Supplementary Material for "Deformable Shape-aware Point Generation for 3D Object Detection"

Anonymous ACCV 2024 Submission

Paper ID #531

1 Appendix

In this document, we provide additional experiments on point cloud objects at different distance ranges, additional implementation details, and more experimental visualization results.

1.1 Data Statistic On Different Distance

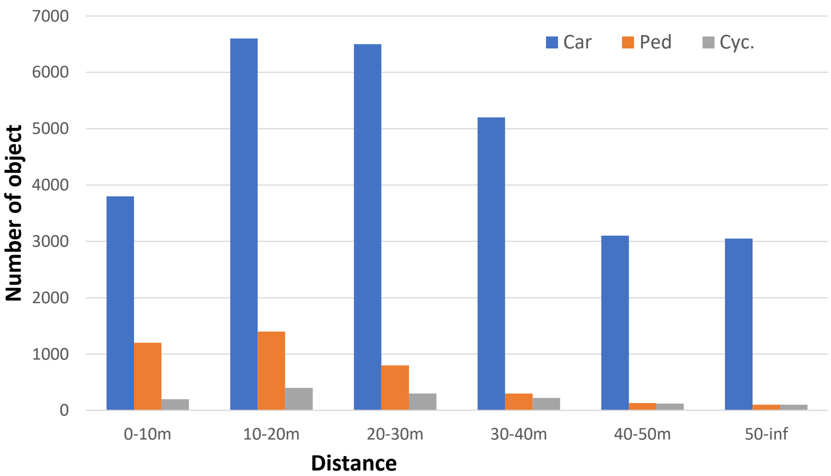

Fig. 1: The distribution of objects in point clouds over different distance ranges on the KITTI validation set.

Taking into account the effect of different distance on 3D object detection, we conduct an investigation into the distance distributions of point cloud objects on the public benchmark. Based on the GT labels provided by the KITTI [1] training set, we divided the scene into six distance ranges, each spanning 10 meters, as the distance increases, the point clouds of the object become increasingly sparse. Fig. 1 presents the distribution of cars, pedestrians, and cyclists within these distance ranges.

Since object detection servers for the perception system of autonomous vehicles, the further away(more 40 meters) an object can be detected, the more time is left for the decision planning system, thus autonomous vehicles will be safer. As shown in Fig.1, objects at distances of more than 30 meters are dramatically reduced, and objects often contain fewer points, posing a serious challenge to detection performance. Most existing 3D detectors are designed for the near distances and perform poorly even at the far distances. Thus, accurate detection of sparse remote targets is a reasonably correct object for 3D object detection.

1.2 More Implementation Details

In this work, before sending to networks, the raw points are first encoded into pillars for heatmap prediction, we define the detection range as [0, 69.12]m for the X-axis, [-39.68, 39.68]m for the Y-axis, and [-3, 1]m for the Z-axis. For the Waymo Open Dataset [2], the detection ranges are set to [-75.2m, 75.2m] for the X and Y axes, and [-2m, 4m] for the Z-axis. The voxel size for each voxel is set to (0.1m, 0.1m 0.15m). We set the pillar size to (0.16m, 0.16m, 4m). We randomly sample 128 proposals for training, and 50% of them are positive samples that have IoU>0.55 with the corresponding ground truth boxes. In the RoI grid pooling step, the dimension of each grid's feature f_{gi} is set to 96 with a grid size G of 6. For each proposal, the point cloud encoder in the detection head extracts an RoI feature vector of dimension 256. The number of points used to calculate foreground score , is set to 2,048.

The experiments on Waymo [2], our point generation supervision is the same as PGRCNN [3], so we approximated the complete shape by utilizing different instances of the same object class to get complete objects on waymo datasets. We used an almost identical network architecture in KITTI for the experiments, except using an increased number of channels of the proposal layers to (128, 256) and 192 for grid feature dimension.

1.3 Details of Point Generation Losses

In this section, we provide further details on the point generation loss $\mathcal{L}_{\text{gen}}$. Our implementation of $\mathcal{L}_{\text{gen}}$ is consistent with those used in PGRCNN [3].

