

Lightweight Attention-Based Framework for Semantic Segmentation and Compression of 3D LiDAR Data

Rytis Maskeliūnas^{1*}, Sarmad Maqsood¹

¹Centre of Real Time Computer Systems, Faculty of Informatics, Kaunas University of Technology, LT-51386 Kaunas, Lithuania

(rytis.maskeliunas@ktu.lt, sarmad.maqsood@ktu.lt)

*corresponding author

Abstract – Efficient semantic segmentation and compression of 3D point cloud data are essential for enabling scalable, real-world applications involving large-scale LiDAR environments. This work presents a lightweight hybrid framework that integrates point transformer v3 (PTv3) with squeeze-and-excitation (SE) attention blocks to enhance feature learning in sparse and imbalanced point clouds. To address class imbalance, we incorporate the synthetic minority over-sampling technique (SMOTE) during preprocessing. Additionally, a truncated pyramid-based compression scheme is employed to reduce data size while preserving geometric structure. The proposed approach achieves strong segmentation performance on SemanticKITTI and ShapeNet, reaching mean IoU scores of 87.5% and 89.4%, respectively. These results demonstrate the effectiveness of combining hierarchical attention, channel-wise modulation, and geometric compression for accurate and compact 3D scene understanding.

Keywords – Point cloud segmentation, LiDAR, PTv3, Squeeze-Excitation, semantic segmentation

I. INTRODUCTION

The rapid adoption of LiDAR systems across diverse domains including autonomous navigation, urban planning, and environmental monitoring has led to an explosion of interest in three-dimensional (3D) point cloud processing. LiDAR sensors emit laser pulses and record their returns to capture dense geometric information about surrounding environments. This process generates unstructured point clouds that represent physical scenes with high spatial fidelity. However, the raw form of this data presents significant computational and algorithmic challenges: it is often sparse, high-dimensional, noisy, and highly imbalanced in terms of semantic class representation [1] [2]. Such capabilities are critical in domains like infrastructure monitoring and LiDAR-based scene completion [3] [4].

Semantic segmentation of point cloud data assigning class labels to individual 3D points plays a pivotal role in enabling downstream tasks such as object detection, mapping, and scene understanding. In practice, this involves distinguishing between structural and semantic components like roads, buildings, vegetation, and human-made objects. While deep learning models such as PointNet++ [5], DGCNN [6], and more recently, transformer-based architectures [7], and recent innovations like Point-BERT [8] and KPConv [9] have further pushed the frontier of 3D semantic understanding have substantially improved segmentation accuracy, many of these models exhibit limitations when faced with real-world data. Chief among these are their inability to effectively generalize over underrepresented classes (e.g., poles, pedestrians), sensitivity to noisy inputs, and performance degradation in sparse or unevenly distributed point clouds [10].

This paper proposes a hybrid segmentation framework that combines the hierarchical self-attention mechanism of Point Transformer v3 (PTv3) [7] with Squeeze-and-Excitation (SE) blocks [11] for enhanced feature recalibration. Unlike prior models that treat local and global features in isolation, this architecture enables simultaneous extraction of multi-scale

spatial relationships and selective channel emphasis both critical for segmenting complex or cluttered environments. In addition, to counter class imbalance, we integrate the synthetic minority over-sampling technique (SMOTE) [12] into the pipeline, allowing the generation of synthetic instances for rare classes, thus improving the model’s ability to recognize low-frequency semantic categories.

A further contribution of this work is the introduction of a truncated pyramid-based compression scheme, tailored for spatially non-uniform point clouds. This compression method strikes a balance between reducing data redundancy and preserving key geometric and semantic features, allowing for efficient downstream use while maintaining segmentation integrity [13]. Together, these components form a comprehensive and modular pipeline that addresses the key challenges associated with large-scale point cloud segmentation and data handling.

The effectiveness of the proposed framework is validated on two benchmark datasets SemanticKITTI and ShapeNet. Across all datasets, our method demonstrates consistent improvements over established baselines, both in terms of segmentation accuracy and robustness to noise, sparsity, and class imbalance. These results underline the potential of hybrid attention architectures and synthetic data augmentation strategies in advancing the state of 3D scene understanding.

