Weak-to-Strong 3D Object Detection with X-Ray Distillation

Alexander Gambashidze*[1,2] Aleksandr Dadukin[2] Maxim Golyadkin[1, 2]
Maria Razzhivina[2] Ilya Makarov[1,3]

Abstract

This paper addresses the critical challenges of sparsity and occlusion in LiDAR-based 3D object detection. Current methods often rely on supplementary modules or specific architectural designs, potentially limiting their applicability to new and evolving architectures. To our knowledge, we are the first to propose a versatile technique that seamlessly integrates into any existing framework for 3D Object Detection, marking the first instance of Weak-to-Strong generalization in 3D computer vision. We introduce a novel framework, X-Ray Distillation with Object-Complete Frames, suitable for both supervised and semi-supervised settings, that leverages the temporal aspect of point cloud sequences. This method extracts crucial information from both previous and subsequent LiDAR frames, creating Object-Complete frames that represent objects from multiple viewpoints, thus addressing occlusion and sparsity. Given the limitation of not being able to generate Object-Complete frames during online inference, we utilize Knowledge Distillation within a Teacher-Student framework. This technique encourages the strong Student model to emulate the behavior of the weaker Teacher, which processes simple and informative Object-Complete frames, effectively offering a comprehensive view of objects as if seen through X-ray vision. Our proposed methods surpass state-of-the-art in semi-supervised learning by 1-1.5 mAP and enhance the performance of five established supervised models by 1-2 mAP on standard autonomous driving datasets, even with default hyperparameters. Code for Object-Complete frames is available here: https://github.com/sakharok13/X-Ray-Teacher-Patching-Tools.

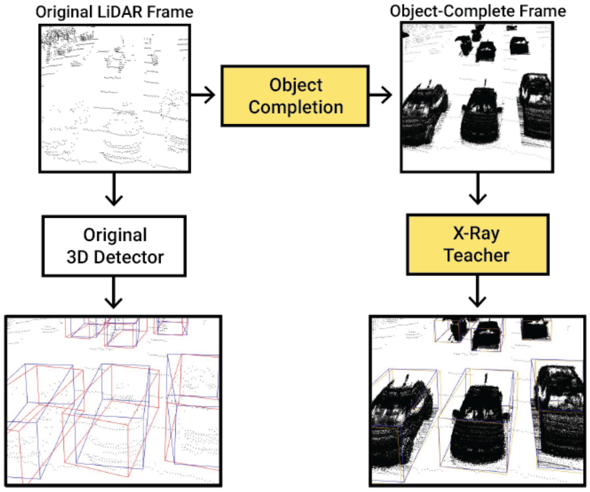

Figure 1. 3D object detection directly from sparse LiDAR data (top left) provides noisy predictions (bottom left). Adding object completion stage (top right) helps to train 3D object detection X- Ray Teacher, which is robust and can be distilled to baseline model. Red, Yellow and Blue colors of bounding boxes are related to classical LiDAR-based object detection, Our model on Object-Complete frames predictions and Ground Truth labels, respectively.

> You're just not thinking fourth dimensionally... *the bridge will exist.*
>
> *Dr. Emmett Brown, Back to the Future*

1. Introduction

3D object detection is a fundamental task in the field of computer vision and autonomous systems, playing a key role in the advancement of self-driving technology [1] and contributing significantly to the robotics industry [38, 51]. Currently, LiDAR-based 3D object detection [6, 17, 36] demonstrates superior performance compared to camera-

*Corresponding author, email: alexandergambashidze@gmail.com
[1] Artificial Intelligence Research Institute
[2] HSE University
[3] ISP RAS

based [19, 39, 40, 45] and radar-based [3, 8, 22, 25] approaches. Furthermore, LiDAR point clouds are the key ingredient of multimodal fusion-based approaches [14, 21, 37], so LiDAR-based 3D detection continues to be a strong focus of the research community.

The point cloud-based 3D object detection challenges the following issues: sparsity, occlusions, and the complexity of 3D data annotation. **Sparsity**: large point clouds are sparse due to the LiDAR sensing process's inherent characteristics that leads to an imprecise representation of the captured scene. Additionally, there is an imbalance in point density. In particular, the point cloud is sparser in the far range and contains less spatial information affecting feature representation and box prediction. **Occlusions**: another problem arises from the frequent occurrence of partial occlusion in LiDAR frames. This primarily happens due to the fact that the frames are acquired from a single fixed point of view, making them essentially 2.5D. To detect and accurately locate a highly occluded object, a detector must recognize the hidden shapes of the object even when a significant portion of its parts is missing. Since the absence of certain shapes inevitably affects object perception, this becomes a critical detection challenge. Our previous approaches focused on depth completion [23], depth inpainting [30] or self-supervised depth pretraining [15, 16] could not address occlusion problems. **Data annotation**: finally, data annotation is a formidable challenge in 3D object detection due to the complexity of annotating objects in three-dimensional space. For example, a skilled annotator can spend weeks annotating just one hour of LiDAR data [9, 24]. This problem is partially addressed by semi-supervised learning approaches in a teacher-student framework [33, 35, 46, 48]. However, the quality of pseudo labels and the performance of these methods are limited due to the aforementioned sparsity and occlusion problems. Therefore, overcoming these inherent challenges is essential to improve the accuracy and reliability of 3D detection systems.

