A TWO-STAGE APPROACH FOR RARE CLASS SEGMENTATION IN LARGE-SCALE URBAN POINT CLOUDS

X. Zhang [1], R. Xue [1,2]*, U. Soergel [1]

[1] Institute for Photogrammetry, University of Stuttgart, 70174 Stuttgart, Germany -
(xinlong.zhang, ruihang.xue, uwe.soergel) @ifp.uni-stuttgart.de
[2] National Lab of Radar Signal Processing, Xidian University, 710071 Xi'an, China - rhxue@stu.xidian.edu.cn

KEY WORDS: Deep Learning, Transformer, Two-Stage Approach, Rare Classes, Imbalanced Classes, Semantic Segmentation, Urban Point Clouds.

ABSTRACT:

Although deep learning has greatly improved the semantic segmentation accuracy of point clouds, the segmentation of rare classes in large-scale urban scenes has not been targeted in available methods. This paper proposes a two-stage segmentation framework with automated workflows for imbalanced rare classes based on general semantic segmentation. The proposed approach includes two stages: general semantic segmentation and object-based refined semantic segmentation. Firstly, general segmentation networks are utilized to segment general large objects. Secondly, refined semantic segmentation is conducted by an automated workflow: 3D clustering and bounding box (BBox) generation are applied to the point cloud of rare fine-grained objects during the training, followed by object detection to extract fine-grained objects. Afterwards, as the constraints, the extracted BBoxes further refine the segmentation results. Our approach is evaluated on the Hessigheim High-Resolution 3D Point Cloud (H3D) Benchmark and obtains state-of-the-art 89.35% overall accuracy and outstanding 75.70% mean F1-Score. Furthermore, rare classes Vehicle and Chimney achieve breakthroughs from zero to 63.63% and 52.00% in F1-score, respectively.

1. INTRODUCTION

Automated semantic segmentation of point clouds is fundamental for various fields of application, including autonomous driving, building information modeling and robotics (Chen et al., 2020). With the advances of recent technology in remote sensing sensors and platforms, especially lightweight LiDAR devices and unmanned aerial vehicles (UAV), which facilitate the availability of fine-grained 3D data. Such data, while revealing the spatial distribution of target objects in high detail, also bring about the problem of significant class imbalance. Efficient methods are needed to fully harness this unpreceded source of information for 3D semantic segmentation.

Established methods like VoxNet (Maturana et al., 2015) voxelizes point clouds to make the data structure suitable for 3D CNNs. But the sparsity of point clouds causes low efficiency of voxel grid arrangement. SSCNs (Graham et al., 2018) takes advantage of sparsity and considers only occupied voxels to improve the efficiency. Schmohl and Soergel (2019) apply it to the large-scale ALS point clouds. But such methods only depend on the voxel boundary and ignore the geometric structure of local regions. PointNet++ (Qi et al., 2017) effectively solves the problem of extracting local features by combining sampling-grouping layer and PointNet (Qi et al., 2017) layer. Nevertheless, features via the pooling operator in each individual dimension have the same weight. The self-attention operator in Point Transformer (Zhao et al., 2021) weights each element adaptively. However, when dealing with imbalanced rare classes in large-scale urban scenes, the above general semantic segmentation methods often fail to extract sufficiently effective semantic features of these classes and perform poorly. To alleviate this problem, surface features based on the local 3D neighborhood (Weinmann et al., 2013) can be utilized to strengthen the local perception of networks.

The goal of 3D object detection is to detect class-imbalanced high-value objects and indicate object location and size attributes in the form of 3D BBoxes. If objects have strong shape cues, detectors can easily locate objects and thus provide valuable information for semantic segmentation (Dong et al., 2014). In general, existing 3D detection methods can be broadly grouped into two categories, i.e., single-stage detection and two-stage detection. Single-stage detection methods regress 3D bounding box directly from extracted features, such as PointPillars (Lang et al, 2019) and 3DSSD (Yang et al., 2020). Two-stage detection methods like PointRCNN (Shi et al., 2019) and PV-RCNN (Shi et al., 2020) generate region-proposal-aligned features in the first stage, and refine predictions in the second stage. Single-stage detection methods usually run faster due to simpler network structures, whereas two-stage detection methods often attain higher precisions benefited from the second refined stage.

