Region-Transformer: Self-Attention Region Based Class-Agnostic Point Cloud Segmentation

Dipesh Gyawali, Jian Zhang and Bijaya B. Karki
School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA 70803, U.S.A.

Keywords: 3D Vision, Class-Agnostic Segmentation, Self-Attention, Point Cloud, Region-Growth.

Abstract: Point cloud segmentation, which helps us understand the environment of specific structures and objects, can be performed in class-specific and class-agnostic ways. We propose a novel region-based transformer model called Region-Transformer for performing class-agnostic point cloud segmentation. The model utilizes a region-growth approach and self-attention mechanism to iteratively expand or contract a region by adding or removing points. It is trained on simulated point clouds with instance labels only, avoiding semantic labels. Attention-based networks have succeeded in many previous methods of performing point cloud segmentation. However, a region-growth approach with attention-based networks has yet to be used to explore its performance gain. To our knowledge, we are the first to use a self-attention mechanism in a region-growth approach. With the introduction of self-attention to region-growth that can utilize local contextual information of neighborhood points, our experiments demonstrate that the Region-Transformer model outperforms previous class-agnostic and class-specific methods on indoor datasets regarding clustering metrics. The model generalizes well to large-scale scenes. Key advantages include capturing long-range dependencies through self-attention, avoiding the need for semantic labels during training, and applicability to a variable number of objects. The Region-Transformer model represents a promising approach for flexible point cloud segmentation with applications in robotics, digital twinning, and autonomous vehicles.

1 INTRODUCTION

Point cloud segmentation is an imperative technique to understand 3D surroundings and objects, with applications in robotics (Ling et al., 2021), automation (Chen et al., 2021b), digital twinning (Mirzaei et al., 2022), VR/AR (Placitelli and Gallo, 2011). Most existing methods perform class-specific segmentation (Qi et al., 2017a) (Qi et al., 2017b) (Yang et al., 2019) (Zhao et al., 2021a) requiring semantic labels. However, class-agnostic segmentation without prior object knowledge is more flexible yet challenging.

Recently, self-attention networks (Zhao et al., 2021a) have shown promise for point cloud tasks by capturing contextual information. And region-growth approaches enable adaptive segmentation determination. However, self-attention has not been explored to enhance region-based segmentation. Our key insight is combining self-attention with region-growth can improve class-agnostic point cloud segmentation.

We propose a Region-Transformer model, which utilizes self-attention in local neighborhoods to iteratively expand/contract segments by adding/removing points likely belonging to the same instance. This model provides two key advantages over previous methods: 1) Attention on local regions captures finer relationships versus global context, and 2) Region growth allows flexible segmentation boundaries using neighborhood information. Our experimental studies show that Region-Transformer significantly outperforms previous class-specific and class-agnostic methods and thus demonstrate the benefits of our proposed approach.

In this work, our main contribution includes the following.

- We leverage the power of the self-attention mechanism combined with the region-growing approach to segment an environment ranging from small-scale to large-scale data completely.
- We don't need semantic labels to train the model that provides flexibility in segmenting any number of objects in an environment.
- We capture local contextual information for each point inside a region that helps identify long-range point cloud data dependencies.

Gyawali, D., Zhang, J. and Karki, B.
Region-Transformer: Self-Attention Region Based Class-Agnostic Point Cloud Segmentation.
DOI: 10.5220/0012424500003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 341-348
ISBN: 978-989-758-679-8; ISSN: 2184-4321

2 RELATED WORKS

Research in point cloud segmentation has primarily focused on semantic and instance segmentation. Semantic segmentation classifies each point within 3D data, while instance segmentation assigns points to a specific instance without class labels. Class-specific segmentation has been more extensively studied than challenging class-agnostic segmentation in varied real-world environments.

Point cloud data features range from XYZ positions to geometric aspects like normals and curvatures. Techniques for analyzing these features include patch stitching and octree-based hierarchical representations (Gumhold et al., 2001) (Guo et al., 2015) (Wang and Yuan, 2010) (Zhou et al., 2021) (Zhao et al., 2019). (Nguyen and Le, 2013) discuss datasets and methodologies for point cloud segmentation.