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{shape}} , \tag{1}$$

$\mathcal{L}_{\text{seg}}$ is a point-level segmentation loss that generates foreground scores for points to determine if the points fall within a ground-truth bounding box. This process assigns them to segmentation labels. We apply Focal Loss on the generated points:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N_p} \sum_j (1 - s_j)^\gamma \log s_j, \tag{2}$$

where s_j , $j = 1, 2, \cdots N_p$ are the foreground score of the sampled points.

Table 1: Performance comparison at different distance ranges on the moderate level car class of KITTI val split set. The results are evaluated with the moderate AP calculated by 40 recall positions. The best performance value is in bold.

Class	Car			Pedestrian			Cyclist		
Distance	0-20m	20-40m	40-inf	0-20m	20-40m	40-inf	0-20m	20-40m	40-inf
VoxelRCNN	94.36	82.59	42.78	69.66	37.25	1.98	90.42	60.48	32.69
SIENet	95.36	84.56	44.21	71.56	39.58	2.69	93.42	62.36	35.89
PGRCNN	96.36	85.48	45.23	74.56	40.26	2.58	94.12	64.56	37.55
DSaPG(Ours)	**96.54**	**86.45**	**46.49**	**75.35**	**44.23**	**2.84**	**94.23**	**66.85**	**38.22**
Improvement	+***0.18***	+***0.93***	+***1.26***	+***0.79***	+***3.97***	+***0.26***	+***0.11***	+**2.29**	+***0.67***

$\mathcal{L}_{\text{shape}}$ supervises of the shape of the generation point cloud. We employed the approximation method proposed in [4] to estimate the complete shape of the object by utilizing other instances of objects within the provided dataset. We employ Chamfer Distance on foreground proposals as follows:

$$\begin{aligned}\mathcal{L}_{\text{shape}} &= \frac{1}{N_{fp}} \sum_r \Bigg(\frac{1}{|\mathbf{P}_r|} \sum_{\mathbf{x} \in \mathbf{P}_r} \min_{\mathbf{y} \in \mathbf{P}_r^*} \|\mathbf{x} - \mathbf{y}\|_2^2 + \\ &\qquad \frac{1}{|\mathbf{P}_r^*|} \sum_{\mathbf{y} \in \mathbf{P}_r^*} \min_{\mathbf{x} \in \mathbf{P}_r} \|\mathbf{y} - \mathbf{x}\|_2^2 \Bigg),\end{aligned} \tag{3}$$

$N_{fp}$is the number of foreground proposals, and $\mathbf{P}_r$ and $\mathbf{P}_r^*$ are the generated and the target point cloud.

1.4 Experimental Analysis At Different Distances

We report the 3D detection performance of the proposed DsaPG compared to VoxelRCNN [5], SIENet [6] and PGRCNN [3] under the distinct distance ranges in Table 1. We can observe 3D detectors can achieve excellent detection performance under the distance of less than 40 meters, however, there is a significant decrease for detection performance at the distance exceeding 30 meters. The reason may be the objects closer to the LiDAR sensor (less than 30 meters) contain rich information under the dense point cloud, while the distant sparse points suffer from incomplete information.

Obviously, compared with VoxelRCNN [5], the point cloud completion method can significantly improve the detection accuracy at a distance. Compared with VoxelRCNN [5], DSaPG improve +3.71%, +0.86% and +5.53% Ap for car, pedestrian and cyclist over 40m distance. This is due to the generated virtual points enhancing the contour of faraway objects, thereby facilitating their detection. Meanwhile, compared with the other two point cloud completion methods, DSaPG improves +0.93%, +3.97% and +2.29% Ap within the specific distance-range of 30-40 meters from LiDAR sensor. This is due to the geometric RPN module, the density-aware of the original point cloud and the deformation learning of the generated point.

Table 2: Performance breakdown over different occlusion levels.

