Shelf-Supervised Cross-Modal Pre-Training for 3D Object Detection

Mehar Khurana[1,*], **Neehar Peri**[2*], **James Hays**[3], **Deva Ramanan**[2]
[1]IIIT Delhi, [2]Carnegie Mellon University, [3]Georgia Institute of Technology

Abstract: State-of-the-art 3D object detectors are often trained on massive labeled datasets. However, annotating 3D bounding boxes remains prohibitively expensive and time-consuming, particularly for LiDAR. Instead, recent works demonstrate that self-supervised pre-training with unlabeled data can improve detection accuracy with limited labels. Contemporary methods adapt best-practices for self-supervised learning from the image domain to point clouds (such as contrastive learning). However, publicly available 3D datasets are considerably smaller and less diverse than those used for image-based self-supervised learning, limiting their effectiveness. We do note, however, that such 3D data is naturally collected in a multimodal fashion, often paired with images. Rather than pre-training with only self-supervised objectives, we argue that it is better to bootstrap point cloud representations using image-based foundation models trained on internet-scale data. Specifically, we propose a *shelf*-supervised approach (e.g. supervised with off-the-shelf image foundation models) for generating zero-shot 3D bounding boxes from paired RGB and LiDAR data. Pre-training 3D detectors with such pseudo-labels yields significantly better semi-supervised detection accuracy than prior self-supervised pretext tasks. Importantly, we show that image-based shelf-supervision is helpful for training LiDAR-only, RGB-only and multi-modal (RGB + LiDAR) detectors. We demonstrate the effectiveness of our approach on nuScenes and WOD, significantly improving over prior work in limited data settings. Our code is available on GitHub.

Keywords: Shelf-Supervised 3D Object Detection, Semi-Supervised Learning, Vision-Language Models, Autonomous Vehicles

1 Introduction

3D object detection is an integral component of the robot perception stack. To facilitate research in this space, the Autonomous Vehicle (AV) industry has released several large-scale 3D annotated multi-modal datasets [1, 2, 3]. However, 3D detectors pre-trained on nuScenes [1] cannot be easily applied to Argoverse [3] due to differences in sensor characteristics across hardware platforms. In practice, training 3D detectors for a new hardware platform requires re-annotating 3D bounding boxes at scale, which can be prohibitively expensive [4] and time consuming [5]. Instead, recent works [6, 7, 8, 9] demonstrate that self-supervised pre-training with unlabeled sensor data can improve downstream detection accuracy with limited labels. Notably, AVs *already* capture large-scale unlabeled multi-modal data localized to HD maps during normal testing [10], motivating the exploration of self-supervised pre-training for outdoor scenes.

Status Quo. Self-supervised learning offers a scalable framework to learn from massive unlabeled datasets. Typically, self-supervised methods establish a pretext task that derives supervision directly from the data (e.g. occupancy prediction, contrastive learning, or masked auto-encoding). Once trained, these self-supervised representations can be adapted for downstream tasks by fine-tuning on

[*]Equal Contribution.

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

Figure 1: **Cross-Modal 3D Detection Distillation with Vision-Language Models.** Existing datasets used for 3D representation learning are considerably smaller and less diverse than those used for image-based self-supervised learning, limiting the effectiveness of pre-training. Therefore, we propose a simple approach for transferring object-centric priors from vision-language models to LiDAR. Specifically, we project 3D LiDAR points onto 2D instance segmentation masks to generate zero-shot 3D bounding boxes [10] that can be used for pre-training 3D detectors.

a limited set of annotations. Given the success of image-based self-supervised learning [11, 12, 13], recent approaches revisit contrastive learning in the context of 3D sensor data [9, 14]. However, existing 3D datasets are considerably smaller and less diverse than those used in image-based self-supervised learning. Instead of solely relying on self-supervised objectives for pre-training, we advocate for a more practical approach that embraces recent advances in image-based foundational models by bootstrapping point cloud representations with vision-language models (VLMs) trained on internet-scale data. Since off-the-shelf vision-language models *already* encode semantics and object priors, our work focuses on distilling these 2D foundational priors into 3D detectors.