Deep learning has emerged as a significant method for 3D point cloud segmentation, with approaches including projection-based, discretization-based, point-based, and proposal-based methods (Ahn et al., 2022) (Qi et al., 2017a) (Guo et al., 2020). (Guo et al., 2020) further explore neural networks for 3D tracking, shape classification, detection, and segmentation.

Few-shot learning, neighborhood information, and class-agnostic approaches are also being explored for segmentation (Zhao et al., 2021b) (Engelmann et al., 2019) (Nunes et al., 2022) (Sharma et al., 2020). Unsupervised methods and region-growing approaches address segmentation without labels, focusing on features and iterative calculations (Xiao et al., 2023) (Kang et al., 2020) (Chen et al., 2021a).

Recently, transformers have been applied to point cloud data, leveraging their success in NLP and Computer Vision (Gyawali, 2023) (Zhao et al., 2021a). Our research combines self-attention mechanisms with region-growing approaches for class-agnostic segmentation, utilizing the transformer architecture's adaptability to varying input data. Consequently, applying self-attention operations to 3D data is a logical choice, given that point clouds are collections of points within 3D space.

3 METHODOLOGY

The methodology includes problem formulation, point transformer block, interaction of self-attention and region growth, data preparation and inference.

3.1 Problem Definition and Formulation

We formulate point cloud segmentation as an iterative region-growing problem using a learned neural network function f. Given a point cloud $\mathbf{P}$ with N points represented by F features, the goal is to assign an instance label $\mathbf{L}$ to every point. The region-growth starts from a seed point $p_{\text{seed}} \in \mathbf{P}$ and progressively adds points $P^* \subset \mathbf{P}$ belonging to the same instance to expand the region. At each step, f transforms the input points C_{k} to output points $C_{\text{k+1}}$. Initially, $C_0 = p_{seed}$, until $C \rightarrow P^*$, indicating the full instance is segmented. The point cloud has 13 features, including XYZ positions, RGB colors, normals, and curvatures. Normals and curvatures are computed using PCA (Asao and Ike, 2022) on local neighborhoods. Room dimensions also normalize XYZ coordinates. In total, each point is represented by a 13D feature vector.

3.2 Network Architecture

The core component of the network architecture is the point transformer (Zhao et al., 2021a), a neural network designed to capture both local and global contextual information of each point, considering its neighboring points. This information is crucial in determining whether neighboring points should be included or excluded in the segmentation process. As shown in Figure 1, the network consists of two branches - an inlier branch and a neighbor branch - which receive inputs of inlier and neighbor point sets. The sets pass through encoder blocks B1 and B2 to generate latent feature vectors per set. The concatenated vectors are broadcasted and decoded by B3 (Chen et al., 2021a).

The Point Transformer block facilitates the exchange of information between localized feature vectors, allowing adaptation to spatial arrangements and features in 3D space. The core Point Transformer layer utilizes a self-attention mechanism to relate each point to its local neighborhood points, as shown in Figure 2. This captures contextual information to determine whether to include or exclude points during segmentation.

The Point Transformer layer captures each point's local and global contextual information by considering its neighboring points. For this, the layer utilizes a self-attention mechanism. Specifically, self-attention is applied locally within a predefined neighborhood (e.g., k-nearest neighbors) around each point. This allows focusing on and aggregating features from the most related subset of neighbors(16) rather than all points. The self-attention procedure uses map-

