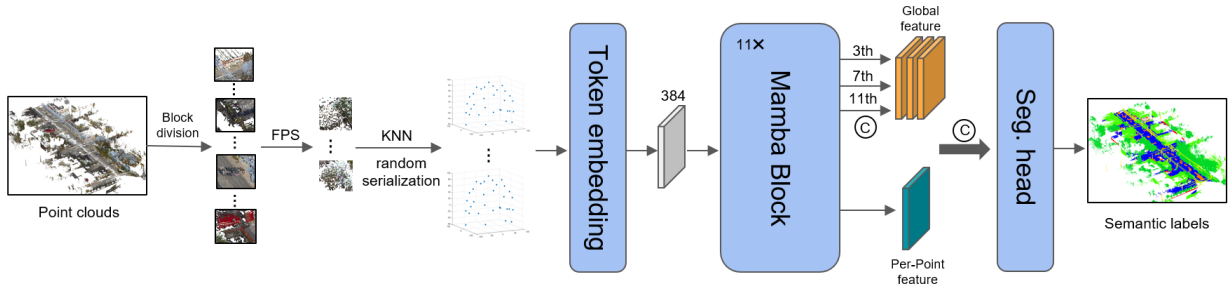

Figure 1. Architecture of the PointMamba network.

derstanding of data patterns and enabling precise identification of both global and local relationships among categories or features in specific scenarios. Currently, point cloud semantic segmentation methods can be categorized into two approaches: point-based methods and Transformer-based methods.

Point-Based Segmentation: The inherent disorder and translational invariance of point clouds in 3D space pose challenges to the direct application of traditional 2D and 3D convolutional neural networks. To solve this problem, point-based segmentation methods have been proposed. PointNet, as a pioneering work in directly processing point clouds, leveraged permutation invariance and utilized shared MLPs and symmetric pooling to capture global features (Qi et al., 2017a). However, its reliance on max pooling restricted its ability to model local structures. PointNet++ addressed this limitation by introducing a hierarchical framework that segmented point clouds into overlapping regions, extracted local features using modified PointNet operations, and hierarchically aggregated them to capture global representations (Qi et al., 2017b). Techniques such as farthest point sampling (FPS) and ball-query algorithms enable PointNet++ to effectively handle varying densities and scales, excelling in tasks like object recognition and semantic segmentation. Further advancements, such as point convolution, adaptively learn weight functions from geometric information. For instance, RSNet employed 1×1 convolutions to extract point-wise features, which were processed through a Local Dependency Module (LDM) to capture local context, thereby enhancing segmentation accuracy in complex 3D scenarios (Huang et al., 2018).

Transformer-based approaches: Transformers are well-known due to their attention mechanisms, with self-attention as the core component and positional encoding enabling the modeling of token order within sequences. Positional encoding is crucial for capturing relative positional relationships. Accordingly, Point Transformer integrated MLP-based positional encoding into vectorized attention and incorporated a KNN-based downsampling module to reduce point resolution (Zhao et al., 2021). Point Transformer v2 refines the baseline by introducing encoding multipliers for relational vectors and a partition-based pooling strategy to better align geometric information (Wu et al., 2022). FastPointTransformer simplified the architecture with a lightweight local self-attention module that efficiently learned positional information while reducing spatial complexity (Park et al., 2022). Point Transformer v3 further enhanced efficiency by replacing KNN-based neighborhood searches with serialized neighborhood mapping, enabling broader receptive field coverage (Wu et al., 2024).

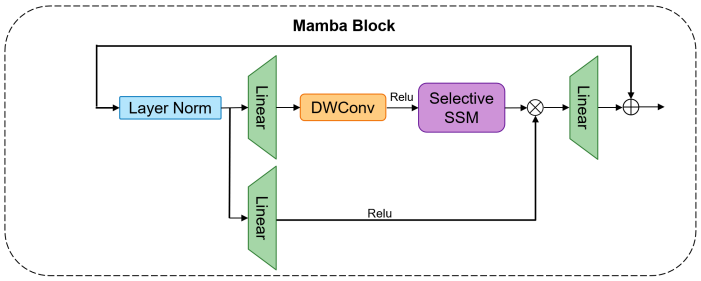

Figure 2. Structure of the Mamba Block.

