Regional-to-Local Point-Voxel Transformer

Subjects: Computer Science, Artificial Intelligence
Contributor: Shuai Li, Hongjun Li

Semantic segmentation of large-scale indoor 3D point cloud scenes is crucial for scene understanding but faces challenges in effectively modeling long-range dependencies and multi-scale features. Researchers present RegionPVT, a novel Regional-to-Local Point-Voxel Transformer that synergistically integrates voxel-based regional self-attention and window-based point-voxel self-attention for concurrent coarse-grained and fine-grained feature learning. The voxel-based regional branch focuses on capturing regional context and facilitating inter-window communication. The window-based point-voxel branch concentrates on local feature learning while integrating voxel-level information within each window.

Keywords: point cloud ; semantic segmentation ; regional-to-local

1. Introduction

Semantic segmentation of large-scale point cloud scenes is a crucial task in 3D computer vision, serving as the core capability for machines to comprehend the 3D world. It has found extensive applications in autonomous driving [1][2], robotics [3][4], and augmented reality [5][6]. In particular, deep learning has made striking breakthroughs in computer vision over the past few years. Enabling reliable semantic parsing of point cloud data using deep neural networks has become an emerging hot research direction and attracted wide interest [7]. Unlike 2D images, 3D point clouds are intrinsically sparse and irregularly scattered in a continuous 3D space. They are unstructured in nature and often at a massive scale. These unique properties impose difficulties in directly adopting convolution operations, which have been the mainstay for 2D image analysis [8][9]. In recent years, convolutional networks (CNNs) [10][11][12] and Transformer [13][14][15] architectures have led to striking advances in semantic parsing of 2D visual data. However, efficiently learning discriminative representations from disordered 3D point sets using deep neural networks, especially at large-scale indoor scenes, remains a challenging open problem.

Abundant methods have explored the comprehension of 3D point clouds and obtained decent performance. In order to leverage convolutional neural networks (CNNs) for point cloud analysis, one category of approaches [16][17][18][19] first transforms the 3D points into discrete representations such as voxels, before applying CNN models to extract high-dimensional features. Another line of work [9][20][21][22][23], pioneered by PointNet [8], directly processes points in the native continuous space. Through alternating steps of grouping and aggregation, PointNet-style models are able to capture multi-scale contextual information from unordered 3D point sets. However, most of these existing methods concentrate on aggregating local feature representations but do not explicitly model long-range dependencies, which have been shown to be vital for capturing contextual information from distant spatial locations [24].

Transformers [25] based on self-attention come naturally with the ability to model long-range dependencies, and the permutation and cardinality invariance of self-attention in Transformers make them inherently suitable for point cloud processing. Recently, inspired by the transformer's remarkable success [13][14][15][26][27][28] in the 2D image domain, a number of studies [29][30][31][32] have investigated adapting Transformer architectures to process unstructured 3D point sets. Engel et al. [29] proposed a kind of point transformer algorithm, which incorporates standard self-attention to extract global features for capturing point relationships and shape information in the 3D space. Guo et al. [31] presented offset-attention that computes the offset difference between self-attention features and input features in an element-wise manner. Concurrently, a spectrum of scholars have explored embedding self-attention modules in diverse point cloud tasks, witnessing noteworthy successes as showcased in works like [30][33]. Despite the promising advancements in point cloud transformers, a clear limitation persists. These models need to generate expansive attention maps due to the use of conventional self-attention mechanisms, placing a high computational complexity (quadratic) and consuming a huge number of GPU memory. This methodology, while rigorous, becomes implausible when scaling up to expansive 3D point cloud datasets, thereby hindering large-scale modeling pursuits.

Furthermore, in an effort to aggregate localized neighborhood information from point clouds, Zhao et al. [30] introduced another kind of point transformer algorithm, which establishes local vector attention within neighboring point sets. Guo et

al. [31] proposed the use of neighbor embedding strategies to enhance point embedding. The PointSwin, as presented by Jiang et al. [34], employs self-attention based on a sliding window to capture local details from point clouds. While the two point transformers, PCT and the PointSwin, have achieved significant advancements, certain challenges continue to hinder their efficiency and performance. These methods fall short of establishing attention across features of different scales, which is crucial for 3D visual tasks [35]. For instance, a large indoor scene often encompasses both smaller instances (such as chairs and lamps) and larger objects (like tables). Recognizing and understanding the relationships between these entities necessitates a multi-scale attention mechanism. Moreover, when delving into large-scale scene point clouds, an optimal blend of both coarse-grained and fine-grained features becomes pivotal [36]. Coarse-grained features present a bird's eye view, providing a general overview of the scene, whereas fine-grained ones are key in identifying and interpreting small details. Integrating both these feature dimensions can significantly amplify the potential and accuracy of point cloud semantic segmentation, particularly in heterogeneous and complex scenarios.