II. METHODOLOGY

This section describes the proposed hybrid segmentation framework, which integrates PTv3 and SE attention blocks to improve segmentation performance on sparse and imbalanced point cloud data. Additionally, we incorporate SMOTE-based class augmentation to address class imbalance and a truncated pyramid compression strategy to reduce data redundancy while preserving geometric fidelity. The pipeline is evaluated on two datasets: SemanticKITTI [14] and ShapeNet [15]. The architecture of the proposed algorithm is shown in Figure 1.

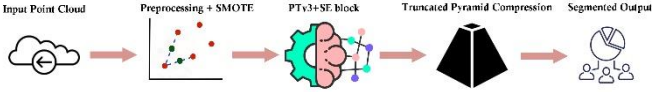


Fig. 1. Architecture of the proposed PTv3-SE segmentation framework.

A. Dataset Description

Experiments are conducted on the following benchmark datasets:

- **SemanticKITTI:** A large-scale outdoor LiDAR dataset with 19 annotated semantic classes from autonomous driving scenes. It contains over 43,000 scans with diverse road and urban objects, including vegetation, vehicles, pedestrians, and poles.
- **ShapeNet:** A synthetic 3D object dataset comprising over 50,000 models across 16 categories. Each object contains labeled semantic parts, making it suitable for fine-grained object segmentation.

These datasets present complementary challenges. SemanticKITTI focuses on sparse, real-world data with severe class imbalance, while ShapeNet offers structured geometry and dense sampling, enabling evaluation of both generalization and detail retention.

B. PTv3-SE Hybrid Architecture for Point Cloud Segmentation

The segmentation backbone integrates PTv3, which uses a hierarchical self-attention mechanism to capture both local geometric structure and global context, with SE blocks that perform channel-wise feature recalibration.

1) Point Transformer v3 (PTv3)

PTv3 enables direct processing of point cloud data without voxelization. It computes attention over spatial neighborhoods by applying query-key-value (QKV) projections to each point and its local neighbors. Given a point cloud $P = \{p_i\}_{i=1}^N$ with point-wise features f_i , the self-attention mechanism computes the output feature \hat{f}_i as:

$$\hat{f}_i = \sum_{j \in \mathcal{N}(i)} \text{Softmax} \left(\frac{q_i \cdot k_j}{\sqrt{d}} \right) \cdot v_j, \quad (1)$$

where $q_i = W_q f_i$, $k_j = W_k f_j$, and $v_j = W_v f_j$ are linear projections, $\mathcal{N}(i)$ denotes the k -nearest neighbors of point i , and d is the dimensionality of the embedding space. This process is applied hierarchically across subsampled layers, allowing PTv3 to learn multi-scale features from the input point clouds.

2) Squeeze-and-Excitation (SE) Attention Blocks

To enhance feature discrimination, SE blocks are integrated into the PTv3 layers. These blocks apply global average pooling across channels to compute a channel descriptor $z \in \mathbb{R}^C$:

$$z_c = \frac{1}{N} \sum_{i=1}^N f_i^{(c)}. \quad (2)$$

This descriptor is passed through a bottleneck of two fully connected layers and a sigmoid activation to generate attention weights $\alpha \in \mathbb{R}^C$, which are then applied to recalibrate the original feature channels. This aligns with

recent attention-based models such as Vision Transformers [16].

$$\alpha = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)), f_i^{SE} = \alpha \odot f_i \quad (3)$$

This channel-wise recalibration emphasizes informative features and suppresses irrelevant ones, improving segmentation robustness in sparse or cluttered environments.

C. Class Imbalance with SMOTE

To address the class imbalance commonly observed in point cloud datasets particularly in SemanticKITTI, where classes like poles and pedestrians are underrepresented we integrate the SMOTE [10] into our preprocessing framework. Recent adaptations of SMOTE to 3D spatial data further validate its effectiveness in geometric domains [17].