Class	Car			Pedestrian			Cyclist		
Occlusion	Level-0	Level-1	Level-2	Level-0	Level-1	Level-2	Level-0	Level-1	Level-2
VoxelRCNN	92.38	79.24	56.73	67.75	28.21	7.88	91.19	25.53	2.03
PGRCNN	92.73	80.07	57.25	68.44	34.23	8.94	93.84	30.06	2.53
DSaPG(Ours)	**93.08**	**81.21**	**57.92**	**72.07**	**35.78**	**9.09**	**94.19**	**31.24**	**2.73**

1.5 Experimental Analysis under different levels of occlusion

We compare DsaPG with other detectors on different occlusion levels. The results shown in Table 2. For car detection, our DSaPG achieves higher accuracy for highly occluded objects. For the two difficult detection categories of cyclist and pedestrian, DSaPG still brings consistent and significant improvements on different levels even in extremely difficult cases.

1.6 Experimental Analysis of hyperparameter values

Figure 2 displays the ablation study of hyperparameters λ and ω. The observation indicates that both λ and ω reach their peak performance at 1.

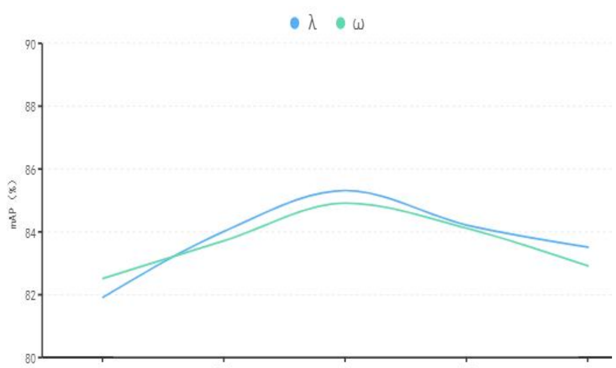

Fig. 2: λ and ω ablation studies

1.7 More Visualization Results

In order to specifically observe the detection performance of our proposed DSaPG, we visualize the point generation experimental results of 3D object detection on the public benchmarks comparing the result of PGRCNN [3]. We also visualize the effect of the addition of different modules(Geometry-guided RPN, Density-aware point generation, Deformation learning) on the generation of point clouds.

Analysis on Point Generation Results Here, we compare the qualitative results of the proposed method on KITTI val data with a previous point cloud completion method, PGRCNN [3]. Fig.3 illustrates some of the point generation

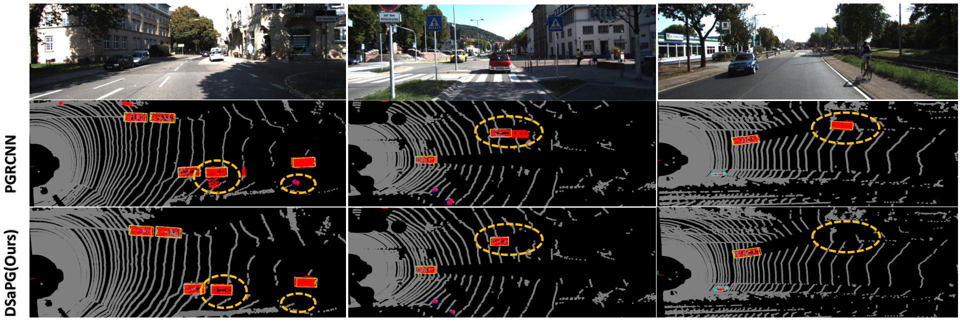

Fig. 3: Illustration of the point generation and 3D detection results on the KITTI validation set of PGRCNN and DSaPG(Ours). The green, cyan, and blue boxes are prediction boxes for cars, cyclists, and pedestrians, respectively.

and detection results of DSaPG and PG-RCNN. The first line show the each point cloud scene. Other two rows of Fig.3 display the outputs in a bird's-eye-view. We marked the inadequacies in PGRCNN [3] with yellow dotted boxes, It can be clearly seen that in some scenes, PGRCNN often results in wrong object recognition and point cloud generation due to the lack of prior guidance and reasonable geometric perception. In contrast, DSaPG can recover accurate object shapes at reasonable locations. In the third column of the second row, PGRCNN [3] causes false detection and unreasonable point cloud generation. In contrast, our method achieves reasonable recognition, thanks to the fact that our geometry-guided RPN module can generate accurate proposals. In general, DSaPG(ours) can faithfully recover the shape of the object in a reasonable position and improve the detection performance.

