TARGET-GUIDED LEARNING FOR RARE CLASS SEGMENTATION IN LARGE-SCALE URBAN POINT CLOUDS

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

[1]Institute for Photogrammetry, University of Stuttgart, 70174 Stuttgart, Germany -
(xinlong.zhang, ruihang.xue, uwe.soergel) @ifp.uni-stuttgart.de
[2]State Key Laboratory of Geo-Information Engineering, Beijing, China - lindong_hb59@163.com

KEY WORDS: Deep Learning, Transformer, Instance Features, Rare Classes, Semantic Segmentation, Large-Scale Point Clouds.

ABSTRACT:

In large-scale urban areas, the diversity of objects and the complexity of scenes pose challenges to semantic segmentation of point clouds. In particular, the data imbalance problem often results in poor performance for rare classes in large scenes. This paper proposes a rare class segmentation method based on the target-guided transformer network. In the network, all the feature extraction and segmentation procedures are realized by attention mechanisms. The self-attention blocks are embedded in U-Net-like structure to gradually integrate the features from local to global. Then, under the supervision of our target-guided block, the instance features of data-imbalanced rare classes are mapped onto the multi-scale features. At last, a multi-layer perceptron is utilized to convert the fused features to the segmentation logits for generating the semantic labels. Experiments using the Hessigheim High-Resolution 3D Point Cloud Benchmark indicated that our approach considerably outperforms the baseline network by up to 11.66% in terms of mean F1 score. In particular, the rare classes Vehicle and Chimney obtain outstanding F1-scores of 82.40% and 82.51%, respectively. Furthermore, our method achieves an overall accuracy of 87.63%, which increases by 1.09% compared to the baseline model.

1. INTRODUCTION

Driven by the fast development of lightweight LiDAR devices and unmanned aerial vehicles, large-scale high-resolution point clouds are automatically acquired by highly integrated platforms. To comprehensively interpret a complex point cloud scene, an indispensable solution is to obtain category information of each point, which is referred to semantic segmentation. Such point-wise semantics are valuable cues across a variety of remote sensing tasks, such as collapsed building detection (Xiu et al., 2020), cultural heritage segmentation (Pierdicca et al., 2020), and tree species classification (Michałowska and Rapiński, 2021).

In recent years, various deep learning models have been developed with outstanding performance for semantic segmentation on point clouds. Inspired by 2D convolutional neural networks (CNNs), voxel-based methods (Maturana and Scherer, 2015, Wu et al., 2015) convert point clouds into 3D volumetric representations to make the data structure suitable for 3D CNNs. Although encouraging performance has been achieved, these methods cannot cope with real-time application scenarios since the computation and memory footprint grow cubically with the resolution. Sparse convolutional neural networks (Liu et al., 2015) and its 3D application submanifold sparse convolutional networks (Graham et al., 2018) are proposed to take advantage of the sparsity, operating only on voxels that are not empty. Another branch such as MVCNN (Su et al., 2015) and SnapNet (Boulch et al., 2018), is to project 3D point clouds into 2D images from multiple perspectives, aggregated features from the well-established 2D CNNs then are reprojected into 3D space to achieve the task of semantic segmentation. Both of them convert point clouds into regular structures, i.e. 3D voxels or 2D grids. However, the quantization error introducd in data conversion is a non-negligible problem, i.e., the boundaries of voxels and grids cause information loss and the fine geometric structures are ignored.

In contrast, point-based methods directly consum raw point clouds without any voxelization or projection. PointNet (Qi et al., 2017a) extracts pointwise features independently with multi-layer perceptrons (MLPs) and aggregates global features with the max-pooling layer. Since the local structural information between points cannot be captured, a hierarchical feature aggregation scheme based on the set abstraction is designed by PointNet++ (Qi et al., 2017b) to abstract the local features layer by layer. Nevertheless, the max-pooling layer in the PointNet family only retains the maximum elements as the global descriptors, resulting in information loss about the initial spatial distribution of the input point set. The self-attention operator in Point Transformer (Zhao et al., 2021) weights each element adaptively, while invariant to permutation and cardinality of the input elements. This weight based attention mechanism is superior to the symmetric function, i.e., the maximum pooling. Moreover, point clouds are essentially sets embedded in 3D space, which makes the positional encoding in transformer framwork a quite natural process.

