Do You Remember . . . the Future? Weak-to-Strong generalization in 3D Object Detection

Alexander Gambashidze[1,2], **Aleksandr Dadukin**[2], **Maxim Golyadkin**[1,2], **Maria Razzhivina**[2] and **Ilya Makarov**[1,3]
[1]Artificial Intelligence Research Institute
[2]HSE University
[3]ISP RAS Research Center for Trusted Artificial Intelligence
{amgambashidze, aodadukin, mvrazzhivina}@edu.hse.ru, {gambashidze, golyadkin, makarov}@airi.net

Abstract

This paper demonstrates a novel method for LiDAR-based 3D object detection, addressing major field challenges: sparsity and occlusion. Our approach leverages temporal point cloud sequences to generate frames that provide comprehensive views of objects from multiple angles. To address the challenge of generating these frames in real-time, we employ Knowledge Distillation within a Teacher-Student framework, allowing the Student model to emulate the Teacher's advanced perception. We pioneered the application of weak-to-strong generalization in computer vision by training our Teacher model on enriched, object-complete data. In this demo, we showcase the exceptional quality of labels produced by the X-Ray Teacher on object-complete frames, showing our method distilling its knowledge to enhance object 3D detection models.

1 Introduction

In the rapidly advancing fields of computer vision and autonomous driving, 3D object detection is crucial for safe vehicle navigation and interaction with the environment. LiDAR, with its detailed 3D environmental data capture, stands out among sensing technologies. However, LiDAR data faces challenges like sparsity and occlusion, affecting 3D detection efficiency. Sparsity results from LiDAR's point cloud data, which, despite its detail, often misses the continuous coverage seen in camera images or self-supervised learning based reconstructed dense depth maps [Karpov and Makarov, 2022; Indyk and Makarov, 2023; Luginov and Makarov, 2023]. Occlusions further complicate detection, as objects can be hidden by obstacles.

Our X-Ray Teacher framework [Gambashidze *et al.*, 2024] offers a novel solution by utilizing the temporal dimension of LiDAR data to construct Object-Complete frames from multiple viewpoints and use it for weak-to-strong knowledge distillation, presented on Figure 1. This approach effectively mitigates sparsity and occlusion, enabling the distillation of comprehensive object knowledge to our detection system.

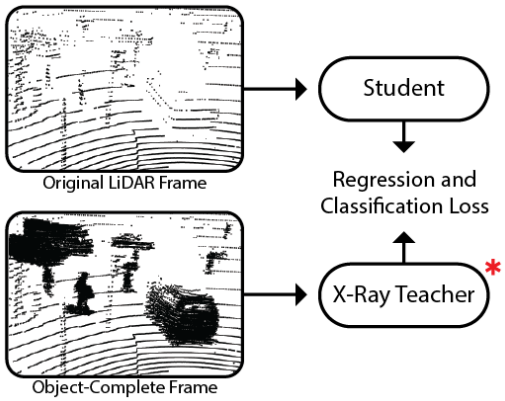

Figure 1: We radically simplify point clouds with object completion to train the X-Ray teacher that provides much more accurate predictions if used on object-complete frames. This allows us to overcome baseline models via distillation in the supervised setting.

Central to our innovation is the Teacher-Student framework, where the Teacher model, trained on Object-Complete frames, imparts its enhanced environmental understanding to the Student model. This knowledge transfer significantly improves detection performance, especially since real-time generation of Object-Complete frames is impractical due to the need for future viewpoints.

Validated across leading autonomous driving datasets, our method integrates seamlessly with any model, consistently boosting performance of supervised models. In this demonstration, we highlight the Teacher model's precise predictions and the X-Ray Student's ability to surpass original base models, setting new accuracy and reliability standards in 3D object detection.