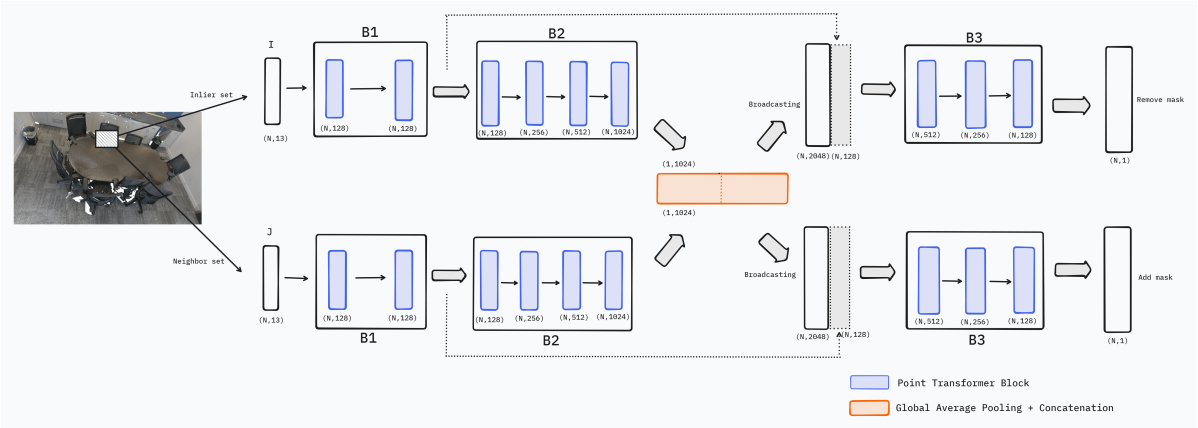

Figure 1: Region-Transformer network architecture for class-agnostic segmentation. Block B1 generating (128,128) features and B2 generating (128,256,512,1024) features act as encoders for inlier and neighbor sets. Block B3 generating (512,256,128) features acts as a decoder. Points from B1 and B2 are average-pooled, and inlier and neighbor sets are concatenated together to form a bottleneck. The encoded features are broadcasted into N number of points, and features from B1 output are concatenated to broadcasted features to get positional information of each point. In the last layer, the add and remove mask predictions are made.

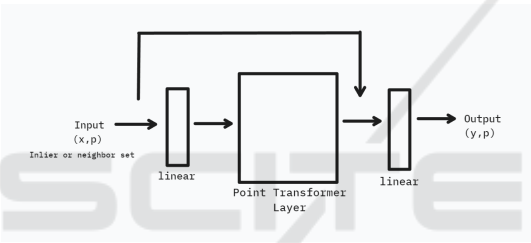

Figure 2: Point Transformer Block.

per functions to transform input point features into queries, keys, and values (Vaswani et al., 2023). Attention weights between the query and keys are calculated. These attention weights determine how much each value vector contributes to the output aggregated feature for the central point. In addition, a relative positional encoding δ is applied to retain positional information of each point in the 3D space. This encoding uses a parameterized function of the difference between point coordinates given as

$$\delta = \beta(p_i - p_j) \tag{1}$$

where p_i and p_j represent 3D position values for points i and j, and β represents three fully connected layers with two nonlinear ReLU in between. The point transformer is based on self vector-attention (Zhao et al., 2020) given as

$$y_i = \sum_{x_j \in X(i)} \rho(\gamma(\varphi(x_i) - \psi(x_j) + \delta)) \odot (\alpha(x_j) + \delta) \tag{2}$$

The self-attention procedure is defined in Equation 2, where φ, ψ, α transform input features, ρ normalizes, δ encodes position, and γ maps attention weights to aggregate features. $X(i) \subset X$ represents the local neighborhood of point $\mathbf{x_i}$. Self-attention is applied to each point's neighborhood to focus on similar local regions rather than global contexts. The mapper γ uses multilayer perceptrons to generate attention weights.

The output is a new feature vector for each point with selectively aggregated contextual information from its local neighborhood. These features can better determine relationships between nearby points to aid the region segmentation process. The ability of self-attention to capture dependencies based on feature similarity rather than spatial proximity helps the region grow according to semantic instance boundaries. The localization also allows finer segmentation precision.

3.3 Self-Attention in Region-Growth

Our method includes encoder-decoder network which is trained to learn effective region-growth with a binary cross-entropy loss on the addition and removal predictions. The loss compares predicted point inclusion/exclusion probabilities with ground truth labels given as

$$\begin{aligned} \mathcal{L} = &-\frac{1}{I}\sum_{1}^{I}[x_i \log \hat{x_i} + (1 - x_i)\log(1 - \hat{x_i})] \\ &-\frac{1}{J}\sum_{1}^{J}[y_j \log \hat{y_j} + (1 - y_j)\log(1 - \hat{y_j})] \end{aligned} \tag{3}$$

Our key insight enabling improved segmentation performance is using self-attention within the context

of iterative, neural network-guided region-growth. Local point neighborhoods are defined around seed points using a radius threshold. Self-attention is applied to every neighborhood, enabling each point to aggregate features from its local context. It captures nuanced geometric relationships between a point and its neighbors that standard features miss. This higher-order feature representation is input to the Point Transformer network to predict iterative growth decisions. So self-attention does the "heavy lifting" to equip the network with finer-grained neighborhood characterization for superior growth predictions. The Point Transformer network analyzes attention-enhanced local features to predict binary masks, indicating which neighborhood points should be added/removed to grow the region. Based on these iterative add/remove decisions, new local neighborhoods are extracted around the grown region's updated seed points. Self-attention and neural feature processing are repeated in the new neighborhoods, further evolving the regions to capture more points belonging to the same instance.