However, most of these cutting-edge approaches developed in the computer vision community have focused on general majority classes in ground scans with limited space or indoor scenes. The complexity of large-scale scenarios and the diversity of objects in urban areas are neglected, without taking into account specialized segmentation solutions for imbalanced rare classes in large-scale point clouds. Actually, learning from imbalanced data is still a challenging problem in point cloud segmentation (Guo et al., 2020). To this end, we propose a segmentation model based on the strong shape cues of targets for imbalanced rare classes, namely target-guided transformer network. The symmetric encoder-decoder network first

* Corresponding author

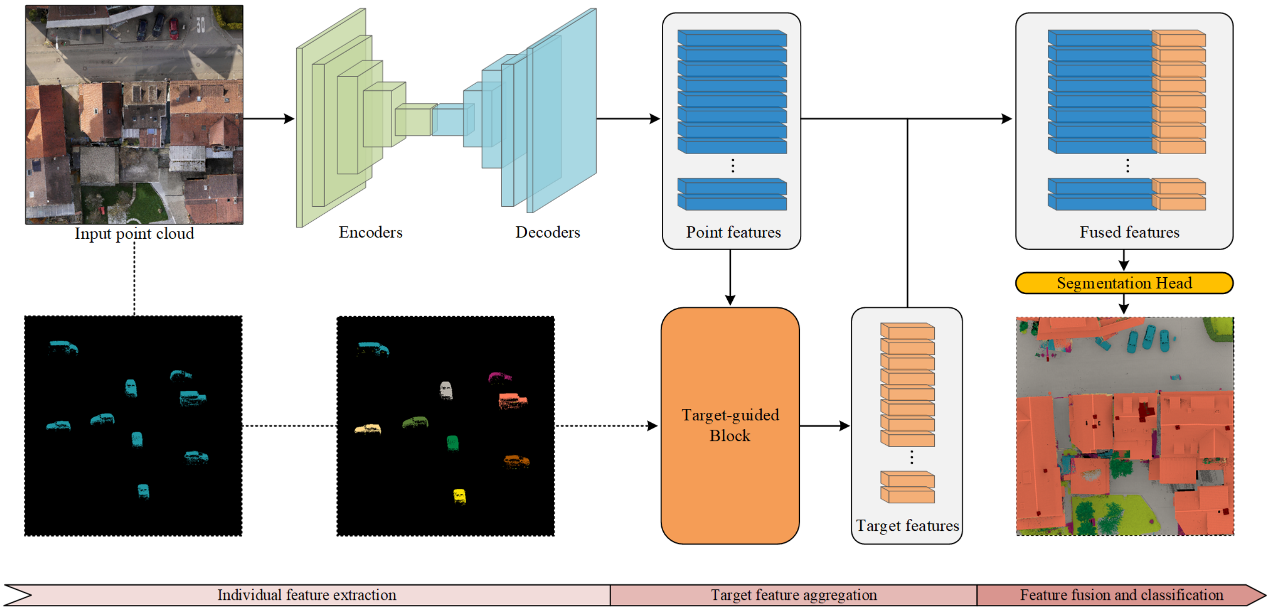


Figure 1. Structure of the proposed network for rare class segmentation. A symmetrical network based on self-attention blocks extracts the point-wise features of the point cloud. Meanwhile, the target features are supervised by the guided-block and aggregated in the point domain. Afterwards, the fused features are fed to the segmentation head to obtain semantic labels.

extracts point-wise features aggregated by connection of multiple self-attention layers. Then these features are passed to the target-guided block to extract target features for rare classes and mapped onto point-wise features. Finally, a MLP is adopted to obtain semantic labels. The proposed model is an End-to-End network, and its target-guided block can simultaneously operate on multiple rare classes. In the experiments of point cloud segmentation on the Hessigheim High-Resolution 3D Point Cloud Benchmark (H3D) (Kölle et al., 2021), the proposed network has achieved better mean F1-score and overall accuracy than the baseline model without the target-guided block. Remarkably, the F1-scores of rare classes have improved significantly.