Bootstrapping 3D Representations with 2D Priors. VLMs [15, 16, 17] are trained on internet-scale data and show impressive zero-shot detection performance across many application domains [18]. Therefore, we posit that pre-training with noisy pseudo-labels distilled from vision-language models can outperform contrastive features learned from scratch (cf. Fig. 1). However, distilling image-based foundational knowledge to LiDAR remains an open challenge [19, 20, 21]. We address this challenge by explicitly "inflating" 2D instance masks to 3D bounding boxes via unlabeled multimodal (RGB + LiDAR) training data, using a careful combination of LiDAR cues, HD maps, and shape priors [10]. This allows us to produce accurate 3D pseudo-labels which can then be used to pre-train uni-modal or multi-modal 3D detectors. One significant advantage of our *shelf*-supervision over self-supervision is that it naturally produces a more aligned pretext task; instead of pre-training 3D detectors with contrastive learning, we learn from 3D bounding box pseudo-labels.

Contributions. We present three major contributions. First, we propose Cross-Modal 3D Detection Distillation (CM3D), a zero-shot approach for generating 3D bounding boxes using off-the-shelf vision-language models. Notably, we show that pre-training detectors with 3D bounding box pseudo-labels achieves higher downstream detection accuracy than prior contrastive learning objectives. We conduct extensive experiments to ablate our design choices and demonstrate that our simple approach achieves state-of-the-art unsupervised and semi-supervised 3D detection accuracy on the nuScenes and WOD benchmarks.

2 Related Works

Unsupervised 3D Object Detection has gained significant interest in recent years as a way to auto-label large datasets without human annotators. Dewan et al. [22] introduced a model-free approach to detect and track the visible portions of objects by leveraging motion cues from LiDAR. More recently, Wong et al. [23] identified unknown instances through supervised segmentation and clustering [24]. However, both [22, 23] only reason about the visible extent of objects and are unable to generate amodal bounding boxes. Cen et al. [25] used a supervised detector to generate bounding

box proposals for unknown categories, and Tian et al. [26] exploited the correspondence between images and point clouds to generate object proposals. Similarly, Wilson et al. [10] employed an HD map and shape priors to "inflate" 2D detections to 3D cuboids. Utilizing priors such as surface normals [27, 28, 29] and motion cues [30, 31, 32, 33] has been shown to improve unsupervised object detection. You et al. [34] integrated cues from temporal scene changes (e.g. point ephemerality from multiple traversals) to detect mobile objects [35, 36]. Najibi et al. [37] used scene flow and clustering to generate 3D bounding boxes for moving objects. However, this method does not assign semantic labels to cuboids and cannot detect static objects. Similarly, [38, 39] use scene flow to cluster points and generate pseudo-labels.

Self-Supervised Learning for 3D Representations has been applied to many domains, including object-centric point clouds [40, 41, 42], indoor scenes [43, 44, 45, 46, 47, 48], and outdoor environments [9, 7, 49]. Current techniques for self-supervised point cloud pre-training can be broadly characterized by their receptive field. Scene-based contrastive learning [6, 50] (often used in the context of representation learning for indoor scenes) may not capture the necessary fine-grained details required for recognizing small objects. In contrast, voxel-based contrastive learning (often used in the context of representation learning for outdoor scenes) inherently struggles with a restricted receptive field, limiting its ability to encode larger structures. Region-based representations [7, 51] attempt to strike a balance between coarse-grained scene representations and fine-grained voxel representations. However, most regions in outdoor scenes (e.g. ground plane and buildings) lack informative cues, making it difficult to learn robust features. More recently, multi-modal contrastive learning methods [8, 9] demonstrated that paired multi-modal sensors can provide complementary (self-)supervision for pre-training. PointContrast [6] established correspondences between different views of point clouds using a contrastive loss. DepthContrast [14] treated different depth maps as contrastive instances and discriminates between them. STRL [52] extracted invariant representations from temporally correlated frames in a 3D point cloud sequence. More recently, SimIPU [8] and CALICO [9] delve into multi-modal contrastive learning by utilizing paired RGB and LiDAR data. GeoMAE [49] frames point-cloud representation learning as a masked auto-encoding task, and uses geometric pretext tasks including occupancy, normals, and curvature estimation.