The key novelty is using the neural predictions from intermediate attention-augmented features to actively determine how regions evolve rather than relying solely on hand-crafted similarity metrics. This dynamic interaction helps address limitations of both attention and region growth in isolation.

3.4 Data Preparation and Simulation

The S3DIS (Armeni et al., 2017) and Scannet (Dai et al., 2017) datasets containing point cloud labels generate simulated training data. The simulation follows a region-growth approach based on (Chen et al., 2021a) with PyTorch implementation. Data augmentation is applied, including random flipping, rotation, and introducing mistake probability(θ) noise. 844 million point clouds from 3.5 million instances are generated as S3DIS training data, in addition to 17 million validation sets. Similarly, 741 million point clouds from 5.0 million instances are generated as Scannet training sets. The simulation grows regions starting from random seed points, iteratively aggregating nearby points sharing the same instance label. Noise is gradually reduced over the region growth iterations to promote convergence while preventing overfitting. This process creates training data mimicking realistic instance segmentations. The validation data evaluates generalization. Each instance is unique despite identical class labels. The approach synthesizes sufficiently large and diverse labeled data for effectively learning the region growth transformations.

3.5 Inference

The inference technique and conditions are derived from (Chen et al., 2021a). During inference, segmentation is performed by iteratively adding and removing points from an initial seed region until all points are labeled. The trained transformer network outputs addition and removal predictions to grow regions. The process continues until one termination criteria:

- No neighboring points are remaining to be assigned to the region.
- The points set to be added are empty.
- There is no expansion of region for two consecutive steps.

On termination, the final region is assigned an instance label and then reset with a new seed. Seeds are strategically selected as the point with the lowest curvature for consistency (Dimitrov and Golparvar-Fard, 2015). For robustness, segments with few points (less than 8) do not form new instances. Instead, points adopt the label of nearest neighbors. This prevents the loss of points between larger segmented instances (Hu et al., 2019) (Xie et al., 2021).

In summary, inference progressively segments the point cloud into instance-labeled regions by learned prediction of what points to add/remove at each iteration. Termination and seed selection strategies maximize completeness, consistency, and efficiency.

4 EXPERIMENTS AND RESULTS

The Region-Transformer model was tested for segmentation on indoor (S3DIS, Scannet) and outdoor scenes, using clustering metrics (ARI, AMI, NMI) and general metrics (mean IOU, Precision, Recall). Comparisons were made with class-agnostic and class-specific segmentation methods. Implemented in Pytorch, the model used Adam Optimizer, which was trained over 90 epochs on NVIDIA RTX A6000 GPU, and the training took about seven days.

The study conducted extensive experiments to evaluate the Region-Transformer, comparing it against previous segmentation methods using the S3DIS and ScanNet datasets. These datasets represent different environments: S3DIS focuses on office settings, while ScanNet covers home environments. This diverse testing revealed that the Region-Transformer excelled over other methods across almost every metric, as shown in Tables 1 and 2.

A significant aspect of this success is attributed to the model's use of local neighborhood informa-