2 Related Works

In the realm of 3D object detection, the field has evolved from directly processing point clouds with foundational models like PointNet and PointNet++ [Qi *et al.*, 2017a; Qi *et al.*, 2017b], to adopting voxel-based representations for efficiency, leveraging 3D sparse convolutions [Yang *et al.*, 2018; Yin *et al.*, 2021; Zhou *et al.*, 2022; Xu *et al.*, 2022], and incorporating advanced techniques like modified self-attention

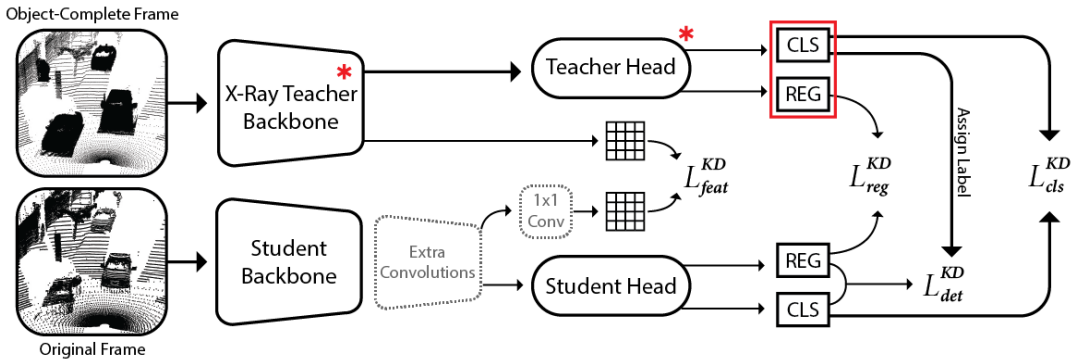

Figure 2: Overall X-Ray Knowledge Distillation for Supervised Learning.The idea is to make the Student model mimic the X-Ray Teacher's behaviour like it also sees Object-Complete frames.

layers for enhanced performance [Wang *et al.*, 2023].

The concept of Knowledge Distillation (KD), introduced by Hinton et al. [Geoffrey Hinton, 2015], employs a teacher-student model to transfer knowledge, typically focusing on logits/regression distillation and feature map distillation. Unlike conventional KD applications in 3D detection that prioritize efficiency [Cho *et al.*, 2023; Yang *et al.*, 2022], our X-Ray Teacher model adopts a hybrid distillation approach to enhance detection accuracy, setting a new precedent for leveraging KD to surpass state-of-the-art performance.

3 Methodology

Our framework, designed for easy integration with any object detection model, leverages sequential LiDAR data for innovative enhancements. We hypothesized that training on object-complete data-aggregating points from all sequence frames-would significantly improve predictions and supervised distillation. Our X-Ray Teacher, a weaker network, effectively guides a more robust student model to higher accuracy. Validation on the NuScenes dataset showed the object-complete model achieving 79.5% mAP (trained and validated on object-complete frames), while baseline model has 59.2% mAP (trained and validated on original clouds).

3.1 Object Complete Frames

To make point clouds much simpler and address the occlusion and sparsity problems, we need to aggregate all possible information about each object in each point cloud from all frames in a sequence. In the supervised setting the task is quite clear: all objects have their own instance IDs and well annotated boxes, so we only need to iterate over all frames and find all appearances of an object that we currently have processed.

3.2 X-Ray Teacher

It is evident that a model trained on original point clouds will not perform better on object-complete frames. In the case of the NuScenes dataset, we validated our original CenterPoint model on an object-complete validation set and achieved only a 31.6% mAP score, which confirms the necessity of training X-Ray Teachers either from scratch or by fine-tuning them from an original pretrained checkpoint. This underscores the critical importance of the X-Ray Teacher stage where we train teacher models from scratch. This step transforms the Teacher model into a weaker one that guides the stronger (student) model to make even better predictions.

3.3 Knowledge Distillation in Supervised Setting

Having prepared object-complete point clouds, we proceeded to train the baseline model (Student) to minimize Knowledge Distillation losses, aligning it with the X-Ray Teacher's insights on complete data. Distillation involves matching the Teacher's and Student's outputs across several dimensions: backbone encoder embeddings, bounding box regression labels, class distributions for classification tasks, and intermediate features from regression and classification heads before label assignment. The pipeline is shown on Figure 2.

