Beyond Virtual Points: Depth-Enhanced LiDAR-only 3D Object Detection with Semi-Supervised Learning (Student Abstract)

Jisu Kang [*] [†], **Wooseok Shin, Jin Sob Kim, Hyun Joon Park, Yujin Ham, Sung Won Han** [‡]

Department of Industrial and Management Engineering, Korea University,
145, Anam-ro, Seongbuk-gu, Seoul 02841, South Korea
{ji_soo_o, wsshin95, jinsob, winddori2002, yjham, swhan}@korea.ac.kr

Abstract

The task of 3D object detection is crucial for various applications that rely on identifying objects in three-dimensional space using inputs like LiDAR point clouds and images. However, LiDAR-based detection faces challenges due to the sparsity of point clouds, especially at greater distances. To address this, depth completion models have been used to generate virtual points from RGB images, but they struggle with real-time applications due to high computational costs. Our work eliminates the depth completion process, significantly improving processing speed while minimizing performance degradation. Consequently, our method has achieved an optimal balance between speed and accuracy on the KITTI leaderboard.

Introduction

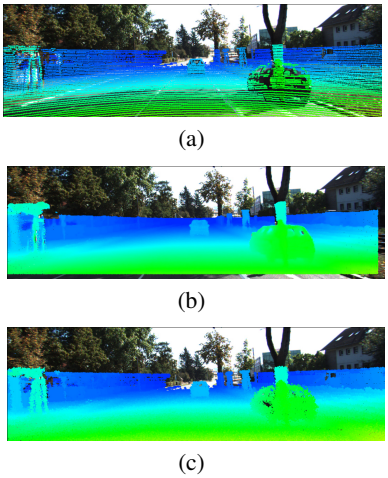
The task of 3D object detection involves identifying objects in three-dimensional space using inputs like LiDAR point clouds and images. Early models relied solely on LiDAR data, but this single-modality approach showed performance limitations, particularly in environments where LiDAR point clouds were sparse. This led to the development of multimodal techniques that integrate data from sources like RGB images and virtual points. VirConv (Wu et al. 2023), a representative multimodal model, generates virtual points from RGB images and achieves state-of-the-art performance. Specifically, VirConv follows a two-stage process, with the depth completion process in the first stage and object detection process in the second stage. While methods have been proposed to reduce the number of virtual points and improve processing speed (Wu et al. 2023), these approaches assume the availability of virtual points, which undermines real-time operation.

In this study, we aim to eliminate the depth completion process used in previous research and devise alternative methods to replace it. The contributions of this research are significant in two main aspects. First, we achieve a substantial speed improvement by eliminating the need for a depth completion process, exceeding a twofold increase. Second, we propose a solution named depth enhancement to mitigate the performance degradation that could arise from the absence of the depth completion process.

Methodology

Depth Enhancement

The depth enhancement method computes percentiles of the input data, achieving a similar effect to using virtual points in depth completion but with faster processing speed, serving as a simple yet effective replacement for the depth completion model. Specifically, depth enhancement targets only the k farthest LiDAR points within the sample points. For each of these k distant points, the process involves adding n random noise points repeatedly, with distribution range σ. This approach aims to enhance the density of the point cloud in the regions farther from the sensor. Enriching the point cloud at greater distances improves the model's ability to process detailed spatial information, ultimately leading to better detection performance. As shown in Figure 1(c), applying depth enhancement results in a denser point cloud, similar to that achieved with virtual points. Compared to the



(a) (b) (c)

Figure 1: (a) LiDAR point clouds (b) virtual point clouds, and (c) LiDAR point clouds with depth enhancement

[*]Phone number; +82-(0)-10-6428-9072
[†]Supplementary e-mail; jisu.kang1997@gmail.com
[‡]Corresponding author.