2. METHODOLOGY

The overall structure of the proposed method is illustrated in Figure 1, which comprises three distinct stages: individual feature extraction, target feature aggregation, feature fusion and classification. Firstly, the point cloud is fed to a symmetrical encoder-decoder network for pointwise feature extraction. Then, the target features of the imbalanced rare classes are aggregated by the target-guided block. Finally, the target features converted to the point domain and pointwise features are concatenated along the channel dimension, and the fused features are further passed through a MLP for generating semantic labels. The detailed process of the network is given below.

2.1 Self-attention Layer

Self-attention layer is the core of the encoder-decoder network to achieve individual feature extraction, which is formed by cascading 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.

Given an input point set X, the subset $X(i) \in X$ is the attention scope of the point x_i, which is obtained by the the k nearest neighbors algorithm. Hence the self-attention calculation of the point x_i is defined as:

$$y_i = \sum_{x_j \in X(i)} \alpha(\beta(\varphi_q(x_i) - \varphi_k(x_j) + p)) \odot (\varphi_v(x_j) + p) \quad (1)$$

where y_i is the output feature. φ_q, φ_k, φ_v are all linear mappings for adapting to different feature dimensions, which perform pointwise feature transformations. β is the attention mapping function that produces attention vectors for feature aggregation, which is implemented by a MLP, i.e., two linear layers and one ReLU layer. α is the softmax activation function and $\odot$ denotes to the dot product, namely the element-wise multiplication, which plays the role of feature aggregation. p is the positional encoding, which is a linear mapping from the relative coordinate of the natural points embedded in 3D space:

$$p = \gamma(\zeta_i - \zeta_j) \quad (2)$$

where ζ_i and ζ_j are respectively the 3D coordinates of points i and j. γ is the encoding function with two linear layers and one ReLU layer (Glorot et al., 2011).

U-Net-style architecture (Ronneberger et al., 2015) is adopted to connect self-attention layers, which aims to extract features at different scales, as shown in Figure 2. Each feature encoder has a self-attention layer connected by a down-sampling layer, while each feature decoder has a self-attention layer connected by a up-sampling layer. The down-sampling layer is realized by the farthest point sampling and KNNs searching. The up-sampling layer is implemented by trilinear interpolation. Besides, the features of the corresponding encoder layer are added to the up-sampled new features by skip connections. Multi-scale feature extraction is realized by stacked encoders and decoders. The number of encoders N directly determines the size of the network, which can be varied according to the application.