The challenges posed by sparsity and occlusion have led to the formulation of several methodologies. Some of them [42] are aimed at improving the computational efficiency of processing sparse data without improving the quality of the predictions. Others, including our work, focus on improving detection performance on sparse occluded data [18, 20, 26, 41]. Li et al. [18] enforce shape constraints to improve object localization with explicit shape priors obtained from a database of CAD models. Najibi et al. [26] and Li et al. [20] do the same by implicitly introducing priors with novel modules pretrained for point cloud completion and SDF approximation. Xu et al. [41] propose a Shape Occupancy Probability estimation module to refine predicted bounding boxes for occluded objects. However, all of these mentioned approaches *utilize new modules and adapt the model architecture to take advantage of them*. As a result, these methods have limited applicability to new architectures that will emerge as the field evolves.

In this work, *we address all three major challenges in 3D object detection* with our novel framework called *X-Ray Distillation* with Object-Complete Frames that is easily plugged in existing approaches and new architectures. It is designed for universal application to any LiDAR-based detector, improving performance on sparse and occluded objects. Our approach exploits the properties of existing large-scale autonomous driving datasets, which consist of sequences of LiDAR frames. Such property makes it possible to reconstruct complete shapes for occluded objects using other occurrences of these objects in the sequence, ensuring that all objects are equipped with points from all available viewpoints in a scene. We then use this completed data in the Teacher-Student framework for both semi-supervised learning and knowledge distillation in a supervised setting. We train our Teacher on extremely informative Object-Complete frames thus making it a weaker model [4]. Then, we use it to extract features from such simple Object-Complete frames and distills this knowledge to a stronger Student, which operates with the original data, to guide him on how to extract rich features from occluded objects. To generate Object-Complete frames, we leverage ground truth object tracking labels. Since there are no labels for the majority of data in the semi-supervised setting, we propose an Objects Temporal Fusion block to detect, track, and use point cloud registration techniques to construct Object-Complete frames.

We validate the proposed X-Ray Teacher framework on nuScenes [5] and Waymo Open Dataset [32] 3D object detection benchmarks for SECOND [42], CenterPoint [47], and DSVT [36] models in a supervised learning paradigm. Experiments show a steady improvement by 1-2 mAP with minimal or no impact on time and computational resources during the inference stage. For semi-supervised performance evaluation, we use ONCE [24] benchmark.

Applying these novel ideas for 3D object detection, we provide the following contributions:

1. We propose the X-Ray Teacher framework for semi-supervised learning, which achieves state-of-the-art performance on the ONCE benchmark.
2. We show that our approach improves the quality of four supervised learning models, including the current state-of-the-art model, and demonstrates the potential to improve the performance of any supervised model trained on sequential data.
3. We suggest the Objects Temporal Fusion block to generate Object-Complete frames for data that lacks ground truth tracking labels.

2. Related work

2.1. Object Detection for point clouds

Modern 3D detectors predominantly use point clouds, voxel representations, or a combination of both. The pioneering PointNet model family [28, 29] provided a significant step forward in 3D recognition. However, direct processing of point clouds poses several challenges: it requires point sampling, grouping, and computation of point-wise features, which can be computationally intensive. To integrate insights from 2D computer vision, a transition to a pixel equivalent, namely voxels, is required. Methods such as [41, 43, 47, 49] first convert point clouds to voxels, followed by the use of 3D sparse convolutions [12]. Recent advances also include the integration of modified self-attention layers, achieving state-of-the-art results [36].

2.2. Semi-supervised 3D detection

Although Semi-supervised 3D object detection is not as extensively researched as 2D detection, it has seen some significant contributions. Among these, SESS [48] stands out for using the general Mean Teacher approach with data augmentations and consistency loss. Similarly, 3DIoUMatch [35] is notable for its unique localization strategies and 3D IoU-guided techniques for box filtering. Proficient Teachers [46] is notable for its box voting and contrastive losses. We do not propose yet another standalone method for Semi-supervised 3D Object Detection; rather, our emphasis is on creating a plug-in technique designed to augment and enhance the performance of existing methods.

2.3. Knowledge Distillation

The concept of Knowledge Distillation (KD) was first introduced by Hinton et al. [10]. KD describes a learning approach where a larger teacher network guides the training of a smaller student network for various tasks [13]. Broadly, KD methodologies are categorized into two types: logits/regression distillation and feature map distillation. Our focus is on a hybrid approach, combining these methods: matching feature maps, regression, and classification heads with pseudo labels generated by the teacher model.

In 3D object detection, Knowledge Distillation is mainly utilized to minimize parameters and FLOPS while aiming to preserve box prediction quality [7, 44]. Yet, methods that aim to outperform state-of-the-art models are uncommon. In contrast to this trend, our X-Ray Teacher model challenges the norm by providing the quality improvement.