So far, those 3D segmentation methods developed in the computer vision community have mostly been used for general large classes in ground scans with limited space, or indoor scenes. To our knowledge, specialized segmentation methods for imbalanced rare classes in large-scale urban point clouds have not yet been investigated. In this paper, we unify object detection models into the framework of general semantic segmentation, and present a two-stage segmentation framework for imbalanced rare classes.

* Corresponding author

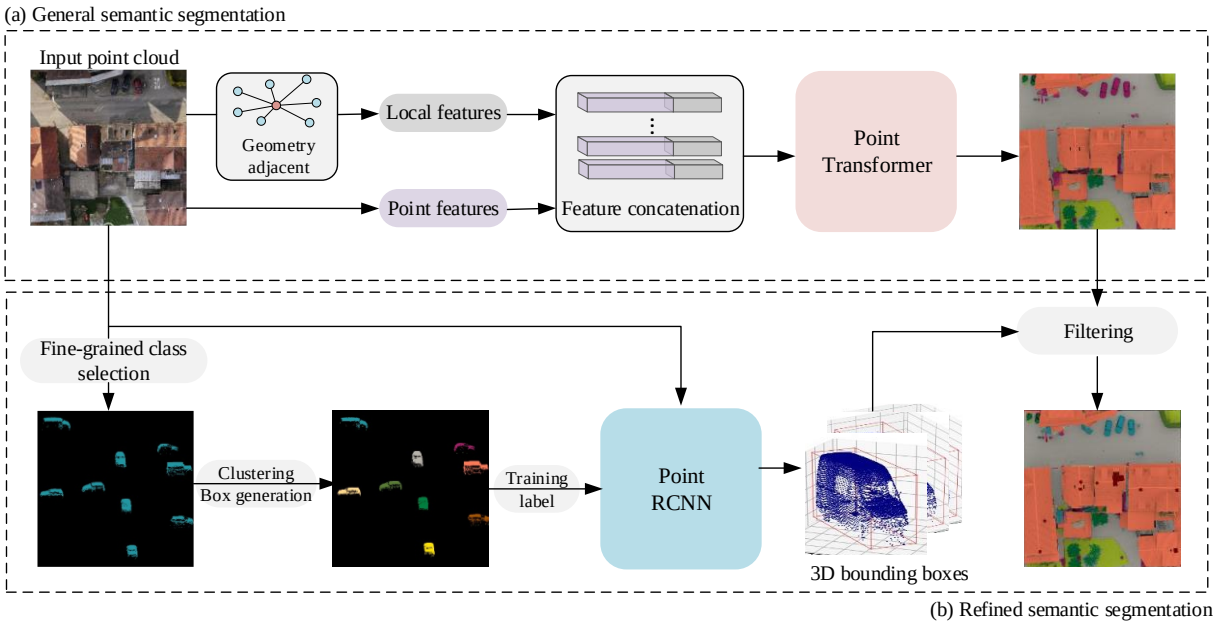

Figure 1. Structure of the proposed approach.

The rest of the paper is organized as follows. In Section 2, the overall structure of the proposed two-stage segmentation framework is introduced in detail. Section 3 shows the experimental details on the H3D dataset, and analyzes the results of general semantic segmentation and our two-stage segmentation. Section 4 is left for conclusion and outlook.

2. METHODOLOGY

In this section, we present our proposed two-stage segmentation framework for imbalanced rare classes. The overall structure is illustrated in Fig. 1, which consists of general semantic segmentation stage and refined semantic segmentation stage. General semantic segmentation is used to extract general large classes, based on which the refinement of imbalanced rare classes is performed in the second stage.

2.1 General Semantic Segmentation

Since Point Transformer is invariant to permutation of the input elements due to the inherent set-level operation of the self-attention structure, which is consistent with the disordered distribution of point clouds, it is quite natural to choose the network as the main component of general semantic segmentation stage. But unlike the original Point Transformer, not only point original features but also local surface features of points are fed into the network. In this way, the local perception of the network can be enhanced to a certain extent.