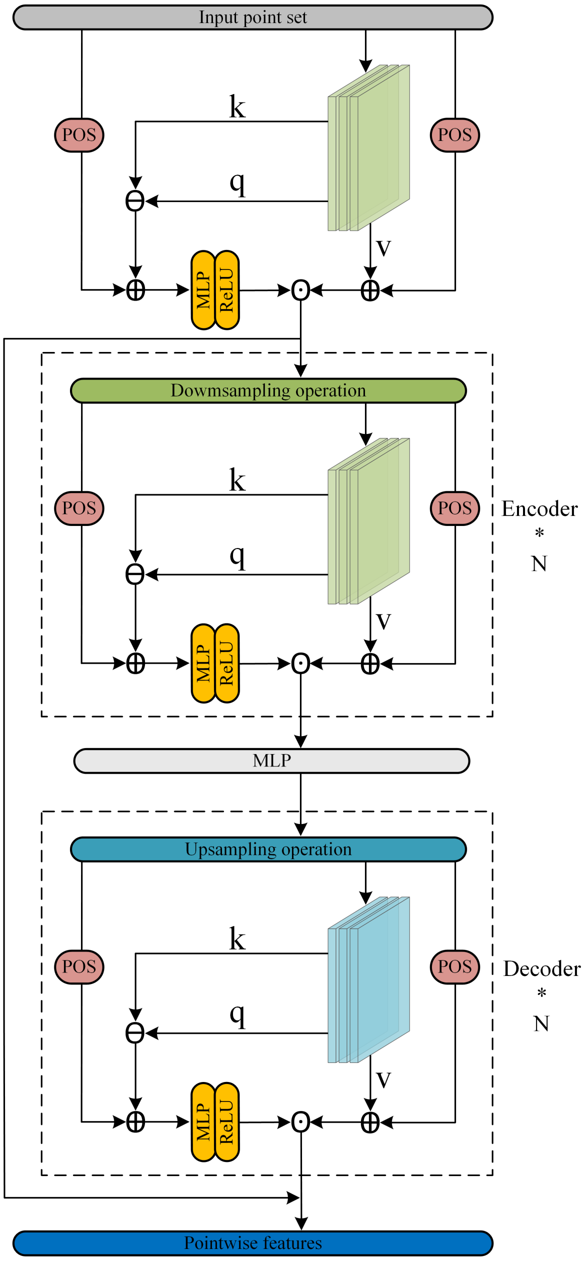


Figure 2. Self-attention layer in U-Net-style architecture.

2.2 Target-guided Block

The strong shape cues of the target object can provide valuable information for semantic segmentation (Dong et al., 2014), distinguishing the contour features of each rare target at the individual level and facilitating the reinforcement of rare class representation. The simplified outlines of rare targets can be represented in the form of $\{w, l, h\}$, providing approximate descriptions of their three-dimensional size. In addition, position features $\{u, v, d, r\}$ are described by the ground center and heading of targets, with $\{u, v, d\}$ denoting the location and $\{r\}$ denoting the orientation. In the top view feature map, the orientation of the target is simplified to the yaw angle $\{\theta\}$ on the ground plane.

We construct the target-guided block for target feature aggregation of rare classes, which is visualized in Figure 3. The pointwise features are first transformed into the regular representations using 3D voxelization, which guarantees the efficiency of the target-guided block. Specifically, the pointwise features within a voxel are combined through averaging to create a voxel feature. Then the features in height dimension are flattened to obtain a top-view feature map, which aims to eliminate height redundancy to narrow the search space for rare targets. Afterwards, the collapsed features are passed forward to the target-guided layer for generating one target heatmap with multiple channels, each corresponding to a distinct rare class. Note that the peak of each heatmap indicates the 2D ground center $\{\hat{u}_i, \hat{v}_i\}$ of the i_{th} target. In order alleviate the imbalance between center points and background points, the variant of focal loss (Lin et al., 2017) is chosen to update the network as follows:

$$loss_{map}(p_c) = -\alpha_c(1-p_c)^{\gamma} \cdot log(p_c), \\ \text{where } p_c = \begin{cases} p, & \text{if center points} \\ 1-p, & \text{otherwise} \end{cases} \tag{3}$$

where the positive supervision for target centers is enlarged by the Gaussian rendering encoded in p_c at each ground truth center.

In the end, the heat map is utilized to mask the top-view feature map, enabling the extraction of the target's features from the corresponding location. In an effort to increase the efficiency of the block, the pointwise features are quantized into voxels, which results in the sacrifice of intricate geometric details. To counteract this, the sub-voxel location refinement $\{\delta_i^u, \delta_i^v\}$ is designed to compensate for the quantization error caused by voxelization. Moreover, the difference between positive and negative orientation is not necessary to distinguish in the semantic segmentation task, so only the sine of the yaw angle $\{\sin(\theta_i)\}$ can indicate the orientation. Thus, the regression parameters of the target contour can be constructed as

$$\hat{T}_i = \{\hat{u}_i + \delta_i^u, \hat{v}_i + \delta_i^v, \hat{d}_i, \sin(\hat{\theta}_i), \hat{w}_i, \hat{l}_i, \hat{h}_i\} \tag{4}$$