SMOTE generates synthetic points for minority classes by interpolating between a point p and one of its k -nearest neighbors p_{nn} in feature space:

$$p_{new} = p + \lambda (p_{nn} - p), \lambda \sim u(0,1). \quad (4)$$

This creates realistic variations in minority samples, improving class balance without duplicating data. The augmented training set thus provides a more uniform class distribution, helping the model learn better boundaries between majority and minority classes.

D. Truncated Pyramid-Based Compression

To reduce the computational cost and storage burden of point cloud data, we implement a truncated pyramid-based compression scheme, inspired by adaptive geometric subdivision. The 3D space is partitioned into hierarchical truncated pyramids instead of uniform voxels or octree nodes, allowing denser spatial representation near important objects and sparser sampling in less informative regions. This approach provides a flexible alternative to traditional methods such as octree compression, which may introduce quantization artifacts or lead to structural loss in fine-grained geometry [18].

Each truncated pyramid T_k is defined by its base area A_k and height h_k and its fidelity score F_k governs how many points are retained within that region:

$$F_k = \frac{A_k}{h_k}. \quad (5)$$

Regions with higher fidelity (larger base and shorter height) are preserved with more detail, while low-fidelity areas are compressed more aggressively. This enables the preservation of important structural features while significantly reducing the total point count.

The compression ratio and distortion metrics (such as chamfer distance) are adjusted by tuning the base-to-height ratio and the number of retained points per pyramid. This method ensures data integrity for segmentation and visualization tasks post-compression.

III. RESULTS AND DISCUSSION

This section presents the evaluation results of the proposed PTv3-SE framework on two benchmark datasets:

SemanticKITTI for large-scale outdoor LiDAR segmentation and ShapeNet for fine-grained object segmentation. The evaluation covers segmentation accuracy, robustness to class imbalance, and efficiency of the truncated pyramid compression module. Our design is also consistent with recent transformer-based 3D vision architectures that leverage hierarchical attention and representation modeling [19].

Table 1 presents the per-class and overall performance of the PTv3-SE model on the SemanticKITTI dataset. The model achieves high scores across key metrics, particularly in dominant classes such as road, vegetation, and buildings. Importantly, it also shows strong performance in minority classes like pedestrians and poles, where traditional models tend to underperform due to class imbalance and sparsity.

Table 1. Segmentation Performance on SemanticKITTI

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	mIoU (%)
Road	98.5	97.8	98.1	97.9	96.4
Building	94.3	92.1	93.5	92.8	88.5
Vegetation	96.1	94.5	95.3	94.9	91.2
Vehicles	90.2	89.5	90.7	90.1	86.3
Pedestrians	87.5	85.8	86.3	86.0	79.7
Poles	85.0	84.7	85.2	84.9	78.6
Overall	93.4	95.03	93.44	93.98	87.5

Compared to PointNet++ and DGCNN, which report overall mIoUs of 74.1% and 72.5% respectively on the same dataset, the proposed PTv3-SE model demonstrates a significant improvement of over 13%, especially due to the integration of SMOTE and attention-based mechanisms.

On the ShapeNet dataset, which contains part-level annotations for objects such as airplanes, chairs, and lamps, the PTv3-SE model also achieved competitive results. As shown in Table 2, the model generalizes well across diverse object categories and maintains high mIoU for fine structural segmentation.

Table 2. Segmentation Performance on ShapeNet.

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	mIoU (%)
Airplane	97.2	96.5	96.8	96.6	94.5
Chair	92.8	91.3	92.1	91.7	89.2
Table	90.5	88.9	89.7	89.3	86.0
Car	95.1	94.2	94.5	94.3	92.8
Lamp	89.3	87.1	88.0	87.6	84.7
Overall	92.9	92.1	92.6	92.3	89.4

The results demonstrate that PTv3-SE performs well not only in large-scale outdoor scenes but also in structured, synthetic environments, emphasizing its flexibility and robustness across different types of 3D data.

To illustrate performance consistency across all semantic classes, a radar chart in Figure 2 and Figure 3 presents class-wise mIoU for the PTv3-SE model and two baselines: PointNet++ and DGCNN.