Local surface features provide the attributes of the local approximate surface of each point (Weinmann et al., 2013), which can be calculated based on the local 3D neighborhood. Only descriptors with strong semantic interpretation are selected to construct the local features of each point p, which are described by one 6-tuple

$$\mathrm{F}_p = (n_{px}, n_{py}, n_{pz}, d_p, f_p, r_p) \tag{1}$$

where n_{px}, n_{py}, n_{pz} are the parameters of the normal vector, d_p is the distance from the origin to the fitted plane of point p, f_p is change of curvature, and r_p is the residual from point p to its fitted surface. Normals distinguish flat and inclined surfaces, d_p is the association with global information, f_p represents the local surface variation and r_p describes the local roughness. This kind of geometric adjacency can enhance the local perception of the network.

As part and parcel of general semantic segmentation stage, point transformer layer is formed by two linear mappings and a self-attention calculation. The linear mapping converts the input-output dimension, and the self-attention estimates the internal relationship among the input points. The self-attention calculation of each point x_i is expressed by

$$y_i = \sum_{x_j \in X(i)} \alpha\left(\beta\left(\varphi_k(x_i) - \varphi_q(x_j) + pos\right)\right) \mathrm{e}\ \varphi_v(x_j + pos) \tag{2}$$

where y_i is the output feature vector, $x_j \in X(i)$ is the neighborhood of the point x_i, which is obtained by the KNN algorithm. α is the softmax activation function, and β is the attention mapping function, which is implemented by a multilayer perceptron (MLP), i.e. 2 linear layers and a ReLU (Glorot et al., 2011) activation function. φ_k, φ_q and φ_v are all linear mappings for adapting to different feature dimensions, and e denotes an elementwise multiplication. pos is the positional coding, which is a linear mapping from the relative coordinate of the points:

$$pos = \varphi_p\left(cor_i - cor_j\right) \tag{3}$$

where cor_i and cor_j are respectively the 3D coordinates of point i and point j, φ_p is a MLP.

The U-Net-like architecture (Ronneberger et al., 2015) is applied to connect point tranformer layers (Figure 2), which

consists of 4 encoder layers and 4 decoder layers. Transition down is implemented by the farthest point sampling and KNNs searching. Transition up is realized by trilinear interpolation. For the semantic segmentation task, a MLP maps the point feature to the label space y_k at the last layer. All the learnable parameters of the network could be updated by optimizing the cross-entropy loss function.

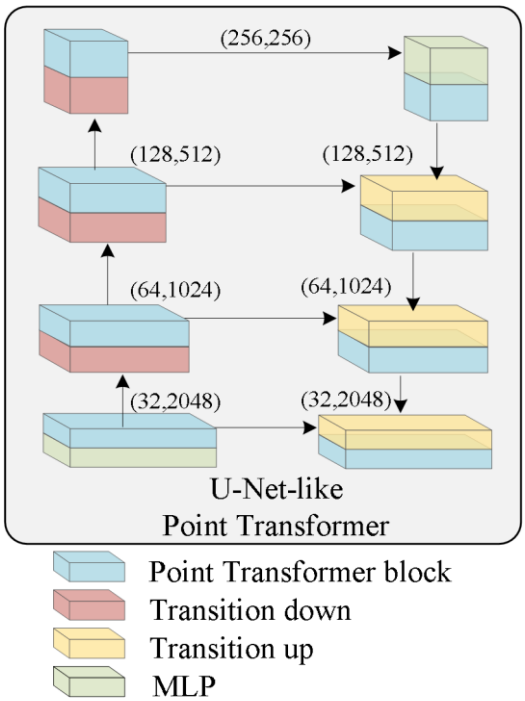

Figure 2. The network architecture of the general semantic segmentation block.

2.2 Refined Semantic Segmentation

In the stage of refined semantic segmentation, the training process and the inference process are separated. During training, automated label generation in the form of 3D BBox is essential to unify object detection models into the framework of semantic segmentation. Firstly, fine-grained rare classes are selected individually to avoid confusion with general classes. Then, considering their discrete distribution, density-based spatial clustering of applications with noise (DBSCAN) method (Ester et al., 1996) is utilized to divided the point cloud of rare classes into separate reliable clusters. Afterwards, the vertices of each cluster corresponding to the convex hull are calculated and adjusted to the vertices of 3D BBox. After the automated process above, the generated BBox labels of rare classes and the original point cloud can be fed into the object detection block.