In addressing the challenges discussed previously, researchers present a novel dual-branch block named the Regional-to-Local Point-Voxel Transformer Block (R2L Point-Voxel Transformer Block), specifically engineered for the semantic segmentation of large-scale indoor point cloud scenes. This block is designed to effectively capture both coarse-grained regional and fine-grained local features within large-scale indoor point cloud senses with linear computational complexity. The method has two key components, including a voxel-based regional self-attention for coarse-grained features modeling and a window-based point-voxel self-attention for fine-grained features learning and multi-scale feature fusion. More specifically, researchers first spatially partition the raw point clouds into non-overlapping cubes, termed "windows", following the concept similar to that of the Swin Transformer [14]. Then, researchers voxelize the point clouds using a window size unit and establish a hash table [37] between the points and voxels. Voxel-based regional self-attention is subsequently applied among the nearest neighboring voxels to obtain coarse-grained features. Finally, the aggregated voxels serving as special "points" participate in the window-based point-voxel self-attention with their corresponding points to obtain fine-grained features. The voxel-based regional self-attention achieves information interaction between different windows while aggregating voxel features. Meanwhile, the window-based point-voxel self-attention not only focuses on learning fine-grained local features, but also captures high-level voxel information, enabling multi-scale feature fusion by treating voxels as specialized points.

Building upon the R2L Point-Voxel Transformer Block, researchers propose a network for large-scale indoor point cloud semantic segmentation, named RegionPVT (Regional-to-Local Point-Voxel Transformer), as depicted in **Figure 1**.

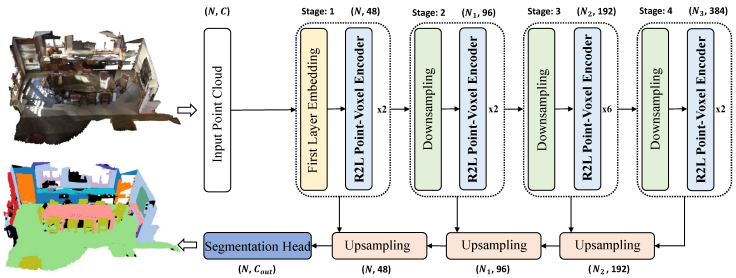

Figure 1. Network structure of the proposed RegionPVT. R2L Point-Voxel Encoder represents the proposed Regional-to-Local Point-Voxel Transformer Encoder. An encoder–decoder architecture is employed, comprising multiple stages connected via downsampling layers to learn hierarchical multi-scale features in a progressive manner. The numbers of point clouds and feature dimensions for each stage are provided on the top and below of the model.

2. Semantic Segmentation on Point Clouds

In the realm of 3D semantic segmentation on point clouds, methods can be divided into three predominant paradigms: voxel-based approaches [18][19][32], point-based techniques [8][9][20][38][39][40], and hybrid methodologies [41][42][43][44]. Voxel-based strategies strive to transform the inherently irregular structure of point clouds into a structured 3D voxel grid, leveraging the computational strengths of 3D CNNs. To enhance voxel efficiency, notable frameworks such as OctNet [16], O-CNN [17], and kd-Net [45] shift their focus to tree structures for non-empty voxels. Meanwhile, SparseConvNet [18] and MinkowskiNet [19] promote the use of discrete sparse tensors, making it easier to create efficient, fully sparse convolutional networks designed for fast voxel processing. However, the granularity of voxel-based methods, constricted by resolution constraints, occasionally sacrifices minute geometric details during the voxelization phase. On the other hand, point-based methods aim to create advanced neural networks that can process raw point clouds. Leading the way in this field, PointNet [8] pioneered the approach of using raw point clouds as clean inputs for neural networks. This was followed by a series of creative efforts [9][20][38][40] that focused on using hierarchical local structures and incorporating valuable semantic features through complex feature combination methods. While these techniques are excellent at

capturing detailed local structures and avoiding issues related to quantization, they come with significant computational costs, especially for large-scale situations. Connecting the two approaches, hybrid techniques cleverly combine both point-based and voxel-based features. By combining the advantages of both approaches, they use the precise details provided by point clouds and the broader context provided by voxel structures. For instance, frameworks like PVCNN [41] and DeepFusionNet [43] smoothly blend layers from both approaches, cleverly avoiding any potential issues that could arise from voxelization.