3. Methods

Fig.1 illustrates the PointMamba architecture designed for point cloud semantic segmentation in urban scenarios. The following sections provide a detailed explanation of each component of the PointMamba.

3.1 Block Division

Given the large volume of point cloud data, the scene is first divided into smaller blocks to accelerate computation during the training process. Cross-validation testing determines the block size to be 10m × 10m, and 16,000 points are randomly sampled within each block, potentially including duplicates. The normals are calculated using a 0.8 m radius and a neighborhood size of 30 points. This radius balances between preserving fine-grained geometry and ensuring computational efficiency in densely sampled urban scenes. While optimized for the Toronto3D dataset, this parameter may need adjustment when applying the method to sparser or denser environments.The input features for each point include the coordinates offset from the UTM origin, the normal vector, and the point intensity. Although the number of returns is also informative, it is not included in the Toronto3D dataset used in this study.

Additionally, the proposed block division method spatially adjacent points into the same data block, which facilitates the model's ability to more effectively learn local features, such as surface textures and geometric shapes, contributing to capturing fine-grained spatial structures.

3.2 FPS and KNN

After the block division, the Farthest Point Sampling (FPS) method is applied to uniformly sample points within each

Proceedings of the International Cartographic Association, 7, 16, 2025 | https://doi.org/10.5194/ica-proc-7-16-2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.

block, ensuring a spatially even distribution of the sampled points (Qi et al., 2017a), which is widely recognized for its efficiency in downsampling large-scale point cloud data. Subsequently, the K-Nearest Neighbors (KNN) algorithm is employed to construct local neighborhood relationships for each sampled point, facilitating the capture of local geometric features. To reduce overfitting to point order and enhance generalization, the input order of the point cloud is randomly permuted within each batch, while the XYZ coordinates themselves remain unchanged. This technique increases the robustness of the model to point ordering, which is arbitrary in unstructured point clouds.

3.3 Token Embedding

The PointMamba framework maps unbiased local patches into the feature space through a lightweight PointNet-based point embedding layer (Qi et al., 2017a), resulting in serialized point tokens $E_{0h} \in \mathbb{R}^{n \times C}$, which are derived from a random sequence representation.

3.4 The Mamba Block

After the token embedding layer, the output is fed into the Mamba Block, which is a core component of the PointMamba network. As illustrated in Fig.2, the Mamba Block is designed to efficiently capture both local and global feature interactions. It follows a structured pipeline that integrates normalization, linear transformations, depthwise convolutions, and selective state-space modeling. This architecture refines spatial features and enhances the network's representational capacity. The detailed process is as follows:

Layer Norm. The Layer Normalization (LN) operation normalizes the input feature $\mathbf{x}$ to stabilize training and prepares the data for subsequent transformations. The normalization process is mathematically defined as follows:

$$\mathrm{LN}(\mathbf{x}) = \gamma \odot \frac{\mathbf{x} - \hat{\mu}}{\hat{\sigma}} + \beta \tag{1}$$

where $\hat{\mu}$ and $\hat{\sigma}$ represent the mean and standard deviation of the input $\mathbf{x}$, respectively. γ and β are learnable parameters that scale and shift the normalized inputs, respectively. $\odot$ denotes the element-wise multiplication.

Linear Transformation. The normalized input is passed through a linear transformation to project it into a higher-dimensional space. The transformation is represented by:

$$Y_{n \times m} = X_{n \times o} W_{o \times m} + b_{n \times 1}, \tag{2}$$

where $X_{n \times o}$ represents the normalized input features. $W_{o \times m}$ and $b_{n \times 1}$ are learnable parameters, denoting the weight matrix and bias term, respectively. $Y_{n \times m}$ is the output feature matrix projected into a higher-dimensional space.