Thanks to its satisfactory detection precision in large-scale complex scenes, PointRCNN is chosen as the detection block. The main components of the network are 3D proposal generation and 3D BBox refinement. 3D proposal generation performs the rough segmentation of foreground points, based on which 3D BBox proposals are constructed. PointNet++ with multi-scale grouping is utilized as the backbone network to learn discriminative point-wise features of the raw point clouds. In order alleviate the class imbalance problem between foreground points and background points, the focal loss (Lin et al., 2017) is chosen to update the network as follows:

$$FL(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t) \tag{4}$$

where p_t is the estimated probability for the class with ture label (foreground point), $\alpha_t (1-p_t)^{\gamma}$ is a self-adaptive modulating factor that not only balances the importance of positive/negative examples, but also differentiate between easy/hard examples. α_t controls the rate of change of the weighting factor $(1-p_t)^{\gamma}$, γ denotes focusing parameter that smoothly adjusts the rate at which simple examples are down-weighted.

In the stage of 3D BBox refinement, when the 3D intersection over union (IoU) between a ground-truth BBox and a BBox proposal is greater than 0.6, the point-wise features and associated features for each positive 3D proposal are fed to PointNet++ for refining the 3D Bbox locations as well as the foreground object confidence. All the learnable parameters could be updated by optimizing the following loss function:

$$Loss = \frac{1}{\beta}\sum_{i\in\beta} L_{cls}(prop_i, label_i) + \frac{1}{\beta_{pos}}\sum_{i\in\beta_{pos}} L_{reg}(prop_i^{pos}, label_i^{pos}) \tag{5}$$

where β is the set of 3D proposals and β_{pos} stores the positive proposals. $prop_i$ is the estimated confidence of the i_{th} proposal and $label_i$ represents the corresponding label, $prop_i^{pos}$ and $label_i^{pos}$ denotes the i_{th} positive 3D proposal and its BBox ground truth. The loss function L_{cls} can supervise the prediction confidence of foreground objects and L_{reg} is utilized to refine the BBox locations, which are the cross-entropy loss and the bin-based regression loss (Shi et al., 2019), respectively.

Finally, the 3D BBoxes of rare fine-grained objects are predicted in the inference stage and these high precision BBoxes are utilized as constraints for rare class segmentation.

3. EXPERIMENTS

3.1 Data Description

The experiments are based on the public H3D dataset (Kölle et al., 2021). The dataset was collected by a Riegl VUX-1LR Scanner and two oblique-looking Sony Alpha 6000 cameras

Split	H3D Classes [%]										
	Low Veg.	I. Surf	Vehicle	U. Furn.	Roof	Facade	Shrub	Tree	Soil	V. Surf	Chimney
Train	35.96	17.53	0.43	1.95	10.56	2.02	1.81	13.60	14.45	1.64	0.04
Validation	25.85	22.21	1.27	3.15	21.10	3.82	2.36	15.34	4.10	0.70	0.11

Table 1. Comparison of class occurrences in H3D dataset.