3. Methodology

3.1. Overview of X-Ray Teacher

We introduce a novel training framework to address the challenges of sparsity and occlusion in 3D Object Detection based on LiDAR data. This framework is not limited to any specific object detection model and has the potential for applications across various deep learning architectures. Our method is designed to process the LiDAR data structured as a sequence of frames.

The two core elements of our approach are Object Complete Frames Generation and Teacher-Student Knowledge Distillation. Our approach for 3D object detection can be applied in both, supervised and semi-supervised settings, with minor differences in the elements implementation.

Object Complete Frames Generation. In this step, we reconstruct the complete shapes of objects presented in the scene by utilizing information from the other frames within the same sequence. Given that autonomous driving datasets are composed of sequential data, we can efficiently leverage their temporal nature: we add points from both the future and the past when objects are observed from different viewpoints. It allows us to reconstruct the complete shapes of objects without shape databases or reconstruction modules.

In order to verify the validity of our approach, we trained a CenterPoint [47] model on both, the original and object-complete NuScenes [5] datasets, and then evaluated their performance on the respective validation sets. The models trained on the original and the object-complete frames achieved mAP scores of 59.2% and 79.5%, respectively. This difference of 20 mAP suggests that 1) it would be beneficial to transform unlabeled original point clouds into Object-Complete ones and to annotate them with a corresponding X-Ray Teacher pretrained on such informative frames 2) the features extracted by a weaker X-Ray Teacher might be distilled to a stronger student to share the knowledge of complete shapes.

Teacher-Student Knowledge Distillation. The necessity for this step arises because we cannot generate Object Complete Frames during the online inference stage, as it is not possible to access future data. Therefore, we need to encourage the model to behave as if it were observing shape-complete objects, even when dealing with occluded ones. A well-known method for enabling a deep learning model to imitate another model's behavior involves using Knowledge Distillation within a Teacher-Student framework. However, for conventional Knowledge Distillation, both the Teacher and Student models typically process data of the same complexity (the only difference may lie in the complexity of augmentations [31]). In contrast to the standard knowledge distillation, we enrich the data for training Teacher, which significantly improves its performance in 3D object detection. Then, we teach the Student to extract important information from less detailed data by distilling knowledge from the Teacher model.

Instead of simplifying the Student model, as it is usually done in standard Knowledge Distillation, we take an opposite approach and design the Student to be more complex

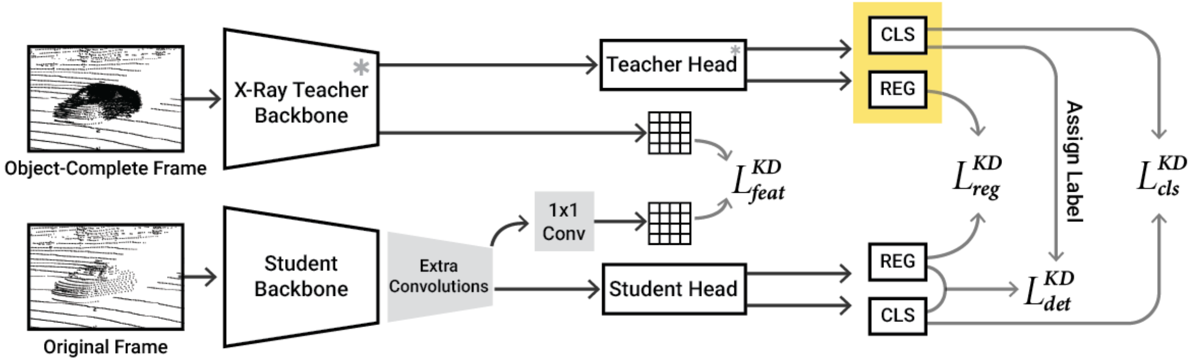

Figure 2. Overall X-Ray Knowledge Distillation for Supervised Learning. X-Ray Teacher is frozen and pretrained on Object-Complete frames which are taken as input. The Student is guided to mimic the Teacher's behaviour through Knowledge Distillation losses: L_{feat} for intermediate embeddings matching, L_{reg} for bounding box regression, L_{det} for basic detection, and L_{cls} for classification.

than the Teacher. It helps to extract high-quality information from more intricate and ambiguous data and demand a Student model to be more robust and to have a more complex receptive field capacity.

In what follows, we provide detailed descriptions of how we implement Object Complete Frames Generation and Teacher-Student Knowledge Distillation steps in both supervised and semi-supervised settings.

3.2. Supervised X-Ray Teacher

In the supervised setting of 3D Object Detection, models are trained and evaluated using datasets that provide labeled data, including precise bounding boxes and instance IDs. Object-Complete Frame Generation for labeled data involves aggregating objects based on their instance IDs and merging different views into a unified point cloud (see Section 4.2.2 for details).