Depthwise Convolution (DWConv). A depthwise convolution is applied to enhance local feature interactions, focusing on nearby spatial relationships using the following equation:

$$h_{\mathrm{dwconv}}(t) = \mathrm{ReLU}(\mathrm{DWConv}(h_{\mathrm{linear1}}(t))) \tag{3}$$

Selective State Space Model. Inspired by control theory, the state-space model (SSM) is conceptualized as a linear time-invariant system that transforms an input sequence $x(t) \in \mathbb{R}^L$ into an output sequence $y(t) \in \mathbb{R}^L$. Formally, it is characterized by a set of ordinary differential equations (ODEs) as follows:

$$\dot{h}(t) = Ah(t) + Bx(t), \tag{4}$$

$$y(t) = Ch(t) + Dx(t), \tag{5}$$

Residual Connections. The block incorporates residual connections to preserve information flow and improve gradient propagation. Two residual paths are used. One path adds the output of the SSM module back to the input after an additional linear transformation as follows:

$$h_{\mathrm{res1}}(t) = W_2 h_{\mathrm{SSM}}(t) + b_2 \tag{6}$$

Another path directly connects the input of the block to the final output as follows:

$$h_{\mathrm{out}}(t) = h_{\mathrm{res1}}(t) + h(t) \tag{7}$$

3.5 The Feature Fusion

After feature extraction through eleven stacked Mamba Blocks, features from the 3rd, 7th, and 11th layers are fused to form global features, denoted as f_i. These global features are subsequently integrated with per-point features f_i^k, extracted using a lightweight PointNet++ module. The integration is expressed by:

$$f_i^h = concat(f_i, f_i^k) \tag{8}$$

where f_i indicates the global feature aggregated from the 3rd, 7th, and 11th layers of the Mamba Block. f_i^k represents the local per-point feature extracted by the PointNet++ module. f_i^h is the final fused feature, which combines global and local features. The fused features f_i^h enable precise and fine-grained semantic segmentation of the point cloud.

4. Experiments Results And Discussion

4.1 Dataset

In this study, the Toronto3D dataset (Tan et al., 2020), which was a LiDAR point cloud dataset designed for urban scene analysis, was used to train and evaluate the PointMamba method. It was collected using a Mobile LiDAR Scanner (MLS) system, representing urban streetscapes and spanning a wide area encompassing both sides of the road. This comprehensive coverage provides an excellent benchmark for assessing the ability of the PointMamba method to extract detailed urban facade information effectively. More specifically, L001, L003, and L004 tiles were selected as the training set, and the L002 tile was used as the validation and testing set.

Proceedings of the International Cartographic Association, 7, 16, 2025 | https://doi.org/10.5194/ica-proc-7-16-2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.

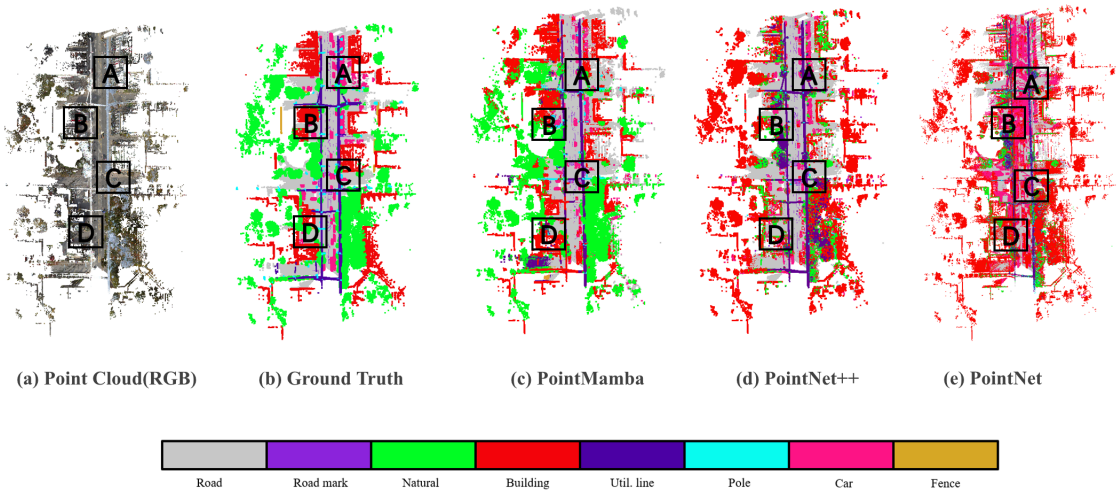

Figure 3. Comparison of visualization results on the Toronto3D dataset.