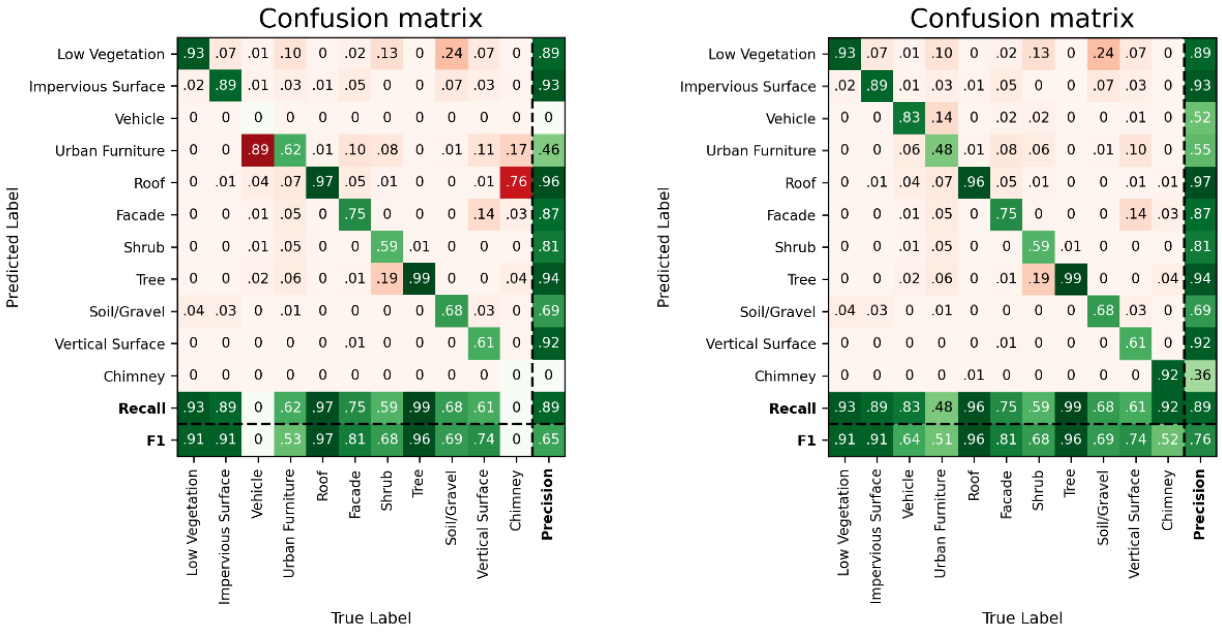

Figure 3. Confusion matrices on (a) the single-stage segmentation (b) our proposed two-stage approach.

integrated on a RIEGL UAV platform. The mean point density is 800 points/m² enriched by RGB colors and the ground sampling distance (GSD) of images is 2-3 cm. In addition, the points have been manually labelled with the following 11 classes: Low vegetation, Impervious surface, Vehicle, Urban furniture, Roof, Facade, Shrub, Tree, Soil/Gravel, Vertical surface, Chimney. However, this fine-grained class catalog leads to data imbalance.

Detailed statistics of class occurrences in H3D dataset is shown in Table 1, the most underrepresented classes are Vehicle and Chimney, which only occupy 0.43% and 0.04% in the training set, respectively. The significant data imbalance makes the semantic segmentation of rare classes a challenging task.

3.2 Implementation Details

Our implementation of the two-stage segmentation approach is realized on a NVIDIA RTX2080Ti GPU with the framework of Pytorch 1.0. According to the analysis in section 3.1, In order to reduce the computational burden, the training data and the test data are cropped into 49 splits and 22 splits, respectively.

In the stage of general semantic segmentation, the configuration of the feature encoder is (32, 2048) (64, 1024) (128, 512) (256, 256), where (32, 64, 128, 256) represents the feature dimension in the corresponding layer, and the output point number is (2048, 1024, 512, 256). In the point transformer block, the decoder has a symmetrical configuration with the encoder. The Adam optimizer is employed in the network. We train the network for 20 epochs with batch size 4 and an initial learning rate of 0.0005.

In the stage of refined semantic segmentation, Vehicle and Chimney are treated as imbalanced rare classes for refined semantic segmentation according to the section 3.1. For the backbone network PointNet++ in the process of 3D proposal generation, we subsample 65536 points from each split as the inputs of the training. Then 4 set-abstraction layers with multi-scale grouping are used to subsample points into groups with sizes 4096, 1024, 256, 64. For the 3D BBox refinement network, 512 points are randomly selected from each 3D proposal as the input and 3 set abstraction layers with group sizes 128, 32, 1 are used to generate a single feature vector for the BBox refinement. The proposal generation network is trained for 300 epochs with batch size 8 and learning rate 0.002, while the BBox refinement network is trained for 200 epochs with batch size 4 and learning rate 0.002.

3.3 Segmentation Results

Our proposed approach is evaluated on the H3D Benchmark dataset. The segmentation results are evaluated by overall accuracy (OA) and F1-score.

The semantic segmentation results and confusion matrix are shown in Figure 3(b), where the overall accuracy achieves state-of-the-art 89.35% and the mean F1-score achieves outstanding 75.70%. The visualization of the corresponding result on test set is shown in Figure 4(c), where the point cloud with RGB is shown in Figure 4(a). The confusion mainly exists between Vehicle and Urban furniture, and Soil/gravel are often inferred as Low vegetation. These ambiguities are caused by their limited inter-class distances and scarce appearances.