Effect of geometry-guided RPN on point cloud completion Fig.4 visualizes the effect of different RPN on point cloud completion in some cases. We respectively use the RPN module of VoxelRCNN [5] and our proposed geometric guided RPN module for training. The first row shows that the original RPN module tends to produce a large orientation deviation from the GT box, despite the presence of foreground supervision. However, the direction of the generated points is still biased towards the direction corresponding to the proposal, resulting in wrong completion. The second row shows the point cloud completion results of geometry-guided RPN. Although it is not guaranteed that all initial boxes can have reasonable orientation alignment, due to our orientation supervision and heat map supervision, the orientation of initial boxes can be roughly guaranteed to ensure the rationality of generated points.

Effect of deformation learning on point cloud completion Figure5 visualizes the impact of deformation learning on point cloud completion, which generates an offset for each generated point through the foreground score of the generated point, which helps to make the location of the generated point reasonable as well as more accurate shape recovery. The red arrow in Figure5 implies that the offset direction of the generating point is the main assumption.

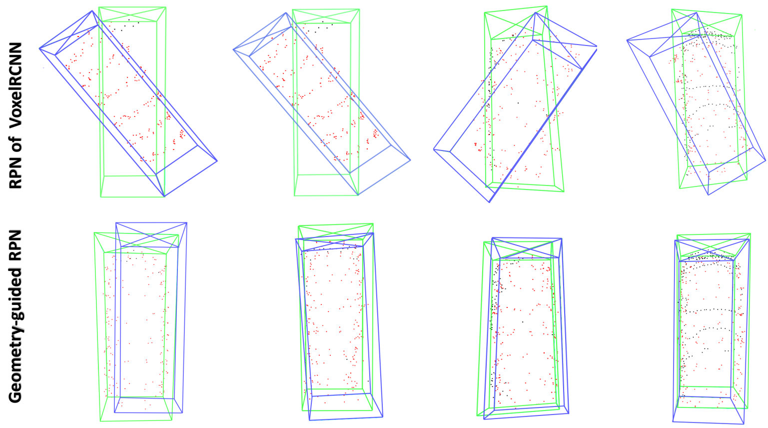

Fig. 4: Illustration of the impact of the RPN module on point cloud completion, we compare the geometry-guided RPN and VoxelRCNN-based RPN, where green, blue, and red represent GT boxes, initial proposal, and generated points, respectively.

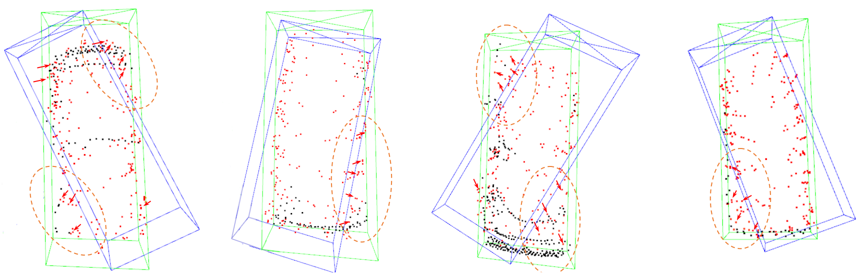

Fig. 5: Illustration of the influence of Deformation learning on point cloud generation. The point cloud in the figure is the normal generation result, the red arrow implies that the moving direction of the point is subjective conjecture, and the green, blue and red represent the GT box, the initial proposal and the generated point respectively.

It can be seen that due to the offset, the generating point can cross the proposal and fall in a reasonable position in the GT box.