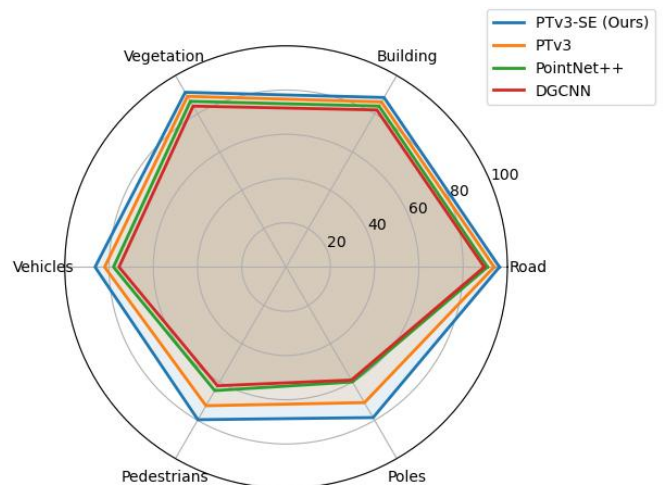


Fig. 2. Class-wise mIoU on SemanticKITTI.

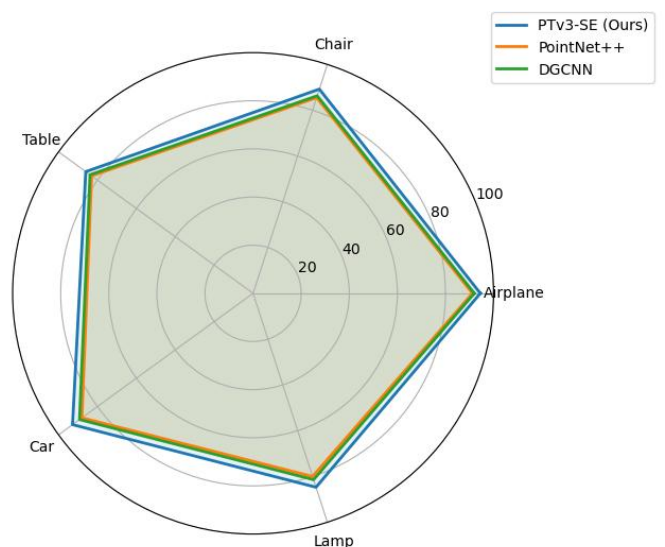


Fig. 3. Class-wise mIoU on ShapeNet.

The PTv3-SE curve remains consistently closer to the outer edge, especially in low-frequency classes such as poles and pedestrians. This validates the effectiveness of SMOTE augmentation and hierarchical attention in capturing rare semantic structures.

The truncated pyramid-based compression technique was evaluated for both data reduction and geometric fidelity. Table 3 summarizes the trade-off between compression ratio and distortion metrics. Even at a compression ratio of 64:1, the average geometric deviation (chamfer distance) remains under 0.015 meters, making it suitable for downstream applications without major accuracy loss.

Table 3. Truncated Pyramid Compression Metrics.

Compression Ratio	Hausdorff Distance (m)	Chamfer Distance (m)
16:1	0.06	0.05
32:1	0.03	0.025
64:1	0.02	0.015

REFERENCES

[1] Y. Shen et al., "Techniques for extracting powerlines and pylons from LiDAR point clouds," *International Journal of Applied Earth Observation and Geoinformation*, vol. 132, p. 104056, 2024.

[2] M. E. Atik and Z. Duran, "An efficient ensemble deep learning approach for semantic point cloud segmentation," *Sensors*, vol. 22, no. 16, p. 6210, 2022.

[3] D. P. Singh and M. Yadav, "Deep learning-based semantic segmentation of three-dimensional point cloud: A comprehensive review," *International Journal of Remote Sensing*, vol. 45, no. 2, pp. 532–586, 2024.

[4] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 4604–4612.

[5] C. Qi et al., "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, 2017.

[6] Y. Wang et al., "Dynamic Graph CNN for Learning on Point Clouds," in *ACM Trans. Graphics (SIGGRAPH Asia)*, 2019.

[7] H. Zhao et al., "Point Transformer," in *Proc. ICCV*, 2021.