3. Vision Transformers

Recently, the Transformer architecture, initially designed for natural language processing, has established itself as a significant player in the computer vision field, demonstrating compelling results. The groundbreaking Vision Transformer (ViT) [26] is proof that using a transformer encoder for image classification can work, competing with traditional Convolutional Neural Networks (CNNs) in terms of performance, especially when provided with plenty of data. Inspired by ViT's discoveries, a series of innovations [13][14][15][26][27][28] began journeys to improve and enhance vision transformer designs. For example, when dealing with the subtle difficulties of tasks like semantic segmentation and object detection that require detailed predictions, Pyramid Vision Transformer (PVT) [13] uses a pyramid structure, aiming to extract hierarchical features while also including spatial reduction attention, which helps reduce the computational load. Battling the inherent quadratic complexity characterizing global attention's computation and memory footprints, the Swin Transformer [14] introduces a partitioned, non-overlapping window-based local attention, further bolstered by a shifted window strategy, fostering inter-window feature exchanges. Expanding the range of perception, the Focal Transformer [15] introduces "focal attention", a skillful mechanism skilled at blending detailed local features with broader global interactions. Adding another layer of sophistication, RegionViT [46] infuses global insights directly into localized tokens via a regional-to-local attention mechanism.

4. Transformer on Point Cloud Analysis

In recent years, the Transformer approach has made a lasting impact on a wide range of point cloud analysis tasks, demonstrating its strength in tasks like semantic segmentation [30][31][32][47], object detection [36][48][49], and registration [50]. In the domain of 3D semantic segmentation, the Point Transformer [30] extends the original PointNet architecture [8]. It cleverly divides point cloud data into smaller groups and performs vector attention computations within these groups. On the other hand, the Fast Point Transformer [32] provides an efficient self-attention mechanism that can incorporate 3D voxel information while reducing computational complexity. On a similar trajectory, the Stratified Transformer [47] computes self-attention within small cubic areas, utilizing a layered key-sampling technique along with a modified window framework. However, even though they have made significant progress in understanding point clouds, these Transformer-based approaches struggle with the inherent computational challenges of self-attention, which grows quadratically. This computational bottleneck often confines their explorations to localized interactions with circumscribed receptive fields, thus leading to an unintended neglect of complex scene structures and the important details of multi-scale features.

References

1. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst. 2020, 22, 1341–1360.
2. Ando, A.; Gidaris, S.; Bursuc, A.; Puy, G.; Boulch, A.; Marlet, R. RangeViT: Towards Vision Transformers for 3D Semantic Segmentation in Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 5240–5250.
3. Alonso, I.; Riazuelo, L.; Montesano, L.; Murillo, A.C. 3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation. IEEE Robot. Autom. Lett. 2020, 5, 5432–5439.
4. Wolf, D.; Prankl, J.; Vincze, M. Enhancing semantic segmentation for robotics: The power of 3-d entangled forests. IEEE Robot. Autom. Lett. 2015, 1, 49–56.
5. Ishikawa, Y.; Hachiuma, R.; Ienaga, N.; Kuno, W.; Sugiura, Y.; Saito, H. Semantic segmentation of 3D point cloud to virtually manipulate real living space. In Proceedings of the 2019 12th Asia Pacific Workshop on Mixed and Augmented Reality (APMAR), Nara, Japan, 23–27 March 2019; pp. 1–7.
6. Yue, X.; Wu, B.; Seshia, S.A.; Keutzer, K.; Sangiovanni-Vincentelli, A.L. A lidar point cloud generator: From a virtual world to autonomous driving. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval,

Yokohama, Japan, 11–14 June 2018; pp. 458–464.

7. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep learning for 3d point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 4338–4364.
8. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
9. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5105–5114.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
12. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 432–448.
13. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 548–558.
14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical Vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
15. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal attention for long-range interactions in Vision transformers. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 30008–30022.
16. Riegler, G.; Osman Ulusoy, A.; Geiger, A. OctNet: Learning deep 3d representations at high resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3577–3586.
17. Wang, P.S.; Liu, Y.; Guo, Y.X.; Sun, C.Y.; Tong, X. O-CNN: Octree-based convolutional neural networks for 3d shape analysis. ACM Trans. Graph. (TOG) 2017, 36, 1–11.
18. Graham, B.; Engelcke, M.; Van Der Maaten, L. 3d semantic segmentation with submanifold sparse convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9224–9232.
19. Choy, C.; Gwak, J.; Savarese, S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3075–3084.
20. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on x-transformed points. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 820–830.
21. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-Net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11105–11114.
22. Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; Ghanem, B. PointNeXt: Revisiting pointnet++ with improved training and scaling strategies. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 23192–23204.
23. Li, Y.; Lin, Q.; Zhang, Z.; Zhang, L.; Chen, D.; Shuang, F. MFNet: Multi-level feature extraction and fusion network for large-scale point cloud classification. Remote. Sens. 2022, 14, 5707.
24. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.

26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.

27. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in Vision transformers. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 9355–9366.

28. Li, Y.; Wu, C.Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; Feichtenhofer, C. Improved multiscale Vision transformers for classification and detection. arXiv 2021, arXiv:2112.01526.

29. Engel, N.; Belagiannis, V.; Dietmayer, K. Point transformer. IEEE Access 2021, 9, 134826–134840.

30. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 16259–16268.

31. Guo, M.H.; Cai, J.X.; Liu, Z.N.; Mu, T.J.; Martin, R.R.; Hu, S.M. PCT: Point Cloud Transformer. Comput. Vis. Media 2021, 7, 187–199.

32. Park, C.; Jeong, Y.; Cho, M.; Park, J. Fast point transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 16949–16958.

33. Mazur, K.; Lempitsky, V. Cloud transformers: A universal approach to point cloud processing tasks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10715–10724.

34. Jiang, C.; Peng, Y.; Tang, X.; Li, C.; Li, T. PointSwin: Modeling Self-Attention with Shifted Window on Point Cloud. Appl. Sci. 2022, 12, 12616.

35. Zhang, C.; Wan, H.; Shen, X.; Wu, Z. Patchformer: An efficient point transformer with patch attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 11799–11808.

36. Dong, S.; Wang, H.; Xu, T.; Xu, X.; Wang, J.; Bian, Z.; Wang, Y.; Li, J. MsSVT: Mixed-scale sparse voxel transformer for 3d object detection on point clouds. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 11615–11628.

37. Pagh, R.; Rodler, F.F. Cuckoo hashing. J. Algorithms 2004, 51, 122–144.

38. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6411–6420.

39. Zhao, H.; Jiang, L.; Fu, C.W.; Jia, J. PointWeb: Enhancing local neighborhood features for point cloud processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5565–5573.

40. Xu, M.; Ding, R.; Zhao, H.; Qi, X. PAConv: Position adaptive convolution with dynamic kernel assembling on point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20-25 June 2021; pp. 3172–3181.

41. Liu, Z.; Tang, H.; Lin, Y.; Han, S. Point-voxel cnn for efficient 3d deep learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 965–975.

42. Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; Han, S. Searching efficient 3d architectures with sparse point-voxel convolution. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 685–702.

43. Zhang, F.; Fang, J.; Wah, B.; Torr, P. Deep fusionnet for point cloud semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 644–663.

44. Zhang, C.; Wan, H.; Shen, X.; Wu, Z. PVT: Point-voxel transformer for point cloud learning. Int. J. Intell. Syst. 2022, 37, 11985–12008.

45. Klokov, R.; Lempitsky, V. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 863–872.

46. Chen, C.F.; Panda, R.; Fan, Q. RegionViT: Regional-to-local attention for Vision transformers. arXiv 2021, arXiv:2106.02689.

47. Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; Jia, J. Stratified transformer for 3d point cloud segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New

Orleans, LA, USA, 19–24 June 2022; pp. 8500–8509.

48. Mao, J.; Xue, Y.; Niu, M.; Bai, H.; Feng, J.; Liang, X.; Xu, H.; Xu, C. Voxel transformer for 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3164–3173.
49. He, C.; Li, R.; Li, S.; Zhang, L. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 8417–8427.
50. Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; Xu, K. Geometric transformer for fast and robust point cloud registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 11143–11152.