The distillation losses are defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{heads}}^{KD} &= \alpha_1 \mathcal{L}_{\text{reg}}^{KD} + \alpha_2 \mathcal{L}_{\text{cls}}^{KD} = \\ &= \alpha_1 D_{KL}(S_{\text{cls}} \| T_{\text{cls}}) + \alpha_2 \text{MSE}(S_{\text{reg}}, T_{\text{reg}}) \quad (1)\end{aligned}$$

$$\mathcal{L}_{\text{feat}}^{KD} = \text{MSE}(T_{\text{back}}, \phi(\omega(S_{\text{back}}))) \quad (2)$$

$$\mathcal{L}_{\text{det}}^{KD} = \mathcal{L}_{detection}(S_{preds}, \tilde{T}_{boxes}) \quad (3)$$

Here, T and S represent the Teacher and Student model outputs, respectively, with S processing the original frame F and T handling the Object-Complete Frame $\tilde{F}$. The terms S_{back} and T_{back} denote the backbone outputs, while S_{reg}, T_{reg}, and S_{cls}, T_{cls} correspond to the regression and classification outputs. $\tilde{T}_{boxes}$ are the Teacher's predicted boxes, and S_{preds} is the Student's overall output. The parameters α_1, α_2 are weights for the loss components, and ϕ and ω adjust feature map dimensions and flexibility, respectively. This approach ensures the Student model learns to extract and utilize rich information from less detailed data, mirroring the Teacher's advanced detection capabilities.

4 Experiments

4.1 Implementation Details

In this demonstration we show results on Waymo Open Dataset [Sun *et al.*, 2020] and NuScenes [Caesar *et al.*, 2020].

For the Waymo dataset, we have refined model architectures due to its longer sequences, which resulted in highly

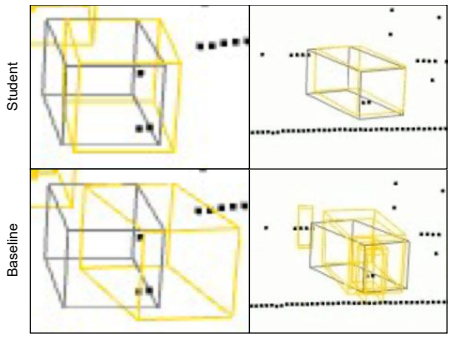

Figure 3: We show typical cases on NuScenes Validation set where student model works significantly better on sparse and occluded objects. Black color denotes ground truth, yellow represents predictions.

Model	mAP/mAPH L1	mAP/mAPH L2	#params
SECOND	67.2/63.1	61.0/57.2	5.3m
X-Ray SECOND	67.0/62.8	60.4/56.7	5.3m
SECOND-Scaled	66.8/62.7	59.4/56.1	6.2m
X-Ray SECOND-Scaled	**68.3/64.3**	**61.9/58.0**	6.2m
CenterPoint	74.4/71.7	68.2/65.8	8.3m
X-Ray CenterPoint	73.2/69.7	67.1/64.5	8.3m
CenterPoint-Scaled	74.1/71.5	67.9/65.3	9.2m
X-Ray CenterPoint-Scaled	**75.2 /72.1**	**68.9/66.3**	9.2m
DSVT Pillar	79.5/77.1	73.2/71.0	8.6m
X-Ray DSVT Pillar	79.2/76.7	72.6/70.3	8.6m
DSVT Pillar-Scaled	79.6/77.2	73.3/71.2	9.5m
X-Ray DSVT Pillar-Scaled	**80.1/77.9**	**73.7/71.4**	9.5m

Table 1: Our method performance on Waymo Validation dataset. We scale student models on this dataset, because Waymo object-complete frames are much more informative compared to NuScenes so we need our student models to be more complex to match much simpler feature maps.

dense complete objects post our object completion procedure, see example in Figure 4. Given that the data for the X-Ray Teacher is significantly simpler, we increased the complexity of the student network to better match the teacher's feature maps distribution. The same principle does not hold for NuScenes as it has less informative object-complete clouds. We used default configuration files for all runs.

4.2 Results

In this demo, we introduce our novel plug-and-play framework for LiDAR-based 3D object detection. Our method demonstrates extreme robustness and improves all models we have tested so far, including the previous state-of-the-art, DSVT [Wang *et al.*, 2023]. We validated our method on the two most popular datasets: Waymo and NuScenes, see results

Model	mAP	NDS
CBGS	50.0	59.2
X-Ray CBGS (ours)	**50.8**	**60.4**
CenterPoint-Voxel	53.4	61.3
X-Ray CenterPoint-Voxel (ours)	**54.3**	**62.9**

Table 2: Our method performance on NuScenes validation set.