All the learnable parameters could be updated by the Smooth L$_1$ Loss (Girshick, 2015):

$$loss_{off} = \frac{1}{N}\sum_{i=1}^{N} L_s(T_i - \hat{T}_i), \\ \text{where } L_s(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{5}$$

To generate supervisory signals T_i for the target-guided block, we design an automation workflow based on the discrete distribution of rare targets. Firstly, fine-grained rare targets are selected individually to avoid confusion with general targets. Then, considering their discrete distribution, density-based spatial clustering of applications with noise method (Ester et al., 1996) is utilized to divided the point cloud of rare targets into separate reliable clusters. Afterwards, the vertices of each cluster corresponding to the convex hull are calculated and adjusted to the vertices of target outlines, which guide the supervision of learnable targets.

2.3 Feature Fusion and Classification

Target features are interpreted as descriptions of contours at the individual level, providing valuable clues between diverse classes. To this end, the attribute is assigned to each point on the corresponding target, which serves as prominent features for

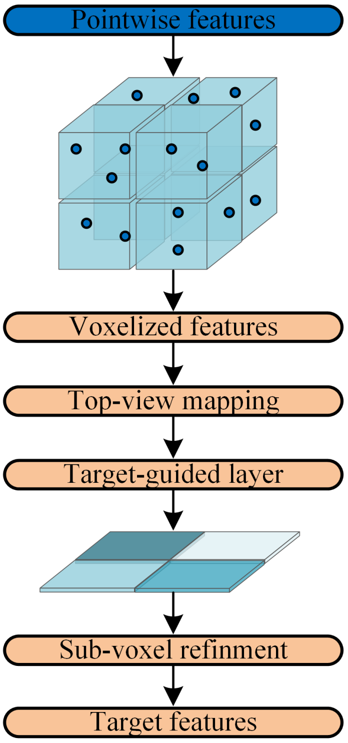


Figure 3. Workflow of the target-guided block.

discriminating between rare and general classes. Specifically, the outlines obtained from the target-guiding block are utilized as the constraints on masks, based on which we concatenate target features with point-wise features along the channel dimension. For general classes outside the mask, we set the pseudo target features as zero tensors of equal length.

For the semantic segmentation task, a MLP maps the fused features to the label space l_k at the last layer. Since the class imbalance in large-scale data can be alleviated by weighting the loss, the class-weighted cross entropy loss (Long et al., 2015, Ronneberger et al., 2015) is optimized to update the network parameters:

$$loss_{seg} = -\sum_{i=1}^{N}\sum_{k=1}^{C} \alpha \cdot y_{ik} \cdot log(p_i) \tag{6}$$

where N and C are the number of points and the number of classes, respectively. p_i represents the probability distribution of the i_{th} point on C classes, which is inferred from the model. y_{ik} denotes the one-hot distribution corresponding to the i_{th} point, indicating whether the point belongs to the k_{th} class. α is a modulating factor that balances the importance of different classes, which is determined by the statistics of points on C classes in training data.

All losses proposed in previous sections participate in backpropagation simultaneously with different weights, i.e. the focal loss with Gaussian kernels for the center points, the smooth L1 loss for the regression and the class-weighted cross entropy loss for the pointwise output:

$$L = loss_{seg} + \xi \cdot loss_{map} + \vartheta \cdot loss_{off} \tag{7}$$

where ξ and ϑ are the loss coefficients to balance the effect of the target-guided losses on the main task loss.

3. EXPERIMENTS

3.1 Data Description

The experiments are based on the large-scale H3D dataset (Kölle et al., 2021). The dataset comprises highly dense LiDAR point cloud around 800 pts/m² enriched by RGB images with the ground sampling distance of about 2-3 cm, which was acquired by a Riegl VUX-1LR Scanner and two oblique-looking Sony Alpha 6000 cameras integrated on a RIEGL Ricopter platform. The format of dataset features $\{x, y, z, r, g, b, i, n_e, e_n\}$ are consist of the coordinate, the texture color, the reflectance, the number of echoes and the echo number. The entire urban scene is subdivided into three distinct areas for training, validation, and testing. This study is based on the training and validation sets with approximately 59.45 million and 14.46 million points, respectively.