For the distillation process (see Figure 2), we train Teacher model on Object-Complete frames and then freeze it. Then, we train baseline model (playing Student role) to directly minimize Knowledge Distillation losses inspired by [44]. The distillation is done by matching Teacher and Student backbone encoders' embeddings, output labels for bounding box regression, classes distribution for classification task (objects like pedestrians, cars, cyclists, etc.), and intermediate features obtained from the outputs of regression and classification heads before postprocessing (assigning labels). Specifically, we define the following losses:

$$\mathcal{L}_{\text{heads}}^{KD} = \alpha_1 \mathcal{L}_{\text{reg}}^{KD} + \alpha_2 \mathcal{L}_{\text{cls}}^{KD} \tag{1}$$

$$= \alpha_1 D_{KL}(S_{\text{cls}} \| T_{\text{cls}}) + \alpha_2 \text{MSE}(S_{\text{reg}}, T_{\text{reg}}) \tag{2}$$

$$\mathcal{L}_{\text{feat}}^{KD} = \text{MSE}(T_{\text{back}}, \phi(\omega(S_{\text{back}}))) \tag{3}$$

$$\mathcal{L}_{\text{det}}^{KD} = \mathcal{L}_{detection}(S_{preds}, \tilde{T}_{boxes}) \tag{4}$$

T and S are the outputs of our Teacher and Student models, respectively. The Student takes the original frame F as input, while the Teacher receives the Object-Complete Frame $\tilde{F}$, so $F \subset \tilde{F}$. S_{back} and T_{back} are referred to the output of the backbone module and S_{reg}, T_{reg}, S_{cls}, T_{cls} are the outputs of regression and classifications heads. $\tilde{T}_{boxes}$ are the X-Ray Teacher's predicted boxes after postprocessing. S_{preds} is an overall Student's output; α_1, α_2 are non-negative hyper-parameters. $L_{detection}$ is a basic detection loss that is used for training 3D object detection models. ϕ is a 1x1 Convolution to better match the Teacher's feature maps. ω refers to some extra convolutions to make the Student more flexible. We discovered that X-Ray Distillation does not need extra convolutions in the case of the NuScenes dataset, so their usage for encoder feature adjustment is also a hyperparameter of the model. By MSE, we mean Mean Squared Error.

Finally, the training objective can be written as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{heads}}^{KD} + \lambda_2 \mathcal{L}_{\text{feat}}^{KD} + \lambda_3 \mathcal{L}_{\text{det}}^{KD} \tag{5}$$

where λ_1, λ_2, λ_3 are non-negative hyper-parameters balancing the contribution of each term.

3.3. Semi-supervised X-Ray Teacher

Semi-supervised learning is characterized by the availability of a small amount of labeled data and a larger pool of unlabeled data, making the use of the Object Complete Frame generation approach proposed for supervised settings infeasible.

In order to overcome this limitation, we introduce the Objects Temporal Fusion block (as shown in Figure 3), which is designed to enable Object Complete Frame generation in situations where ground truth labeling is missing. This block leverages a model pretrained on labeled data to detect and track objects across unlabeled sequences. Subsequently, it employs Point Cloud Registration (PCR) to merge objects points from different views for all detected objects.

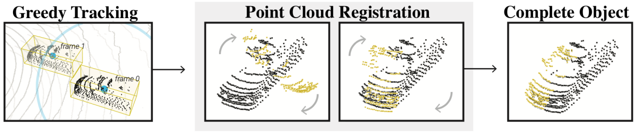

Figure 3. Object-Complete Frame generation process for semi-supervised setting. It consists of tracking and Point Cloud Registration. We track objects across all frames in the whole sequence, then we use Point Cloud Registration to merge points that represent the same object from different views, and finally we replace the original object with the new, complete one.

The provided steps outline a precise and detailed procedure for Objects Temporal Fusion:

1. to label all LiDAR frames using a pretrained model;
2. to greedily track objects across all frames in each sequence using predicted bounding boxes; assign unique IDs to every object instance to facilitate the identification of recurring objects within the scene and to organize them in a sequential manner;
3. for each object within each sequence, to merge different views of the same object by applying a deep learning model for Point Cloud Registration; this process generates a complete point cloud for each object, which is then used to replace the occluded one in each frame;
4. to fine-tune the base model on Object-Complete labeled frames; this refined model forms our X-Ray Teacher.

More details for greedy tracking and Point Cloud Registration models can be found in Section 4.2.2.

The Knowledge Distillation step presented in Figure 4 can be integrated with any semi-supervised 3D Object Detection approach that uses pseudo labeling as a form of self-distillation. All methods in the domain usually follow this paradigm, which underlines the high universality of our approach. From this perspective, the use of Object-Complete frames for pseudo label prediction refines the 3D bounding boxes and improves the quality of self-distillation. The sole limitation is that existing semi-supervised methods [33, 46] typically update the Teacher model's weights using the exponential moving average of the Student's weights. We refrain from this practice because our Teacher is finetuned on Object-Complete frames and thus predicts higher-quality labels (see proofs in Section 4.3).

4. Experiments

This section covers the datasets, implementation details, experiment results, and a detailed analysis of the different components that affect performance. It starts with Section 4.1, which defines the datasets and metrics used to prove the validity of our ideas. Section 4.2 delves into the network architectures and training parameters, while Section 4.4 analyzes the obtained metric values. Finally, Section 4.3 examines various parts of our approach and their influence on the overall performance.