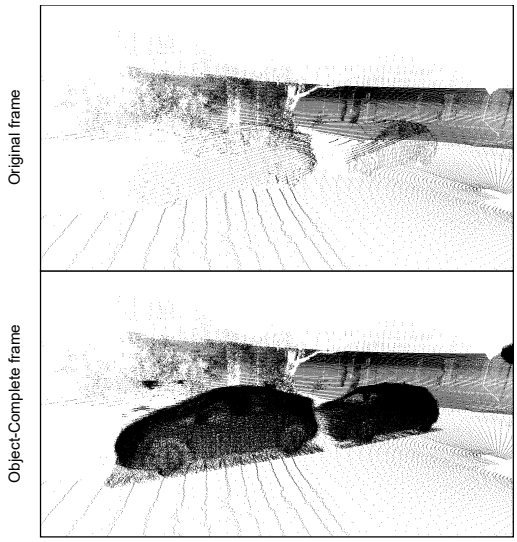

Figure 4: Comparison of object-complete and original frames, Waymo Open dataset.

in Table 1 and Table 2, respectively.

We also show visual side-by-side comparison of the framework performance on NuScenes at Figure 3.

5 Conclusion

In this demonstration, we highlight the capability of combining object points from different frames to create complete object frames. This approach seamlessly integrates knowledge distillation into any 3D object detection pipeline, functioning as a plug-and-play solution. Our results reveal that a Teacher model, trained and validated on object-complete frames delivers highly accurate predictions. The Teacher model's behavior enables the student model to focus on occluded objects and gain a better understanding of sparsely represented ones.

Ethical Statement

Our research on LiDAR-based 3D object detection presents an improvement in detection metrics by 2-3%. While significant within our domain, this advancement requires extensive hyperparameter tuning, leading to increased computational demands and potential environmental impacts.

We utilized open datasets from Waymo and NuScenes, with all privacy concerns managed by the dataset providers. This approach ensures our adherence to data privacy standards and contributes to the reproducibility of our work.

A notable consideration is the impact of autonomous driving technologies on employment within the transportation sector. As these technologies advance, particularly through enhancements like ours, the demand for human drivers may decrease, posing socioeconomic challenges.

In summary, our work seeks to push forward the capabilities of autonomous driving technologies while acknowledging the ethical considerations of increased computation and the potential societal impacts on employment. We emphasize the importance of responsible technological advancement and the need for a balanced approach to innovation.

Acknowledgements

This research was supported in part through computational resources of HPC facilities at HSE University.

The work of I. Makarov on 3D object detection related work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of dated November 2, 2021 No. 70-2021-00142.

References

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Cho *et al.*, 2023] Hyeon Cho, Junyong Choi, Geonwoo Baek, and Wonjun Hwang. itkd: Interchange transfer-based knowledge distillation for 3d object detection. In *CVPR*, 2023.

[Gambashidze *et al.*, 2024] Alexander Gambashidze, Aleksandr Dadukin, Maksim Golyadkin, Maria Razzhivina, and Ilya Makarov. Weak-to-strong 3d object detection with x-ray distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–15, 2024.

[Geoffrey Hinton, 2015] Jeff Dean Geoffrey Hinton, Oriol Vinyals. Distilling the knowledge in a neural network. In *arXiv preprint arXiv:1503.02531*, 2015.

[Indyk and Makarov, 2023] Ilia Indyk and Ilya Makarov. Monovan: Visual attention for self-supervised monocular depth estimation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1211–1220. IEEE, 2023.

[Karpov and Makarov, 2022] Aleksei Karpov and Ilya Makarov. Exploring efficiency of vision transformers for self-supervised monocular depth estimation. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 711–719, 2022.

[Luginov and Makarov, 2023] Albert Luginov and Ilya Makarov. Swiftdepth: An efficient hybrid cnn-transformer model for self-supervised monocular depth estimation on mobile devices. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 642–647. IEEE, 2023.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[Sun *et al.*, 2020] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Wang *et al.*, 2023] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. Dsvt: Dynamic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13520–13529, June 2023.

[Xu *et al.*, 2022] Qiangeng Xu, Yiqi Zhong, and Ulrich Neumann. Behind the curtain: Learning occluded shapes for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36 (3), pages 2893–2901, 2022.

[Yang *et al.*, 2018] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018.

[Yang *et al.*, 2022] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. In *NeurIPS*, 2022.

[Yin *et al.*, 2021] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[Zhou *et al.*, 2022] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022.