Models	OA	Mean F1-score	F1-score	
			Vehicle	Chimney
Single-stage	89.19%	65.36%	0	0
Two-stage	89.35%	75.70%	63.63%	52.00%

Table 2. Performance comparison between the proposed two-stage approach and the single-stage segmentation.

In order to verify the effectiveness of the proposed segmentation approach for imbalanced fine-grained objects, we also compare it with the single-stage segmentation (without refined semantic segmentation). Figure 3(a) shows the detailed confusion matrix and Figure 4(b) displays the visualization result. The performance comparison is shown in Table 2.

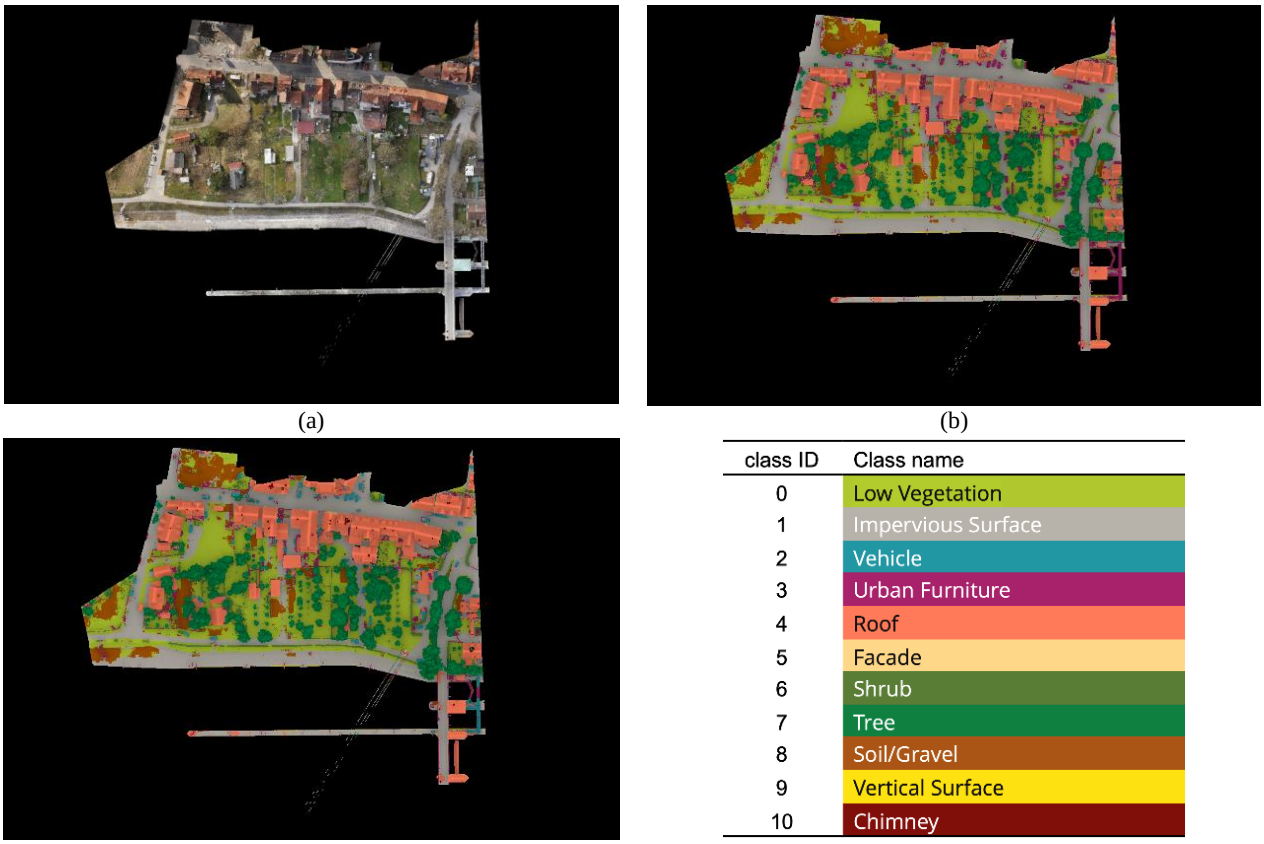

Figure 4. Visualization of the test results on (a) point cloud with RGB (b) the single-stage segmentation (c) our proposed two-stage approach (d) the class catalog of the H3D dataset.