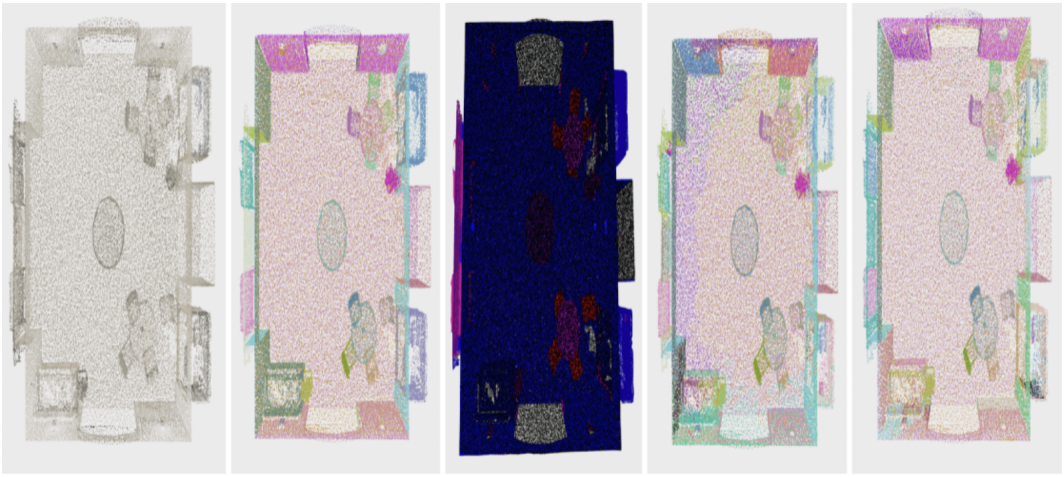

Figure 3: Object-agnostic segmentation results across Scannet a) Raw point cloud original visualization (b) Ground truth original segmentation (c) PointNet++ segmentation (d) LRGNet segmentation (e) Region-Transformer (Our Method).

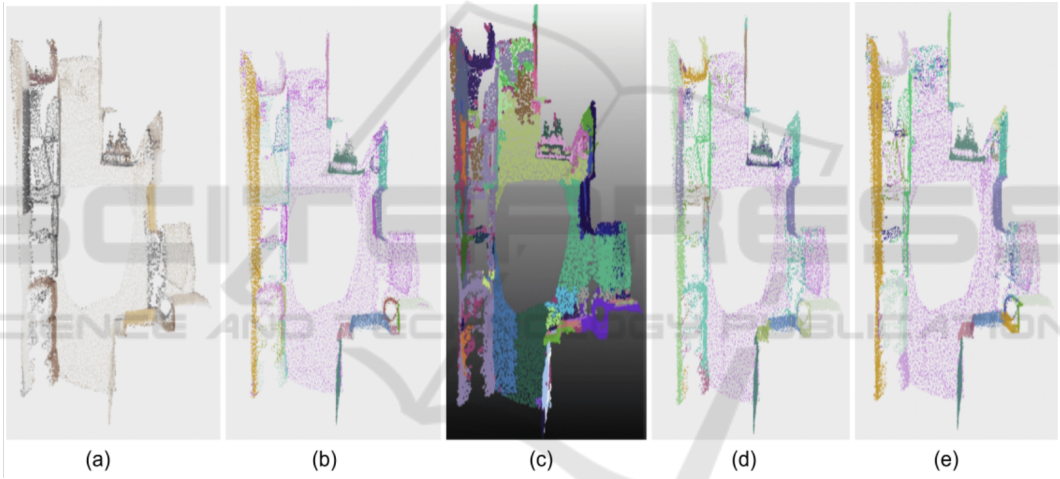

Figure 4: Object-agnostic segmentation results across S3DIS a)Raw point cloud original visualization (b) Ground-truth original segmentation(c) PointNet++ segmentation (d) LRGNet segmentation (e) Region-Transformer (Our Method).

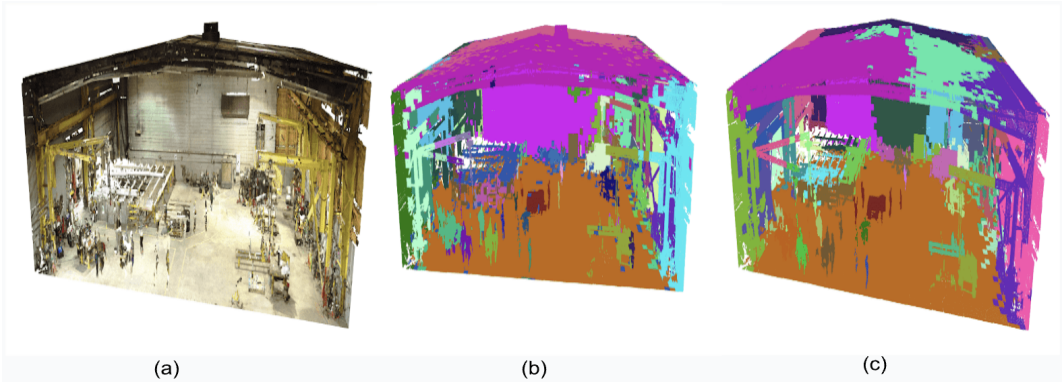

Figure 5: Object-agnostic Segmentation results on real-world, large-scale factory data a) Raw point cloud original visualization (b) LRGNet segmentation (c) Region-Transformer (Our Method).