Eleven semantic categories were manually annotated, including Low Vegetation, Impervious Surface, Vehicle, Urban Furniture, Roof, Facade, Shrub, Tree, Soil/Gravel, Vertical Surface, and Chimney. However, the class imbalance in large-scale data occurs on such fine-grained class catalog. Detailed statistics of class occurrences in H3D dataset is shown in Table 1, where the number of points for different classes considerably varied. 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 rare class segmentation a challenging task.

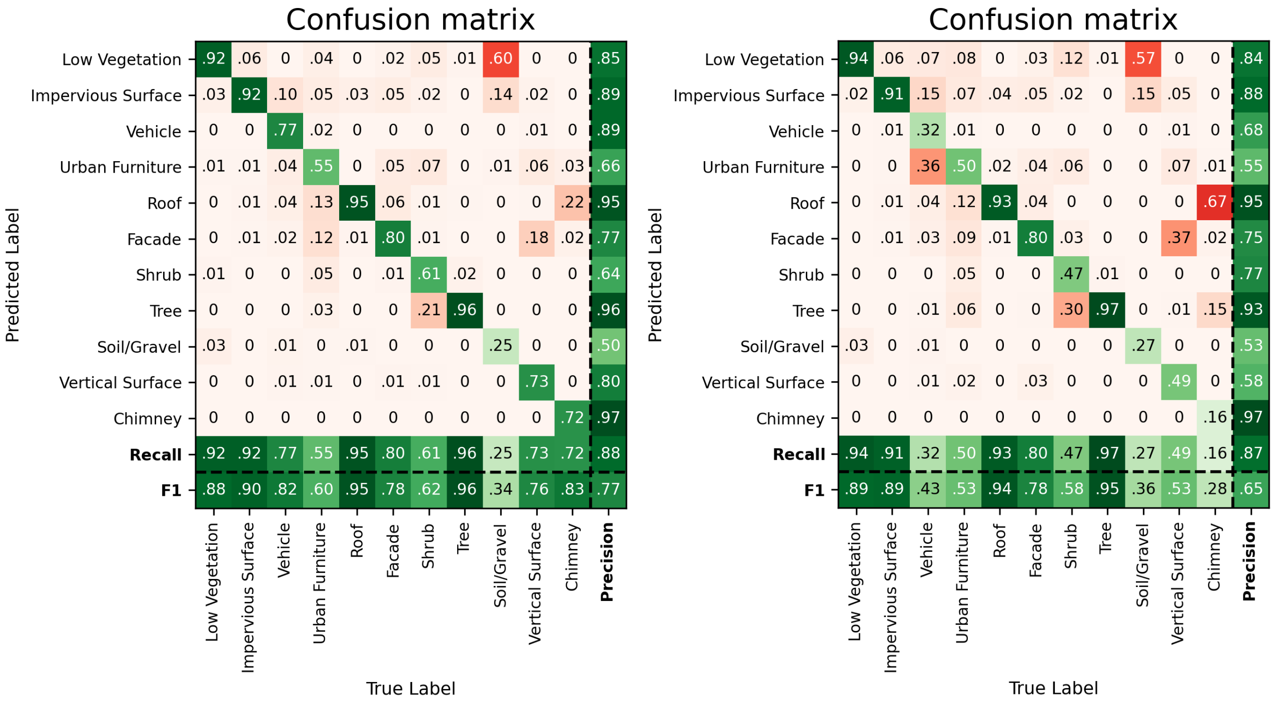
Class	Traning set		Validation set	
	points	%	points	%
Low Vegetation	21375614	35.96	3738743	25.85
Impervious Surface	10419635	17.53	3212988	22.21
Vehicle	258032	**0.43**	183263	1.27
Urban Furniture	1159205	1.95	455389	3.15
Roof	6279431	10.56	3052150	21.10
Facade	1198227	2.02	551996	3.82
Shrub	1077141	1.81	341579	2.36
Tree	8086818	13.60	2218551	15.34
Soil/Gravel	8590706	14.45	592510	4.10
Vertical Surface	974976	1.64	101243	0.70
Chimney	25321	**0.04**	15836	0.11
Total	59445106	1.00	14464248	1.00

Table 1. Class distribution of H3D dataset.