Leveraging 2D Foundational Models for 3D Perception is an active area of research. Sautier et al. [53] introduced SLidR, a 2D-to-3D representation distillation method aimed at cross-modal self-supervised learning on large-scale point clouds. Mahmoud et al. [54] subsequently extended SLidR with a semantically-aware contrastive constraint and a class-balancing loss. SEAL [55] takes inspiration from SLidR and used SAM [56] to generate 3D class-agnostic point cloud segments for contrastive learning. More recently, SA3D [57] extends SAM [56] to segment 3D objects with NeRFs. Anything-3D [58] combines BLIP [16], a pre-trained 2D text-to-image diffusion model, with SAM for single-view conditioned 3D reconstruction. 3D-Box-Segment-Anything utilizes SAM with a pre-trained VoxelNeXt [59] 3D detector for interactive labeling. SAM3D [60] repurposes SAM to directly segment objects from BEV point clouds. Most recently, UP-VL [20] distilled 2D vision-language features [11] into LiDAR point clouds to generate amodal cuboids.

3 Cross-Modal 3D Detection Distillation (CM3D) with VLMs

Given a large unlabelled set of multi-modal (RGB + LiDAR) data, we combine 2D VLMs with 3D domain knowledge (given by 3D shape priors and map constraints) to generate 3D bounding box pseudo-labels for pre-training 3D detectors.

3.1 Generating 3D Bounding Box Pseudo-Labels

2D Mask Generation. The first stage of our pipeline requires producing accurate 2D instance masks for a fixed vocabulary of object categories. We prompt a 2D VLM detector (e.g., Detic [15] or GroundingDINO [61]) with class names (e.g., `car, bus, truck`) to generate 2D box proposals. Although some VLM detectors (i.e., Detic) already produce instance segmentation masks, we found it more effective to prompt separate (foundation) models such as SAM [56] with the predicted 2D