Table 1: Comparison between models using Scannet as training and S3DIS as test.

Type	Method	ARI	AMI	NMI	Precision	Recall	mIoU
class-dependent	PointNet (Qi et al., 2017a)	0.38	0.48	0.58	0.18	0.17	0.25
	PointNet++ (Qi et al., 2017b)	0.40	0.56	0.62	0.15	0.22	0.31
	3D-BoNet (Yang et al., 2019)	0.68	0.72	0.75	0.20	0.29	0.35
	JSIS3D (Pham et al., 2019)	0.63	0.73	0.74	0.28	0.29	0.36
	Point Transformer (Zhao et al., 2021a)	0.69	0.73	0.75	0.44	0.44	0.47
class-independent	FPFH (Rusu et al., 2009)	0.39	0.60	0.62	0.14	0.25	0.32
	Region growing	0.59	0.70	0.71	0.19	0.34	0.38
	Rabbani et.al (Rabbani et al., 2006)	0.62	0.71	0.72	0.17	0.31	0.36
	LRGNet (Chen et al., 2021a)	0.67	0.74	0.75	0.25	0.41	0.43
	LRGNet+ local search (Chen et al., 2021a)	0.68	0.75	0.76	0.34	0.44	0.45
	Region-Transformer	**0.79**	**0.85**	**0.86**	**0.63**	**0.66**	**0.62**

Table 2: Comparison between models using S3DIS as training and Scannet as test.

Type	Method	ARI	AMI	NMI	Precision	Recall	mIoU
class-dependent	PointNet (Qi et al., 2017a)	0.40	0.51	0.57	0.08	0.13	0.26
	PointNet++ (Qi et al., 2017b)	0.47	0.57	0.63	0.15	0.21	0.32
	3D-BoNet (Yang et al., 2019)	0.34	0.54	0.59	0.10	0.13	0.24
	JSIS3D (Pham et al., 2019)	0.31	0.56	0.57	0.15	0.13	0.22
	Point Transformer (Zhao et al., 2021a)	0.56	0.69	0.70	**0.33**	0.34	0.38
class-independent	FPFH (Rusu et al., 2009)	0.28	0.51	0.53	0.10	0.14	0.26
	Region growing	0.44	0.60	0.62	0.17	0.23	0.30
	Rabbani et.al (Rabbani et al., 2006)	0.49	0.62	0.64	0.13	0.24	0.32
	LRGNet (Chen et al., 2021a)	0.54	0.67	0.69	0.25	0.33	0.39
	LRGNet+ local search (Chen et al., 2021a)	0.56	0.68	0.69	0.31	0.33	0.38
	Region-Transformer	**0.61**	**0.70**	**0.72**	0.25	**0.39**	**0.43**

tion combined with a self-attention mechanism. This approach was particularly effective compared to local search techniques like those in LRGNet (Chen et al., 2021a). The research underscores the advantage of applying the self-attention mechanism in a region-based, class-agnostic approach for point cloud segmentation. Unlike methods trained with semantic label information that showed diminished performance when applied to a different dataset, the Region-Transformer demonstrated robust generalization capabilities.

Regarding specific evaluation metrics, the Normalized Mutual Information(NMI) metric assesses the similarity between two clusters, with a value range of 0 to 1. A high NMI score for the Region-Transformer indicates a reduction in the entropy of instance labels and an improvement in the under-segmentation of instance labels. The method's high NMI is attributed to its ability to predict pure clusters that closely match the ground truth. Similar to NMI but adjusted for random chance, AMI ranges from -1 to 1. The Region-Transformer scored high on AMI, signifying its efficiency in creating pure clusters and solving over-segmentation problems. AMI accounts for the number of clusters and dataset size, discounting chance normalization. From -1 to 1, ARI is related to accuracy in measuring the percentage of correct predictions. It corrects for a change from the rand index and is particularly useful for considering unbalanced clustering.

The paper emphasizes that while NMI and AMI are effective for evaluating clustering purity and similarity, they have limitations. For instance, NMI can increase with the number of clusters regardless of actual mutual information. Similarly, AMI might be biased towards unbalanced clustering solutions. Hence, including ARI as a metric provides a more comprehensive and balanced evaluation.