4.1. Data

To evaluate our models, we use three large-scale autonomous driving datasets: NuScenes and Waymo Open Dataset for the supervised setting and ONCE for the semi-supervised setting.

NuScenes [5] is a popular outdoor dataset with diverse annotations for different tasks. It has 40,157 annotated samples. For 3D object detection, we provide NuScenes Detection Score (NDS) and mean Average Precision (mAP).

Waymo Open [32] is also one of the most popular outdoor 3D perception datasets. It contains 1150 point cloud sequences and has more than 200K total frames. All results are evaluated with 3D mean Average Precision (mAP) and its weighted variant (mAPH)

ONCE [24] is a large-scale dataset with 1 million pout cloud samples from LiDAR and only 15K annotated frames that are divided into train, val, and test with 5k, 3k, and 8k samples, respectively. This dataset is designed exactly for semi-supervised learning tasks and simulates real life: annotations are expensive and time-consuming. All not labeled frames are divided into Small (70 sequences), Medium (321 sequences), and Large (560 sequences) parts. We follow the ONCE Benchmark and use mAP over all classes with the 3D IoU thresholds 0.7, 0. 3, and 0.5 for classes ”Vehicle”, ”Pedestrian”, and ”Cyclist”, respectively.

4.2. Implementation Details

4.2.1 Network architectures

For supervised setting, we use SECOND [50], CenterPointVoxel [47], CBGS [50], DSVT [36] within our framework. The implementations of these models are based on the OpenPCDet [34] library, and we adhere to the default configurations suggested by this library for both training and inference. In addition, we present scaled versions of these models, which are essentially the original networks augmented with five additional convolutional layers stacked on top of the BEVEncoder with the following parameters 1x Conv(512, 128), 3x Conv(128, 128), and 1x Conv(128, 512) with Batch Normalizations and ReLU activations after each convolution. With the help of light Grid Search, we choose the following distillation hyperparameters: $\alpha_1 = 2$, $\alpha_2 = 1$, $\lambda_1 = 0.7$, $\lambda_2 = 0.3$, $\lambda_3 = 1$.

For semi-supervised setting, we follow the previous works [35, 46] and use SECOND and CenterPointVoxel models from OpenPCDet for validation of comparison. For training and inference, we use recently proposed refined configurations from work [11]. X-Ray Teacher model is fine-tuned on object-complete frames with the same hyperparameters for 10 epochs.

For computations, we use 4x A100 40GB GPU and AMD EPYC 7702 CPU.

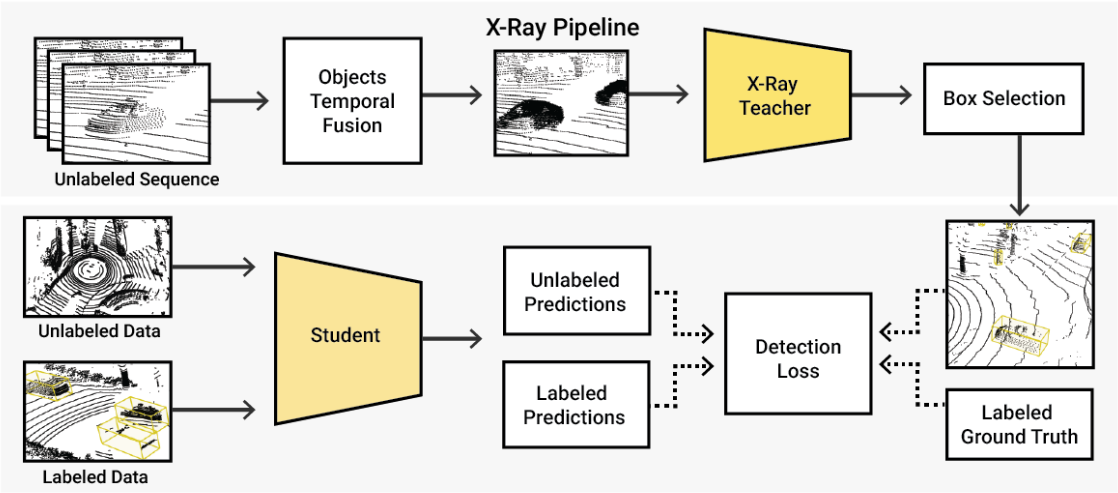

Figure 4. Semi-supervised X-Ray Teacher pipeline for 3D object detection task. Unlabeled sequences are processed by the Objects Temporal Fusion block to create more complete object representations by aggregating information over time. The Student model learns from both pseudo-labeled predictions and actual labeled data with ground truth annotations.

Table 1. Evaluation of the impact of teacher fine-tuning on Object-Complete frames in semi-supervised setting. The results indicate that teacher fine-tuning is essential.