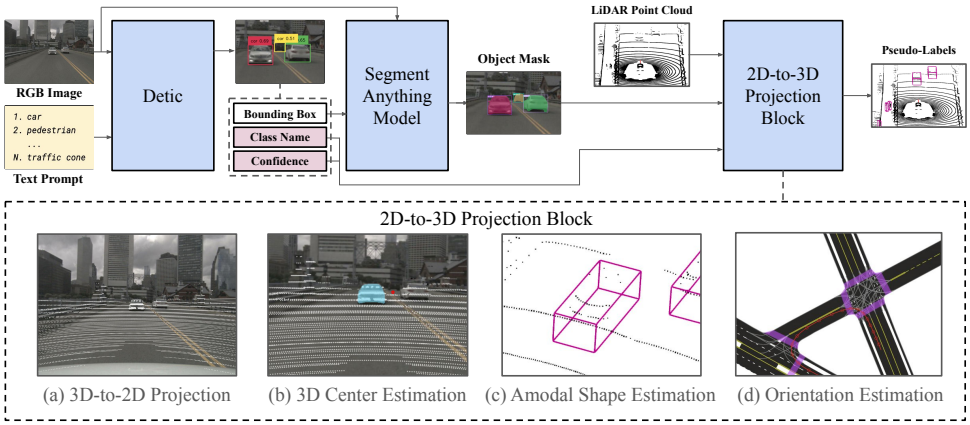

Figure 2: **CM3D Pseudo-Label Pipeline.** First, we prompt an open-vocabulary 2D detector (e.g., Detic [15]) with a class name (e.g., `car`) to generate 2D box proposals. Next, we prompt SAM [56] with the predicted 2D bounding boxes to generate high-quality instance segmentation masks. We then generate an oriented 3D cuboid using the set of LiDAR points that project to a given 2D instance mask. Specifically, we define the *center* of the cuboid to be the medoid of the LiDAR points, the *dimensions* (length, width, height) to be a fixed shape prior (similar to an anchor box) as reported by ChatGPT when prompted with the class name, and the *orientation* to be aligned with lane geometry provided from an HD map.

bounding boxes to generate high-quality segmentation masks. We evaluate the impact of using different VLMs in Appendix F. Notably, we find that Detic [15] produces better 3D pseudo-labels than GroundingDINO [61] in our CM3D pipeline.

3D Projection. Next, we project the LiDAR points onto the 2D image plane and group all points within each instance mask. To produce a final 3D bounding box pseudo-label, we need to generate a 3D centroid, 3D orientation, and cuboid dimensions (width, length, height). Our key technical approach, described below and inspired by [10], is to leverage 3D priors in the form of maps (e.g., car cuboids should typically be aligned with map geometry) and shape priors (e.g., fine-grained categories such as trucks tend to have canonical dimensions).

3D Cuboid Generation. We define an initial candidate 3D centroid to be the medoid of points within each mask, expressed as $m = \text{argmin}_{x \in P_M} \sum_{p \in P_M} ||p - x||_2$, where P_M is the set of LiDAR points that lie within the mask M. Note that this medoid tends to lie on the surface of the object visible to the LiDAR sensor rather than the true object centroid, which implies it will require refinement. To estimate a canonical width-length-height, we simply prompt an LLM (e.g., ChatGPT) with each class name and use the returned value for all instances. Although LLMs do not have access to our 3D training data, they have seen many descriptions of object shapes on the web and can provide reasonable prototypical 3D sizes of common objects (Appendix H). Lastly, we estimate box orientation for vehicles using lane geometry from an HD map and assign the direction of the nearest lane to each cuboid. We assign an arbitrary orientation to all non-vehicles (e.g., `pedestrians`, `barriers` and `traffic cones`).

3D Refinement. Fig. 2 visualizes our overall pseudo-label generation pipeline. Because many components of our pipeline provide only rough 3D estimates (e.g., not all trucks have precisely the same 3D dimension), we find that 3D labels can be refined. Table 1 ablates several strategies for improvement, including prompt engineering to improve VLM zero-shot 2D detections, LiDAR accumulation to improve medoid estimation, mask erosion to remove noisy LiDAR points near mask boundaries, medoid compensation to reduce object center estimation biases, and late-fusion of independent zero-shot 3D detectors to improve orientation and shape estimation. We find that prompt engineering and medoid compensation provide the greatest improvement to pseudo-label quality. We present a detailed overview of each refinement step in Appendix B.