In the qualitative evaluation, the Region-Transformer demonstrates marked improvements in segmenting indoor scenes, as shown in Figures 3 and 4. It effectively resolves under-segmentation on smooth surfaces like floors and accurately distinguishes objects of varying shapes and sizes. However, challenges arise in wall segmentation due to uneven surfaces and corner over-segmentation. Despite these issues, its performance in differentiating unique instances in environments like the S3DIS indoor scenes is notable.

The model's adaptability to large-scale scenes is

Table 3: Computation Time analysis (seconds).

Method	Minimum	Average	Maximum
Region growing	0.4	4.8	18.6
PointNet (Qi et al., 2017a)	0.1	0.6	2.5
PointNet++ (Qi et al., 2017b)	0.1	0.9	3.5
Rabbani et.al (Rabbani et al., 2006)	0.3	4.6	18.3
3D-BoNet (Yang et al., 2019)	1.5	14.1	69.3
FPFH (Rusu et al., 2009)	0.5	4.6	17.8
LRGNet (Chen et al., 2021a)	0.8	64.9	620.9
JSIS3D (Pham et al., 2019)	1.0	539.2	16713.9
Region-Transformer	1.5	57.4	311.5

a key strength. Initially trained on homes and offices, the Region-Transformer shows remarkable capability in segmenting larger environments, including factories and large buildings, as illustrated in Figure 5. This is essential for applications in self-driving cars and digital twinning, demonstrating its practical utility in handling complex, large-scale environments without prior object knowledge in general.

Furthermore, the Region-Transformer significantly improves computational efficiency, particularly in inference time, compared to other class-agnostic, region-based segmentation approaches. Despite the iterative nature of its process, it maintains better efficiency, a finding supported by the average inference time analysis of 50 S3DIS datasets presented in Table 3. This balance of accuracy and computational speed makes it suitable for real-time applications, highlighting its potential in accuracy and efficiency.

5 CONCLUSIONS

We propose a novel region-based transformer model called Region-Transformer for performing class-agnostic point cloud segmentation. Experiments demonstrate that combining self-attention with an iterative region-growing approach significantly improves segmentation performance. Specifically, attention mechanisms effectively capture local contextual relationships between points missed by previous region growth methods. Key advantages of the proposed approach include:

- Attention on local point neighborhoods enables capturing finer feature relationships versus global context. This aids in precisely determining segmentation boundaries.
- The region growth formulation allows flexible, adaptive segmentation based on progressively learned point neighborhood relationships, avoiding strong assumptions.
- The method avoids dependence on semantic class labels, enabling new object segmentation.

The promising performance and flexibility of Region-Transformer represent an important step toward practical point cloud segmentation without prior knowledge. Potential real-world applications span robotic perception, autonomous navigation, digital twinning, and augmented reality.

Future avenues for improving Region-Transformer include reducing training and inference times via model compression techniques tailored for transformers. New spatial attention operators could also be designed to capture geometric relationships in point clouds. An exciting research direction involves extending the approach to perform video segmentation on dynamic point cloud sequences containing moving objects.

ACKNOWLEDGEMENTS

This project is supported in part by NSF grant OIA-1946231 and NASA. We are grateful to Mr. Marc Aubanel for his feedback and providing data for this research.

REFERENCES

Ahn, P., Yang, J., Yi, E., Lee, C., and Kim, J. (2022). Projection-based point convolution for efficient point cloud segmentation. *IEEE Access*, 10:15348–15358.

Armeni, I., Sax, A., Zamir, A. R., and Savarese, S. (2017). Joint 2d-3d-semantic data for indoor scene understanding. *ArXiv e-prints*.

Asao, Y. and Ike, Y. (2022). Curvature of point clouds through principal component analysis.

Chen, J., Kira, Z., and Cho, Y. K. (2021a). Lrgnet: Learnable region growing for class-agnostic point cloud segmentation. *IEEE Robotics and Automation Letters*, 6(2):2799–2806.