Model	Car BEV AP				Car 3D AP				Time (sec)
	Easy	Mod.	Hard	Sum	Easy	Mod.	Hard	Sum	
VirConv-S	95.99	93.52	90.38	279.89	92.48	87.20	82.45	262.13	0.09
VirConv-T	96.11	92.65	90.38	279.14	92.54	86.25	81.24	260.03	0.09
TED	95.44	92.05	87.30	274.79	91.61	85.28	80.68	257.57	0.14
LoGoNet	95.48	91.52	87.09	274.09	91.80	85.06	80.74	257.60	0.11
SFD	95.64	91.85	86.83	274.32	91.73	84.76	77.92	254.41	0.08
CasA++	94.57	91.22	88.43	274.22	90.68	84.04	79.69	254.41	0.14
CLOCs	92.91	89.48	86.42	268.81	89.16	82.28	77.23	248.67	0.07
SE-SSD	95.68	91.84	86.72	274.24	91.49	82.54	77.15	251.18	0.02
BVPConv-L	95.27	91.49	88.93	275.69	91.38	84.40	80.07	255.85	0.01
BVPConv-T	95.24	91.75	89.15	276.14	91.59	84.83	80.38	256.80	0.05

Table 1: Experimental results on the KITTI leaderboard

depth completion process, the training time is significantly reduced when using depth enhancement. Furthermore, since depth enhancement is not applied during inference, it does not introduce any additional computational overhead.

Semi-Supervised Learning

Semi-supervised learning was applied by training the student model with pseudo-labels produced by a teacher model that incorporated both LiDAR data and virtual points. The parameters of the teacher model are kept fixed during this process, which ensures that the generation of labels is carried out independently of the student model's training.

Experiments

Datasets and Evaluation Metrics

The KITTI 3D object detection dataset (Geiger, Lenz, and Urtasun 2012) includes 7,481 LiDAR and image frames for training and 7,518 frames for testing. Testing was conducted on the official KITTI leaderboard. For semi-supervised training, we sampled 10,888 frames in KITTI odometry dataset (Geiger, Lenz, and Urtasun 2012). Evaluation was performed using standard metrics: bird's eye view(BEV) and 3D average precision(AP) with 40 recall thresholds (R40) across easy, moderate, and hard levels.

Implementation Details

Models were trained on dual NVIDIA RTX A6000s using the ADAM optimizer with a learning rate of 0.01 and a one-cycle strategy for 60 epochs. An NMS threshold of 0.8 was used to generate 160 object proposals with a 1:1 positive-to-negative sample ratio. An NMS threshold of 0.1 was applied to remove redundant boxes for testing. The backbone network, VirConv, used feature dimensions of 16, 32, 64, and 64, with depth enhancement hyperparameters set to $n = 20$, $\sigma = 0.05$ and $k = 50$. Inference times were reported per sample under consistent conditions.

Results on the Test Set

We trained the VirConv-T and VirConv-L with depth enhancement and semi-supervised learning and referred to them as BVPConv-T and BVPConv-L. The experimental results on the test data are shown in Table 1. The proposed methods in this study demonstrated both excellent performance and speed within LiDAR-only environments. Notably, BVPConv-L achieved the fastest inference time of just 0.01 seconds per sample while maintaining robust performance. This underscores BVPConv-L's exceptional efficiency and effectiveness, making it highly suitable for real-time applications where both speed and accuracy are critical. On the other hand, BVPConv-T had a slightly longer inference time compared to BVPConv-L due to the use of the transformed refinement scheme. However, it still achieved approximately twice the score when accounting for both inference time and performance compared to VirConv-S.

Overall, the methods proposed in this study not only mitigate the performance degradation typically associated with the absence of depth completion models but also provide a substantial speed advantage. Among the current methods listed on the KITTI leaderboard, our approach stands out as the most effective in balancing performance and speed.

Conclusion

In this study, we explored methods to replace virtual points in 3D object detection. Our findings reveal that integrating depth enhancement with semi-supervised learning leads to significant improvements in both performance and computational efficiency. The proposed methods not only enhance detection accuracy but also reduce the inference time, making them well-suited for real-time applications. Notably, on the KITTI leaderboard, the models BVPConv-L and BVPConv-T, which incorporate both semi-supervised learning and depth enhancement, demonstrated strong performance. Among these, BVPConv-L particularly excelled, offering the most favorable balance between detection performance and processing speed. Future work will explore knowledge distillation and depth enhancement to improve the 3D object detection performance.

Acknowledgements

This study was conducted as part of the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and the National Research Foundation of Korea.

References

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving. In *The KITTI Vision Benchmark Suite, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361.

Wu, H.; Wen, C.; Shi, S.; Li, X.; and Wang, C. 2023. Virtual sparse convolution for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21653–21662.