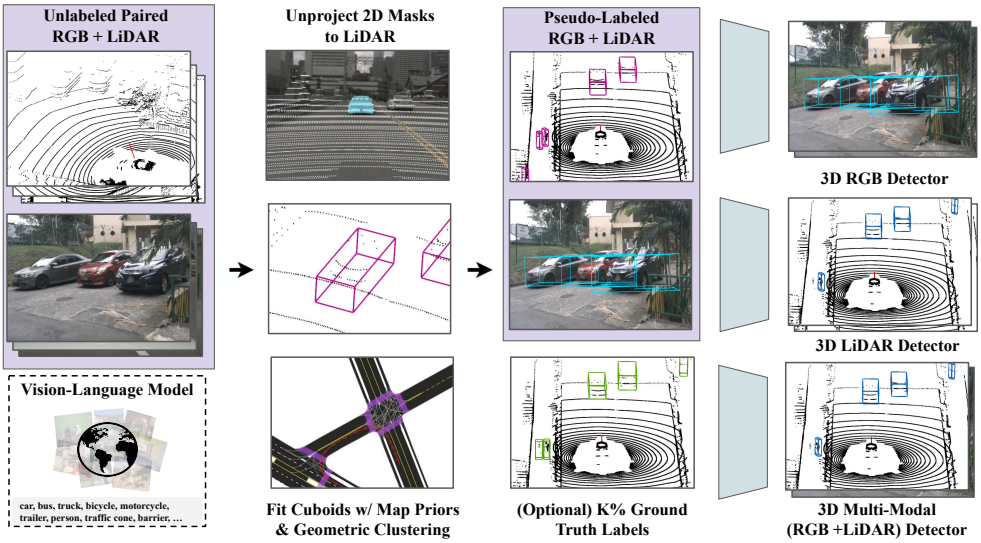

Figure 3: **Pre-Training and Fine-Tuning Pipeline.** Given unlabeled, calibrated, and paired RGB images and LiDAR sweeps, we generate 3D bounding boxes psuedo-labels using foundational 2D priors from VLMs, map priors, and geometric clustering. Using these pseudo-labels, we can train LiDAR-only, RGB-only, or multi-modal 3D detectors.

3.2 Training with Pseudo-Labels

We use our generated pseudo-labels to pre-train 3D detectors that will be fine-tuned on a limited set of labeled data. We present a block diagram of our training pipeline in Fig 3. Notably, it is trivial to train *any* 3D detector with our pseudo-labels, making our proposed framework widely applicable. Further, although our pseudo-label generator requires paired RGB and LiDAR data for pseudo-label generation, we can easily train a LiDAR-only model like CenterPoint, which requires only LiDAR sweeps at inference. In this paper, we evaluate CM3D by training two popular detectors, CenterPoint [62] and BEVFusion [63]. Following CALICO's [9] experiment setup, we train BEVFusion with the CenterHead and turn off class-balanced grouping and sampling. Although copy-paste augmentation is widely used when training 3D detectors, we find that performance degrades with our noisy pseudo-labels (Appendix E). We train the CenterHead following standard 3D detection losses. We use the sigmoid focal loss (for recognition) [64] and L1 regression loss (for localization). Concretely, our loss function is as follows: $L = L_{HM} + \lambda L_{REG}$, where $L_{HM} = \sum_{i=0}^{C} SigmoidFocalLoss(X_i, Y_i)$ and $L_{REG} = |X_{BOX} - Y_{BOX}|$, where X_i and Y_i are the ith class' predicted and ground-truth heat maps, while X_{BOX} and Y_{BOX} are the predicted and pseudo-label box attributes.

Self Training. Although directly pre-training with our pseudo-labels significantly improves mAP, it also yields lower NDS compared to prior work. We posit that this is due to the errors in size and orientation estimation, which can both be attributed to our simplistic 3D priors. Prior work suggests that self-training can be effective for semi-supervised learning [34, 20]; not only can it discover more objects [34, 20], but it may also refine noisy box labels. We ablate the impact of self-training in the Appendix D. We find that even one round of self-training significantly improves average precision and reduces orientation estimation error.

4 Experiments

In this section, we conduct extensive experiments to evaluate our proposed approach. We find that CM3D achieves state-of-the-art unsupervised detection performance and significantly improves over prior works in limited data settings (cf. Table 3). We copy the results of prior work from CALICO and refer readers to [9] for further details.

Table 1: **Pseudo-Label Refinement**. We analyze the impact of each component over the baseline [10] (cf. Fig 2). Importantly, we find that prompt engineering, medoid compensation, and non-maximum suppression have the greatest impact. See Appendix B for additional details.