Method	mAP
X-Ray MT SECOND	60.12
X-Ray PT SECOND	61.48
X-Ray MT SECOND* (ours)	65.26
X-Ray PT SECOND* (ours)	**68.43**
X-Ray MT CenterPoint	59.35
X-Ray PT CenterPoint	62.17
X-Ray MT CenterPoint* (ours)	67.60
X-Ray PT CenterPoint* (ours)	**70.55**

4.2.2 Object-Complete Frame Generation

The key concept of this paper - Object-Complete Frame Generation - combines diverse ideas to achieve optimal Object Completion, encompassing facets such as detection, tracking, and the registration of point clouds.

The detection phase is construed as an elective stage, exclusively implemented on unlabeled data. Within this stage discerned instances are encapsulated into generated 3D bounding boxes. E.g., a substantial proportion of instances within the ONCE dataset lacks accompanying labels.

Subsequently, the greedy tracking procedure traverses the entire set of frames, associating the appearance of instances at the i-th frame with potential candidate appearances in the following frame. The list of candidate instances is built by including all instances from the succeeding frame that fall within a prescribed radius, defined as twice the maximum dimension across their respective bounding boxes. The nearest instance is selected from the list, while the remaining instances are discarded. The series of matches made using this algorithm are combined into a single sequence, called track. When there are no more potential matches in the next frame, the track is considered as terminated. For the given instance in some specific frame there is the only corresponding track across the entire scene.

Those prepared instances are then used for Object Completion, a process executed through several sequential steps:

1. Point clouds associated with instances are extracted from their respective frames, they are translated back to the zero-point of the global basis. The rotation of the instances are also reset to identity.
2. The point clouds corresponding to instances within a common track undergo a merging process, constituting the Point Cloud Registration phase, wherein diverse set of merging approaches is used to glue them into a larger, densely populated point cloud.
3. Later on the corresponding point clouds are replaced with their respective densely populated point cloud from the previous step, restoring their original translation and rotation in a specific frame.

Table 2. Comparison of three distillation strategies: using only classification and regression heads matching, BEV features and heads matching, and the final pipeline that is the SECOND-Scaled model; experiment performed using Waymo Validation set.

Technique	mAP/mAPH L1	mAP/mAPH L2
Heads match	67.5/63.4	61.1/57.2
BEV & Heads	67.8/63.5	61.6/57.4
Full Pipeline (ours)	**68.3/64.3**	**61.9/58.0**

Various merging strategies were empirically tested in generating Object-Complete point clouds:

1. Geometry: assumes the objects (with boxes) are best aligned by the detected bounding box, performs both inverse translation and rotation to clear their geometric transformations and then merges intact point clouds.
2. GeDi: uses GeDi Point Cloud Registration method [27].
3. Greedy Grid: uses Greedy Grid implementation [2].

Object-Complete frames have the potential to become exceedingly large, so we employ a point subsampling strategy.

Table 3. Registration methods comparison for Object Complete Frame Generation on the ONCE Small split. We trained our best SECOND model in the semi-supervised setting with X-Ray Proficient Teacher on three different types of preprocessed data and compared results on ONCE validation set.

Method	**mAP**
Box Geometry	67.88
Greedy Grid [2]	68.17
GeDi [27]	**68.43**

Table 4. Comparison for supervised 3D Object Detection task on Waymo Open Dataset. We compare baseline models, their scaled versions and X-Ray distillation with default hyperparameters.

Model	mAP/mAPH L1	mAP/mAPH L2	#params
SECOND X-Ray Teacher*	85.1/70.3	75.1/64.7	5.3m
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 X-Ray Teacher*	88.3/78.6	76.4/72.9	8.3m
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 X-Ray Teacher*	89.3/79.7	79.1/73.4	8.6m
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

4.3. Ablation Studies

In this section, we provide the comparison of different components of our approach and show how they affect the performance. Specifically, we prove the usefulness of teacher fine-tuning and compare Point Cloud Registration and distillation methods. We also analyse if dealing with sparsity and incomplete shapes really improves the performance.

First, we performed a detailed comparison of the semi-supervised performance of our X-Ray method, both with and without teacher fine-tuning using Object-Complete frames. We use Mean Teacher and Proficient Teacher with SECOND and CenterPointVoxel models trained on ONCE Small split. Table 1 illustrates the substantial impact of teacher fine-tuning on performance, showing that neglecting this step results in a noticeable performance decline.

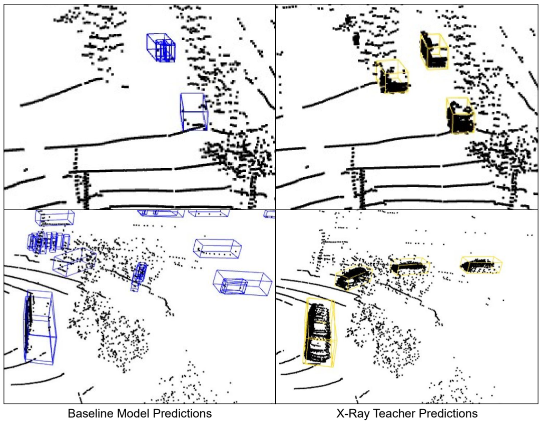

Figure 5. Visual comparison between noisy and poor baseline (original) 3D detector SECOND (left column) and our X-Ray Teacher that perceives Object-Complete frames. We compare two identical timestamps and view angles. The Baseline model fails to detect some objects while X-Ray Teacher does not. This explains why knowledge distillation is indeed beneficial and should improve models.