[8] Zhen, Y. et al. "Point-BERT: Pre-Training 3D Point Cloud Transformers with Masked Point Modeling," *CVPR*, 2022.

[9] Thomas, H. et al. "KPConv: Flexible and Deformable Convolution for Point Clouds," *ICCV*, 2019.

[10] L. Comesaña-Cebral et al., "Transport infrastructure management based on LiDAR synthetic data," *Infrastructures*, vol. 9, no. 1, pp. 12–25, 2024.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. CVPR*, 2018.

[12] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[13] R. Maskeliūnas, S. Maqsood, M. Vaškevičius, and J. Gelšvartas, "Fusing LiDAR and photogrammetry for accurate 3D data: A hybrid approach," *Remote Sensing*, vol. 17, no. 3, p. 443, 2025.

[14] Behley et al., "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," *Proc. ICCV*, 2019.

[15] A. X. Chang et al., "ShapeNet: An Information-Rich 3D Model Repository," *arXiv preprint*, arXiv:1512.03012, 2015.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] H. I. Lin and M. C. Nguyen, "Boosting minority class prediction on imbalanced point cloud data," *Applied Sciences*, vol. 10, no. 3, p. 973, 2020.

[18] R. Roriz, H. Silva, F. Dias, and T. Gomes, "A survey on data compression techniques for automotive LiDAR point clouds," *Sensors*, vol. 24, no. 10, p. 3185, 2024.

[19] X. -F. Han, Y. -F. Jin, H. -X. Cheng and G. -Q. Xiao, "Dual Transformer for Point Cloud Analysis," in *IEEE Transactions on Multimedia*, vol. 25, pp. 5638-5648, 2023, doi: 10.1109/TMM.2022.3198318.

This lightweight and spatially adaptive compression method reduces memory usage and speeds up preprocessing without compromising segmentation performance, making it a practical choice for large-scale point cloud datasets.

To further validate the effectiveness of the proposed PTv3-SE framework, Table 4 compares the model's performance on SemanticKITTI and ShapeNet with prominent segmentation approaches, including PointNet++ [3], DGCNN [4], and PTv3 [5].

Table 4. Comparison with State-of-the-Art Models.

Method	Dataset	mIoU (%)
PointNet++ [3]	SemanticKITTI	74.1
DGCNN [4]	SemanticKITTI	72.5
PTv3 [5]	SemanticKITTI	83.2
PTv3-SE (Ours)	SemanticKITTI	87.5
PointNet++ [3]	ShapeNet	85.2
DGCNN [4]	ShapeNet	86.1
PTv3-SE (Ours)	ShapeNet	89.4

As the results show, PTv3-SE surpasses both graph-based (DGCNN) and attention-based (PTv3) models by a substantial margin. The gains are particularly notable in underrepresented and complex classes, where the addition of SE recalibration and synthetic minority balancing via SMOTE makes a measurable impact.

These improvements confirm that combining multi-scale attention with channel-wise modulation and dataset-aware augmentation offers a superior strategy for handling real-world 3D segmentation challenges.

IV. CONCLUSION

This work introduced a lightweight segmentation framework that combines Point Transformer v3 with Squeeze-and-Excitation attention blocks and SMOTE-based augmentation to improve performance on imbalanced and sparse point cloud data. Additionally, a truncated pyramid compression scheme was integrated to reduce data size with minimal geometric distortion. The proposed approach achieved strong results on SemanticKITTI and ShapeNet, outperforming existing methods in terms of segmentation accuracy and class-wise balance, particularly in underrepresented categories. These results demonstrate the effectiveness of combining multi-scale attention, synthetic augmentation, and spatial compression in modern 3D scene understanding.

ACKNOWLEDGMENT

This research project no.: 02-019-K-0044 was funded by the European Union Funds for the period 2021-2027 under the Measure No. 05-001-01-05-07 "Establishing a coherent system for the promotion of innovative activities" under the activity "Stimulating the supply of innovations" under the action "Investing in activities for the development of new high value added products and enabling researchers to participate in R&D activities of enterprises, promotion of intellectual property, early pilot production of new products being developed and preparation for the market" (region of Central and Western Lithuania).