Method	mAP ↑	NDS ↑
Baseline (Re-implementation of [10])	18.6	17.8
+ Prompt Engineering	19.7 (+1.1)	18.4 (+0.6)
+ LiDAR Accumulation	20.0 (+0.3)	18.6 (+0.2)
+ Mask Erosion	20.2 (+0.2)	19.1 (+0.5)
+ Medoid Compensation	21.8 (+1.6)	21.3 (+2.1)
+ Non-Maximal Suppression	22.8 (+1.0)	21.9 (+0.6)
+ Late Fusion	23.0 (+0.2)	22.1 (+0.2)

Datasets. We evaluate our work using the well-established nuScenes dataset [1]. We follow the suggested protocol in [7, 49, 9] and sample K% of the training data uniformly from the entire training set. Lastly, we also evaluate our approach on the Waymo Open Dataset [2] in Appendix I, and find that our conclusions hold across datasets.

Metrics. Mean average precision (mAP) is an established metric for object detection [65]. For 3D detection, a true positive (TP) is defined as a detection that has a center distance within a distance threshold on the ground-plane to a ground-truth annotation [1]. mAP computes the mean of AP over all classes, where per-class AP is the area under the precision-recall curve drawn with distance thresholds of [0.5, 1, 2, 4] meters. NDS (nuScenes Detection Score) is computed by taking a weighted sum of mAP along with five other metrics, including mATE (mean absolute translation error), mAOE (mean absolute orientation error), mASE (mean absolute scale error), mAVE (mean absolute velocity error), and mAAE (mean absolute attribute error). mAP is assigned a weight of 5, while all other metrics are assigned a weight of 1.

Pseudo-Label Refinement. We ablate our pseudo-label generation algorithm to determine how each component improves the baseline. As discussed above, we find that medoid compensation has the greatest impact on pseudo-label quality, improving mAP by 1.6% and NDS by 2.1%. Tuning the Detic text prompts by including synonyms also improves results significantly. Finally, the 3D distance-based NMS helps remove duplicates present in the overlapping regions of the ring cameras and increases the mAP by 1%.

Class Agnostic Evaluation. Prior pseudo-label generation methods only evaluate class-agnostic performance [39]. For fair comparison with prior work, we evaluate CM3D pseudo-labels without differentiating between classes in Table 2. Prior works use motion cues from scene flow to generate object proposals for moving objects [38]. In contrast, our approach localizes both static and moving objects and classifies them using VLMs. CM3D outperforms LISO-TF [39] by 14.5% AP and 0.6% NDS, highlighting the benefit of using foundational priors and multi-modal inputs.

Table 2: **Class Agnostic Evaluation on nuScenes**. We evaluate class-agnostic performance of CM3D pseudo-labels for fair comparison with prior work. We find that our method significantly outperforms LISO-TF by 14.5% AP.

Method	Modality	AP ↑	NDS ↑	mATE ↓	mAOE ↓	mASE ↓
DBSCAN [24]	L	0.8	10.9	0.980	3.120	0.962
RSF [66]	L	1.9	18.3	0.774	1.003	0.507
Oyster-CP [67]	L	9.1	21.5	0.784	1.514	0.521
Oyster-TF [67]	L	9.3	23.3	0.708	1.564	0.448
LISO-CP [39]	L	10.9	22.4	0.750	1.062	0.409
LISO-TF [39]	L	13.4	27.0	0.628	0.938	0.408
CM3D (Ours)	L + C	27.9	27.6	0.592	0.872	0.428

Table 3: **Semi-Supervised 3D Detection on nuScenes**. In terms of zero-shot accuracy, pseudo-labels from CM3D outperforms prior art (SAM3D) by 20.7% NDS. With 5% supervised training data, pre-training CenterPoint with CM3D pseudo-labels improves over the prior art of PRC [9] by 8.1 mAP / 2.8 NDS. When comparing to Camera-LiDAR methods, pre-training BEVFusion with CM3D pseudo-labels outperforms prior art (CALICO) by 8.6 mAP / 4.6 NDS. We highlight the best LiDAR-only (L) results in blue, and the best Camera-LiDAR (L+C) results in red.