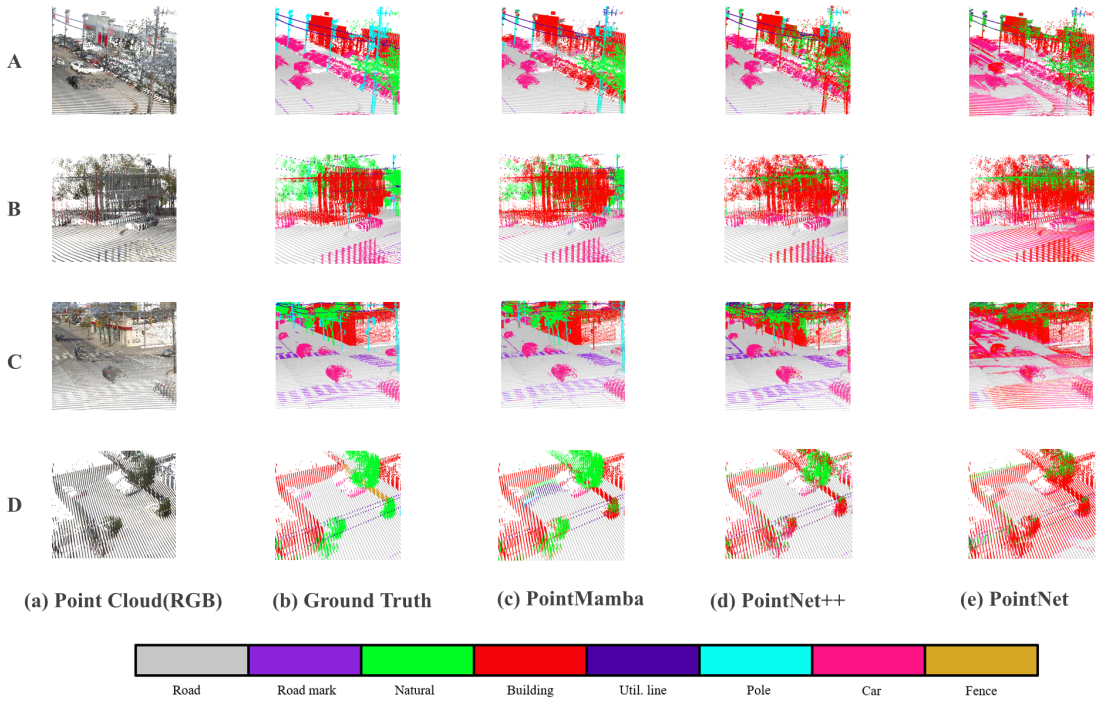

Figure 4. Comparison of visualization results on the Toronto3D dataset, where A to D represent four close-view areas.

Table 1. Quantitative Result of Segmentation Performance compared with PointMamba in OA, mIoU, and class IoUs. Bold values indicate the best model performance.

Method	OA	mIoU	Road	Road mrk	Natural	Building	Util line	Pole	Car	Fence
PointNet	75.71	30.14	89.0	40.6	27.1	22.6	28.8	6.1	24.7	2.2
PointNet++	84.16	43.42	91.9	35.8	56.0	46.2	47.1	18.4	51.9	0
PointMamba	**93.94**	**66.03**	**93.9**	**47.1**	**91.8**	**88.0**	32.3	**78.7**	**81.4**	**15.1**

4.2 Experiments Settings

All experiments were conducted on a desktop computer equipped with an Intel Core i5-14600F CPU, an NVIDIA GeForce RTX 4070 Super GPU (12GB), and 32GB of RAM. The model was implemented using PyTorch 1.13.0 with CUDA 11.7 in a Python 3.9 environment. For PointMamba, the input point number per batch was set to 16,384, with training and testing batch sizes of 2 and 4, respectively. The training process consisted of 200 epochs. The model was optimized using the AdamW optimizer, with a 5% reduction in the learning rate after each epoch. The number

Proceedings of the International Cartographic Association, 7, 16, 2025 | https://doi.org/10.5194/ica-proc-7-16-2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.

of nearest neighbors K was set to 32. The final trained model was evaluated using Overall Accuracy (OA), per-class Intersection over Union (IoU), and mean Intersection over Union (mIoU) to comprehensively assess the performance of PointMamba.