3.2 Implementation Details

In this work, $N = 4$ is set to build the backbone network, where down-sampling rates and feature channels for the encoders are $[4, 4, 4, 4]$ and $[64, 128, 256, 512]$, respectively. Note that the decoder has a symmetrical configuration with the corresponding encoder. During the training, Vehicle and Chimney are treated as imbalanced rare classes for the target-guided block according to the Table 1. We sliced blocks with a length of 10 m, and the selected seed point at each batch is determined by a statistical function. The Adam optimizer is employed in the network with an initial learning rate of 0.001. The model was implemented in the framework of PyTorch and trained on a single NVIDIA RTX2080Ti 11GB GPU.

https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1693-2023 | © Author(s) 2023. CC BY 4.0 License.



Predicted Label	Low Vegetation	Impervious Surface	Vehicle	Urban Furniture	Roof	Facade	Shrub	Tree	Soil/Gravel	Vertical Surface	Chimney	**Precision**
Low Vegetation	.92	.06	0	.04	0	.02	.05	.01	.60	0	0	.85
Impervious Surface	.03	.92	.10	.05	.03	.05	.02	0	.14	.02	0	.89
Vehicle	0	0	.77	.02	0	0	0	0	0	.01	0	.89
Urban Furniture	.01	.01	.04	.55	0	.05	.07	0	.01	.06	.03	.66
Roof	0	.01	.04	.13	.95	.06	.01	0	0	0	.22	.95
Facade	0	.01	.02	.12	.01	.80	.01	0	0	.18	.02	.77
Shrub	.01	0	0	.05	0	.01	.61	.02	0	0	0	.64
Tree	0	0	0	.03	0	0	.21	.96	0	0	0	.96
Soil/Gravel	.03	0	.01	0	.01	0	0	0	.25	0	0	.50
Vertical Surface	0	0	.01	.01	0	.01	.01	0	0	.73	0	.80
Chimney	0	0	0	0	0	0	0	0	0	0	.72	.97
Recall	.92	.92	.77	.55	.95	.80	.61	.96	.25	.73	.72	.88
F1	.88	.90	.82	.60	.95	.78	.62	.96	.34	.76	.83	.77

(a)

Predicted Label	Low Vegetation	Impervious Surface	Vehicle	Urban Furniture	Roof	Facade	Shrub	Tree	Soil/Gravel	Vertical Surface	Chimney	**Precision**
Low Vegetation	.94	.06	.07	.08	0	.03	.12	.01	.57	0	0	.84
Impervious Surface	.02	.91	.15	.07	.04	.05	.02	0	.15	.05	0	.88
Vehicle	0	.01	.32	.01	0	0	0	0	0	.01	0	.68
Urban Furniture	0	0	.36	.50	.02	.04	.06	0	0	.07	.01	.55
Roof	0	.01	.04	.12	.93	.04	0	0	0	0	.67	.95
Facade	0	.01	.03	.09	.01	.80	.03	0	0	.37	.02	.75
Shrub	0	0	0	.05	0	0	.47	.01	0	0	0	.77
Tree	0	0	.01	.06	0	0	.30	.97	0	.01	.15	.93
Soil/Gravel	.03	0	.01	0	0	0	0	0	.27	0	0	.53
Vertical Surface	0	0	.01	.02	0	.03	0	0	0	.49	0	.58
Chimney	0	0	0	0	0	0	0	0	0	0	.16	.97
Recall	.94	.91	.32	.50	.93	.80	.47	.97	.27	.49	.16	.87
F1	.89	.89	.43	.53	.94	.78	.58	.95	.36	.53	.28	.65

(b)

Figure 4. Confusion matrices on (a) our proposed network (b) the baseline model PointTransformer.