As we noted before, the reconstructed objects will not be perfect in the semi-supervised setting, which is precisely why we included Table 3. PCR models trained on domain objects like cars, pedestrians, cyclists, etc., should improve our method even more.

As we mentioned earlier, we combine several techniques for the knowledge distillation: BEV features matching, heads output matching with simple regression and KL divergence and detection loss on the teachers predictions. We also compare partial solutions on the Waymo validation set with a SECOND model in Table 2

Table 5. Comparison on NuScenes dataset. Our method improves baselines without scaling models due to the fact that NuScene's object-complete frames are less informative compared to Waymo.

Model	**mAP**	**NDS**
X-Ray Teacher* CBGS	77.1	72.4
CBGS	50.0	59.2
X-Ray CBGS (ours)	**50.8**	**60.4**
X-Ray Teacher* CenterPoint-Voxel	79.5	77.1
CenterPoint-Voxel	53.4	61.3
X-Ray CenterPoint-Voxel (ours)	**54.3**	**62.9**
X-Ray Teacher* Transfusion-L	81.3	78.2
Transfusion-L	56.1	66.3
X-Ray Transfusion-L (ours)	**56.8**	**66.9**

Table 6. Performance of X-Ray-powered Mean Teacher and Proficient Teacher methods in the semi-supervised setting using SECOND and CenterPoint baselines, ONCE validation set. Our approach consistently outperforms the state-of-the-art for Semi-Supervised 3D Object Detection in terms of mAP across all splits.

Method	SECOND	CenterPoint
Train (5k labeled samples)		
Pretraining	63.22	64.41
Small (5k labeled + 100k unlabeled samples)		
Mean Teacher	64.31	66.47
X-Ray MT (ours)	65.26 (+0.95)	67.60 (+1.13)
Proficient Teacher	67.06	67.72
X-Ray PT (ours)	**68.43** (+1.37)	**68.76** (+1.04)
Medium (5k labeled + 500k unlabeled samples)		
Mean Teacher	64.73	66.87
X-Ray MT (ours)	65.62 (+0.89)	67.78 (+0.91)
Proficient Teacher	67.49	68.54
X-Ray PT (ours)	**68.75** (+1.26)	**69.96** (+1.42)
Large (5k labeled + 1M unlabeled samples)		
Mean Teacher	65.03	67.45
X-Ray MT (ours)	65.97 (+0.94)	68.17 (+0.72)
Proficient Teacher	67.89	69.68
X-Ray PT (ours)	**69.10** (+1.21)	**70.55** (+0.87)

We evaluate various Point Cloud Registration (PCR) methods used in the Object Complete Frame Generation process. This analysis, detailed in Table 3, is conducted on a ONCE Small split. The results indicate that superior PCR leads to the creation of less noisy objects, which in turn contributes to improved overall quality. However, it's important to note that methods, such as the Greedy Grid method [2] and GeDi [27], are computationally more expensive. This introduces a trade-off between computational efficiency and the quality of the results, highlighting the need for a balanced approach in the selection of PCR methods.

4.4. Model comparison

4.4.1 Supervised Learning

We perform model comparisons using SECOND [42], CenterPoint-Voxel [47], and DSVT Pillar [36] on the Waymo dataset, and CBGS [50] and CenterPoint-Voxel on the NuScenes dataset. Additionally, we train scaled versions of these models without the X-Ray Teacher to show that the improvements in detection quality are due to the effectiveness of our method, not just an increase in the number of parameters. We scale Waymo students because of the extremely dense and complete point clouds, unlike NuScenes (see supplementary materials), where simpler data requires fewer parameters to learn meaningful feature representations. The results, presented in Tables 4 and 5, demonstrate that our approach consistently outperforms baseline models by 1-2 mAP.

Table 7. Student performance on ONCE validation small split with different tracking methods used in Objects Temporal Fusion

Tracking method	**SECOND**	**CenterPoint**
Greedy	68.43	68.76
Kalman Filter + IoU	68.57	68.91
ReID + Kalman Filter	**68.79**	**69.12**

4.4.2 Semi-Supervised Learning

To validate the effectiveness of our method in the context of pseudo-label-based semi-supervised learning, we perform a comparative analysis with the Mean Teacher [33] and Proficient Teacher [46] methods, which use the SECOND and CenterPointVoxel models. We compare the results obtained with and without the use of the X-Ray Teacher, as detailed in Table 6. Our results show that the application of our approach consistently improves performance, yielding an improvement of 0.8-1.4 mAP.

5. Conclusion

In our research, we proposed the innovative X-Ray Teacher framework, tailored to improve 3D Object Detection models in supervised and semi-supervised settings. Our extensive results have shown that this approach not only achieves state-of-the-art performance in the semi-supervised setting on the ONCE benchmark, but also consistently improves the quality of supervised models on NuScenes and Waymo Open Dataset. The main contributions of our work include the design of the X-Ray Teacher framework, the development of the Objects Temporal Fusion block for generating Object-Complete frames for data lacking ground truth tracking labels, and the demonstration of the potential of our method to improve the performance of any supervised model trained on sequential data. Future work will focus on optimizing the Objects Temporal Fusion block for more complex environments and exploring the integration of our framework with a broader range of model architectures and applications.