Effect of Density-aware Deformable Point Generation on point cloud completion Fig. 6 and Fig. 7 visualize the output results of the density-aware deformable point generation module. Fig.6 shows the results of point cloud generation from a bird's eye view. The first shows the original point cloud, and the second shows the point cloud completed by DsaPG. It can be seen that our method can reasonably recover the missing shape of the object regardless of whether the object contains more points or fewer points, and the density distribution of concerns generates more uniform and intentional points to facilitate the acquisition of more meaningful spatial information. Fig. 7 intuitively shows the point cloud completion results of the rider and bicycle. The first is the original point cloud, and the second is the view of the completed point cloud. For small target objects, although it is sometimes difficult to form a reasonable shape

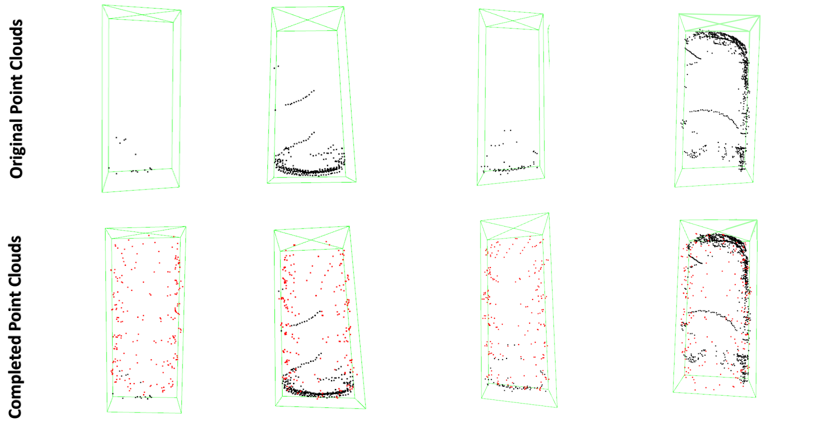

Fig. 6: Examples of completed point clouds in a bird's-eye-view, where green and red represe

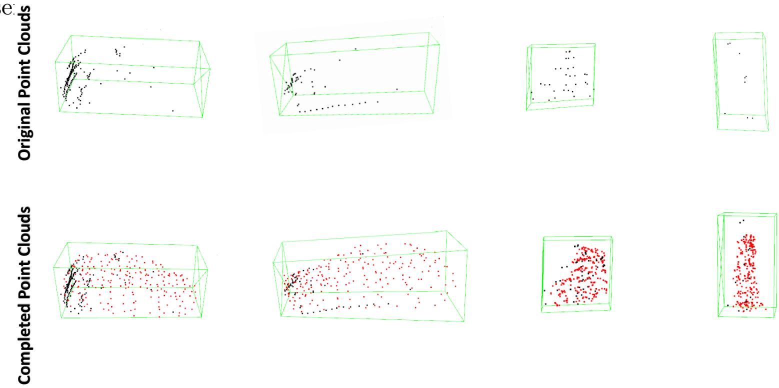

Fig. 7: Examples of completed point clouds for a car, pedestrian, and cyclist, where green and red represent GT boxes and generated points, respectively

due to too few points, the generated points still retain the exact position. All the results prove the effectiveness of our point cloud completion method, which not only focuses on the location of the generated points, but also ensures the effectiveness of the generated points for object shape recovery.

References

1. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012. 1
2. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020. 2

3. I. Koo, I. Lee, S.-H. Kim, H.-S. Kim, W.-j. Jeon, and C. Kim, "Pg-rcnn: Semantic surface point generation for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18142–18151, 2023. 2, 3, 4, 5
4. Q. Xu, Y. Zhong, and U. Neumann, "Behind the curtain: Learning occluded shapes for 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 2893–2901, 2022. 3
5. J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 1201–1209, 2021. 3, 5
6. Z. Li, Y. Yao, Z. Quan, J. Xie, and W. Yang, "Spatial information enhancement network for 3d object detection from point cloud," *Pattern Recognition*, vol. 128, p. 108684, 2022. 3