3.3 Evaluation Metrics

The overall accuracy (OA) and the mean F1-score are used to evaluate the performance of semantic segmentation. OA represents the percentage of correctly classified points. The F1-score is the harmonic mean combining precision and recall, while the mean F1-score denotes the average of F1-scores of all categories:

$$P_k = \frac{TP_k}{TP_k + FP_k}$$
$$R_k = \frac{TP_k}{TP_k + FN_k} \qquad (8)$$
$$F1_k = \frac{2 \cdot P_k \cdot R_k}{P_k + R_k}$$

where TP_k, FP_k and FN_k are true positives, false positives, and false negatives determined for class k, respectively.

3.4 Results and Analysis

The segmentation result of the proposed network in form of confusion matrix is shown in Figure 4(a), where the overall accuracy achieves 87.63% and the mean F1-score achieves outstanding 76.76%. It can be observed that the confusion mainly exists between Low Vegetation and Soil/Gravel, which is caused by their similar rough geometric appearances.

In order to verify the effectiveness of the proposed approach for imbalanced rare classes, we compare it to the state-of-the-art baseline model PointTransformer. The detailed confusion matrix of the baseline model is illustrated in Figure 4(b) and the performance comparison is shown in Table 2. Our approach performs better than the baseline in all evaluation metrics, which is benefited from the specialized target-guided block for rare classes. It is observed that the mean F1-score has boosted for 11.66% than the baseline. Specifically, the F1-scores of the rare classes Vehicle and Chimney have been significantly improved to 82.40% and 82.51%, respectively. However, the designed model only outperforms the baseline by a small margin of 1.09 percentage points, which is consistent with the limited relative class occurences, i.e. the extremely low percentages of rare classes.

Models	F1-Score		Mean F1-Score	OA
	Vehicle	Chimney		
Baseline	43.13%	27.72%	65.10%	86.54%
Ours	82.40%	82.51%	76.76%	87.63%

Table 2. Performance comparison between the proposed network and the baseline model PointTransformer.

4. CONCLUSION

In this work, we proposed an end-to-end segmentation network for imbalanced rare classes in large-scale scenes. Self-attention based transformer was designed as the backbone for the individual feature extraction. And a target-guided block was developed to take advantage of the strong shape cues of rare targets, facilitating the reinforcement of rare class representation. Comprehensive experiments on large-scale urban data indicated that, our method achieves outstanding OA and significantly improves F1-scores of rare classes compared to the scheme without our target-guided block.

However, the proposed solution still has the limitation of uncertain error from automatic supervisory signals, which are significantly contingent upon the spatial distribution of rare targets. Furthermore, the simplified description of contours in the target-guided block makes it only suitable for approximately cuboid instances. The future work will be focused on the automated generation of robust supervisory signals and the refined description of target contours.

ACKNOWLEDGEMENTS

Considerable parts of the research presented in this paper were supported by the China Scholarship Council - No. 202008080145 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2120/1 - 390831618.

REFERENCES

Boulch, A., Guerry, J., Le Saux, B., Audebert, N., 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers & Graphics*, 71, 189–198.

Dong, J., Chen, Q., Yan, S., Yuille, A., 2014. Towards unified object detection and semantic segmentation. *Proceedings of Computer Vision–ECCV 2014: 13th European Conference*, 299–314.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Girshick, R., 2015. Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. *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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9224–9232.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12), 4338–4364.

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, 100001.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, B., Wang, M., Foroosh, H., Tappen, M., Pensky, M., 2015. Sparse convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 806–814.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

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

Michałowska, M., Rapiński, J., 2021. A review of tree species classification based on airborne LiDAR data and applied classifiers. *Remote Sensing*, 13(3).

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S., Frontoni, E., Lingua, A. M., 2020. Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sensing*, 12(6).

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Proceedings of Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, 234–241.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 945–953.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Xiu, H., Shinohara, T., Matsuoka, M., Inoguchi, M., Kawabe, K., Horie, K., 2020. Collapsed building detection using 3d point clouds and deep learning. *Remote Sensing*, 12(24).

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.