Benefited from the specialized segmentation for imbalanced rare classes, the two-stage approach performs better than the single-stage method in all evaluation metrics. Due to the low percentage of the fine-grained rare classes, there is only a limited improvement (0.16%) in overall accuracy. However, Vehicle and Chimney have achieved breakthroughs from zero to 63.63% and 52.00% in F1-score respectively, which has also promoted our two-stage approach outperforms the single-stage segmentation by a large margin of 10.34 percentage points in the mean F1-score.

4. CONCLUSION AND OUTLOOK

In this work we have presented a two-stage segmentation approach for imbalanced rare classes, which has unified object detection models into the semantic segmentation framework. Comprehensive experiments on large-scale urban data demonstrated that the proposed approach have obtained the state-of-the-art overall accuracy and the satisfactory mean F1-score, and have achieved the outstanding F1-scores for imbalanced rare classes. However, the proposed solution also has limitations. The proposed method is only suitable for fine-grained objects with strong discrete distributions, and it requires a considerable amount of computational resources due to the additional training for the refinement network. In future work, we will focus on the feature-level unification of detection networks into segmentation framework, and construct an end-to-end lightweight segmentation network for imbalanced rare classes.

ACKNOWLEDGEMENTS

The research work in this paper has been funded by China Scholarship Council (No. 202008080145).

REFERENCES

Chen, S., Liu, B., Feng, C., Vallespi-Gonzalez, C., Wellington, C., 2020. 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. *IEEE Signal Processing Magazine*, 38(1), 68-86. doi.org/10.1109/MSP.2020.2984780.

Dong, J., Chen, Q., Yan, S., Yuille, A., 2014. Towards unified object detection and semantic segmentation. *In European Conference on Computer Vision*, 299-314. doi.org/10.1007/978-3-319-10602-1_20.

Ester, M., Kriegel, H. P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In kdd*, 96(34), 226-231.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. *In Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315-323.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 9224-9232. doi.org/10.1109/CVPR.2018.00961.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J. D., Ledoux, H., 2021. The Hessigheim 3D (H3D) Benchmark on Semantic Segmentation of High-Resolution 3D Point Clouds and Textured Meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1, 11. doi.org/10.1016/j.ophoto.2021.100001.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O., 2019. Pointpillars: Fast encoders for object detection from point clouds. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12697-12705. doi.org/10.1109/CVPR.2019.01298.

Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *In Proceedings of the IEEE international conference on computer vision*, 2980-2988. doi.org/10.1109/TPAMI.2018.2858826.

Maturana, D., Scherer, S., 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. *In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 922-928. doi.org/10.1109/IROS.2015.7353481.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 652-660. doi.org/10.1109/CVPR.2017.16.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Advances in Neural Information Processing Systems*, 30. doi.org/10.48550/arXiv.1706.02413.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention*, 234-241. doi.org/10.1007/978-3-319-24574-4_28.

Schmohl, S., Soergel, U., 2019. Submanifold Sparse Convolutional Networks for Semantic Segmentation of Large-Scale ALS Point Clouds. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4. doi.org/10.5194/isprs-annals-IV-2-W5-77-2019.

Shi, S., Wang, X., Li, H., 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770-779. doi.org/10.1109/CVPR.2019.00086.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H., 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529-10538. doi.org/10.1109/CVPR42600.2020.01054.

Weinmann, M., Jutzi, B., Mallet, C., 2013: Feature relevance assessment for the semantic interpretation of 3D point cloud data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5(W2), 1. doi.org/10.5194/isprsannals-II-5-W2-313-2013.

Yang, Z., Sun, Y., Liu, S., Jia, J., 2020. 3dssd: Point-based 3d single stage object detector. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11040-11048. doi.org/10.1109/CVPR42600.2020.01105.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259-16268. doi.org/10.1109/ICCV48922.2021.01595.