Chen, S., Liu, B., Feng, C., Vallespi-Gonzalez, C., and Wellington, C. (2021b). 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. *IEEE Signal Processing Magazine*, 38(1):68–86.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Dimitrov, A. and Golparvar-Fard, M. (2015). Segmentation of building point cloud models including detailed architectural/structural features and mep systems. *Automation in Construction*, 51(C):32–45. Publisher Copyright: © 2014 Elsevier B.V. All rights reserved.

Engelmann, F., Kontogianni, T., Schult, J., and Leibe, B. (2019). *Know what your neighbors do: 3d semantic segmentation of point clouds*, page 395–409. Springer International Publishing.

Gumhold, S., Wang, X., and Macleod, R. (2001). Feature extraction from point clouds. In *International Meshing Roundtable Conference*.

Guo, Y., Bennamoun, M., Sohel, F., Lu, M., Wan, J., and Kwok, N. (2015). A comprehensive performance evaluation of 3d local feature descriptors. *International Journal of Computer Vision*, 116.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., and Bennamoun, M. (2020). Deep learning for 3d point clouds: A survey.

Gyawali, D. (2023). Lrtransformer: Learn-region transformer for object-agnostic point cloud segmentation. Master's thesis, Louisiana State University.

Hu, P., Held, D., and Ramanan, D. (2019). Learning to optimally segment point clouds.

Kang, C. L., Wang, F., Zong, M. M., Cheng, Y., and Lu, T. N. (2020). Research on improved region growing point cloud algorithm. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3/W10:153–157.

Ling, C. F., Dang, S. W., Zhang, C. Y., and Chen, Y. (2021). Research and application of semantic point cloud on indoor robots. In *2021 5th International Conference on Communication and Information Systems (ICCIS)*, pages 108–113.

Mirzaei, K., Arashpour, M., Asadi, E., et al. (2022). Automatic generation of structural geometric digital twins from point clouds. *Sci Rep*, 12:22321.

Nguyen, A. and Le, B. (2013). 3d point cloud segmentation: A survey. In *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, page 225–230.

Nunes, L., Chen, X., Marcuzzi, R., Osep, A., Leal-Taixe, L., and Stachniss, C. (2022). 3d point cloud clustering with learnable robust geometric constraints.

Pham, Q.-H., Nguyen, D. T., Hua, B.-S., Roig, G., and Yeung, S.-K. (2019). Jsis3d: Joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields.

Placitelli, A. P. and Gallo, L. (2011). Low-cost augmented reality systems via 3d point cloud sensors. In *2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems*, pages 188–192.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space.

Rabbani, T., van den Heuvel, F., and Vosselman, G. (2006). Segmentation of point clouds using smoothness constraints. In Maas, H. and Schneider, D., editors, *ISPRS 2006 : Proceedings of the ISPRS commission V symposium*, volume 35, pages 248–253. International Society for Photogrammetry and Remote Sensing (ISPRS). ISPRS commission V symposium : image.

Rusu, R. B., Blodow, N., and Beetz, M. (2009). Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217.

Sharma, A., Khan, N., Sundaramoorthi, G., and Torr, P. (2020). Class-agnostic segmentation loss and its application to salient object detection and segmentation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Wang, L. and Yuan, B. (2010). Curvature and density based feature point detection for point cloud data. In *IET 3rd International Conference on Wireless, Mobile and Multimedia Networks (ICWMNN 2010)*, page 377–380.

Xiao, A., Huang, J., Guan, D., Zhang, X., Lu, S., and Shao, L. (2023). Unsupervised point cloud representation learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Xie, C., Xiang, Y., Mousavian, A., and Fox, D. (2021). Unseen object instance segmentation for robotic environments.

Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., and Trigoni, N. (2019). Learning object bounding boxes for 3d instance segmentation on point clouds.

Zhao, H., Jia, J., and Koltun, V. (2020). Exploring self-attention for image recognition.

Zhao, H., Jiang, L., Jia, J., Torr, P., and Koltun, V. (2021a). Point transformer.

Zhao, N., Chua, T.-S., and Lee, G. H. (2021b). Few-shot 3d point cloud semantic segmentation.

Zhao, R., Pang, M., Liu, C., and Zhang, Y. (2019). Robust normal estimation for 3d lidar point clouds in urban environments. *Sensors*, 19(5).

Zhou, J., Jin, W., Wang, M., Liu, X., Li, Z., and Liu, Z. (2021). Fast and accurate normal estimation for point cloud via patch stitching.