6. 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

[1] Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

[2] David Bojanić, Kristijan Bartol, Josep Forest, Stefan Gumhold, Tomislav Petković, and Tomislav Pribanić. Challenging the universal representation of deep models for 3d point cloud registration. In *BMVC 2022 Workshop Universal Representations for Computer Vision*, 2022.

[3] Daniel Brodeski, Igal Bilik, and Raja Giryes. Deep radar detector. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6, 2019.

[4] Collin Burns et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[5] 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.

[6] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M. Alvarez. Focalformer3d: Focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8394–8405, October 2023.

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

[8] Xu Dong, Pengluo Wang, Pengyue Zhang, and Langechuan Liu. Probabilistic oriented object detection in automotive radar. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 458–467, 2020.

[9] J. Fang, D. Zhou, F. Yan, T. Zhao, F. Zhang, Y. Ma, L. Wang, and R. Yang. Augmented lidar simulator for autonomous driving. In *IEEE Robotics and Automation*, 2020.

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

[11] Maksim Golyadkin, Alexander Gambashidze, Ildar Nurgaliev, and Ilya Makarov. Refining the once benchmark with hyperparameter tuning. *IEEE Access*, 12:3805–3814, 2024.

[12] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

[13] Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu. Teacher-student architecture for knowledge distillation: A survey. In *arXiv preprint arXiv:2210.17332*, 2022.

[14] Haotian Hu, Fanyi Wang, Jingwen Su, Yaonong Wang, Laifeng Hu, Weiye Fang, Jingwei Xu, and Zhiwang Zhang. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. *arXiv preprint arXiv:2303.17895*, 2023.

[15] 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.

[16] 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.

[17] Junho Koh, Junhyung Lee, Youngwoo Lee, Jaekyum Kim, and Jun Won Choi. Mgtanet: Encoding sequential lidar points using long short-term motion-guided temporal attention for 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1179–1187, Jun. 2023.

[18] Haoang Li, Jinhu Dong, Binghui Wen, Ming Gao, Tianyu Huang, Yun-Hui Liu, and Daniel Cremers. Ddit: Semantic scene completion via deformable deep implicit templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21894–21904, October 2023.

[19] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Ziyu Li, Yuncong Yao, Zhibin Quan, Jin Xie, and Wankou Yang. Spatial information enhancement network for 3d object detection from point cloud. *Pattern Recognition*, page 108684, 2022.

[21] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781, 2023.

[22] Bence Major, Daniel Fontijne, Amin Ansari, Ravi Teja Sukhavasi, Radhika Gowaikar, Michael Hamilton, Sean Lee, Slawomir Grzechnik, and Sundar Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 924–932, 2019.

[23] Ilya Makarov, Vladimir Aliev, and Olga Gerasimova. Semi-dense depth interpolation using deep convolutional neural networks. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1407–1415, New York, NY, USA, 2017. Association for Computing Machinery.

[24] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021.

[25] Michael Meyer, Georg Kuschk, and Sven Tomforde. Graph convolutional networks for 3d object detection on radar data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3053–3062, 2021.

[26] Mahyar Najibi, Guangda Lai, Abhijit Kundu, Zhichao Lu, Vivek Rathod, Thomas Funkhouser, Caroline Pantofaru, David Ross, Larry S. Davis, and Alireza Fathi. Dops: Learning to detect 3d objects and predict their 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3d local deep descriptors for point cloud regis-

tration. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (early access) 2022.

[28] 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.

[29] 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.

[30] Ilia Semenkov, Aleksei Karpov, Andrey V Savchenko, and Ilya Makarov. Inpainting semantic and depth features to improve visual place recognition in the wild. *IEEE Access*, 2024.

[31] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020.

[32] 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.

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[34] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

[35] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*, 2020.

[36] 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.

[37] Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, and Liwei Wang. Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In *ICCV*, 2023.

[38] Haiyang Wang, Wenguan Wang, Xizhou Zhu, Jifeng Dai, and Liwei Wang. Collaborative visual navigation. *arXiv preprint arXiv:2107.01151*, 2021.

[39] Yan Wang, Bin Yang, Rui Hu, Ming Liang, and Raquel Urtasun. Plumenet: Efficient 3d object detection from stereo images. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3383–3390, 2021.

[40] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[41] 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, pages 2893–2901, 2022.

[42] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[43] 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.

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

[45] Lehan Yang and Kele Xu. Cross modality knowledge distillation for multi-modal aerial view object classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 382–387, June 2021.

[46] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *European Conference on Computer Vision*, pages 727–743. Springer, 2022.

[47] 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.

[48] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020.

[49] 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.

[50] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. In *arXiv preprint arXiv:1908.09492*, 2019.

[51] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3357–3364, 2017.