4.3 Experimental Results and Discussion

Fig.3 presents the qualitative semantic segmentation results across different models, including the proposed PointMamba PointNet++, and PointNet, alongside the ground truth and the original point cloud (RGB). Additionally, four representative regions (A, B, C, D) from the Toronto3D testing set were selected to evaluate the segmentation performance in complex urban scenarios (see Fig.4).

Table 1 provides the quantitative evaluation, showing that PointMamba achieved strong segmentation performance in complex urban environments. The comparative results demonstrated the superiority of PointMamba over PointNet and PointNet++ in semantic segmentation tasks. Quantitatively, PointMamba achieved significantly higher OA (93.94%) and mIoU (66.03%) compared to PointNet (75.71%, 30.14%) and PointNet++ (84.16%, 43.42%). It excelled in key categories such as natural surfaces (91.8%), buildings (88.0%), and cars (81.4%). Additionally, PointMamba exhibited a distinct advantage in segmenting fine-grained structures, including poles (78.7%) and fences (15.1%), with notable reductions in omissions and misclassifications.

Despite its strengths, some limitations remain. In region A, certain pole-like structures were misclassified as buildings, while in region C, parts of trees were incorrectly labeled as buildings. Similarly, in region D, some road points were misclassified as road markings. These errors likely arise from PointMamba's limited capacity to effectively capture local-scale features, posing challenges in segmenting fine-grained structures.

5. Conclusion

This study extends the PointMamba network to semantic segmentation in complex urban environments and systematically evaluates its performance on urban-scale point clouds. Results show that PointMamba achieves excellent performance, with an OA of 93.94% and mIoU of 66.03%, significantly outperforming classical models like PointNet and PointNet++. It excels in key categories such as natural surfaces, buildings, and cars, and shows clear advantages in segmenting fine-grained structures like poles and fences. Future work will incorporate advanced multi-scale feature extraction to improve accuracy and extend PointMamba to other datasets and domains, validating its robustness for autonomous driving and smart cities.

References

Gu, A., Goel, K. and Ré, C., 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Huang, Q., Wang, W. and Neumann, U., 2018. Recurrent slice networks for 3d segmentation of point clouds. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2626–2635.

Li, H., Ma, L., Guan, H., Qin, N., Zang, Y. and Wang, L., 2024. Multi-granularity feature fusion for point cloud semantic segmentation under urban scenes. In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, pp. 8612–8615.

Liang, D., Zhou, X., Xu, W., Zhu, X., Zou, Z., Ye, X., Tan, X. and Bai, X., 2024. Pointmamba: A simple state space model for point cloud analysis. In: *Advances in Neural Information Processing Systems*.

Long, X., Cheng, X., Zhu, H., Zhang, P., Liu, H., Li, J., Zheng, L., Hu, Q., Liu, H., Cao, X. et al., 2021. Recent progress in 3d vision. *Journal of Image and Graphics* 26(6), pp. 1389–1428.

Ma, L., Li, Y., Li, J., Wang, C., Wang, R. and Chapman, M. A., 2018. Mobile laser scanned point-clouds for road object detection and extraction: A review. *Remote Sensing* 10(10), pp. 1531.

Nguyen, A. and Le, B., 2013. 3d point cloud segmentation: A survey. In: *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*, IEEE, pp. 225–230.

Park, C., Jeong, Y., Cho, M. and Park, J., 2022. Fast point transformer. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16949–16958.

Qi, C. R., Su, H., Mo, K. and Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660.

Qi, C. R., Yi, L., Su, H. and Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K. and Li, J., 2020. Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 202–203.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T. and Zhao, H., 2024. Point transformer v3: Simpler faster stronger. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4840–4851.

Wu, X., Lao, Y., Jiang, L., Liu, X. and Zhao, H., 2022. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems* 35, pp. 33330–33342.

Yang, B., Chen, c. and Dong, Z., 2022. 3d geospatial information extraction of urban objects for smart surveying and mapping. *Acta Geodaetica et Cartographica Sinica* 51(7), pp. 1476.

Zhao, H., Jiang, L., Jia, J., Torr, P. H. and Koltun, V., 2021. Point transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268.