Training Data	Method	Modality		mAP ↑	NDS ↑
		Train	Test		
0% (Unsupervised)	SAM3D [60]	L	L	1.7	2.4
	CM3D (Ours)	L + C	L + C	23.0	22.1
	CenterPoint [62] + CM3D (Ours)	L + C	L	16.7	21.4
	BEVFusion-C [63] + CM3D (Ours)	L + C	C	11.7	16.1
	BEVFusion [63] + CM3D (Ours)	L + C	L + C	20.6	23.3
5%	CenterPoint [62] (Rand. Init.)	L	L	33.1	37.4
	PointContrast [6]	L	L	36.7	43.0
	ProposalContrast [7]	L	L	37.0	43.1
	PRC [9]	L	L	38.2	46.0
	CenterPoint [62] + CM3D (Ours)	L + C	L	46.3	48.8
	BEVFusion [63] (Rand. Init.)	L + C	L + C	39.0	43.7
	SimIPU [8]	L + C	L + C	39.1	45.8
	PRC [9] + BEVDistill [68]	L + C	L + C	41.0	47.5
	CALICO [9]	L + C	L + C	41.7	47.9
	BEVFusion [63] + CM3D (Ours)	L + C	L + C	51.3	52.5
10%	CenterPoint [62] (Rand. Init.)	L	L	41.1	48.0
	PointContrast [6]	L	L	42.3	51.2
	ProposalContrast [7]	L	L	42.1	51.1
	PRC [9]	L	L	44.1	53.1
	CenterPoint [62] + CM3D (Ours)	L + C	L	51.0	56.3
	BEVFusion [63] (Rand. Init.)	L + C	L + C	46.2	51.9
	SimIPU [8]	L + C	L + C	47.5	52.4
	PRC [9] + BEVDistill [68]	L + C	L + C	49.7	53.6
	CALICO [9]	L + C	L + C	50.0	53.9
	BEVFusion [63] + CM3D (Ours)	L + C	L + C	53.3	56.5

Evaluating Zero-Shot Performance. We evaluate the quality of CM3D's pseudo-labels (0% Training Data) and find that our method achieves 22.3% mAP and 22.1% NDS. We compare CM3D with SAM3D [60], a recently released zero-shot 3D detector. Notably, both CM3D and SAM3D use SAM [56] to group LIDAR points. However, CM3D groups LiDAR points on the image plane and SAM3D [60] groups LiDAR points in the 2D BEV plane. Our multi-modal zero-shot 3D detector significantly outperforms SAM3D's LiDAR-only approach by 20.7%. Importantly, SAM3D was initially designed as a class-agnostic zero-shot detector. We posit that SAM3D's poor performance can be attributed to nuScenes' diverse classes.

Distilling Multi-Modal Priors into Uni-Modal Models. Although CM3D requires paired RGB images and LiDAR sweeps for pseudo-label generation, we can easily train an RGB-only or LiDAR-only student model (e.g., CenterPoint + CM3D). Importantly, these student models don't require paired RGB-LiDAR data at inference (and are therefore compared against other models that only use LiDAR at inference). However, we find that the CenterPoint + CM3D (0% Training Data) student model performs worse than the pseudo-label generator (CM3D). This can be attributed to learning with noisy labels and not having access to the same sensor data as the teacher. Interestingly, we find that using multi-modal cues for both training and testing yields higher results.

Significant Improvements in Low-Data Setting. Pre-training CenterPoint with CM3D pseudo-labels and fine-tuning on labeled data improves over the prior art of PRC [9] by 8.1 mAP and 2.8 NDS with 5% labeled data. Similarly, pre-training BEVFusion with CM3D pseudo-labels outperforms prior art (CALICO) by 8.6 mAP and 4.6 NDS. Notably all pre-training strategies improve over random initialization (Rand. Init.). This suggests that aligning your pre-training and fine-

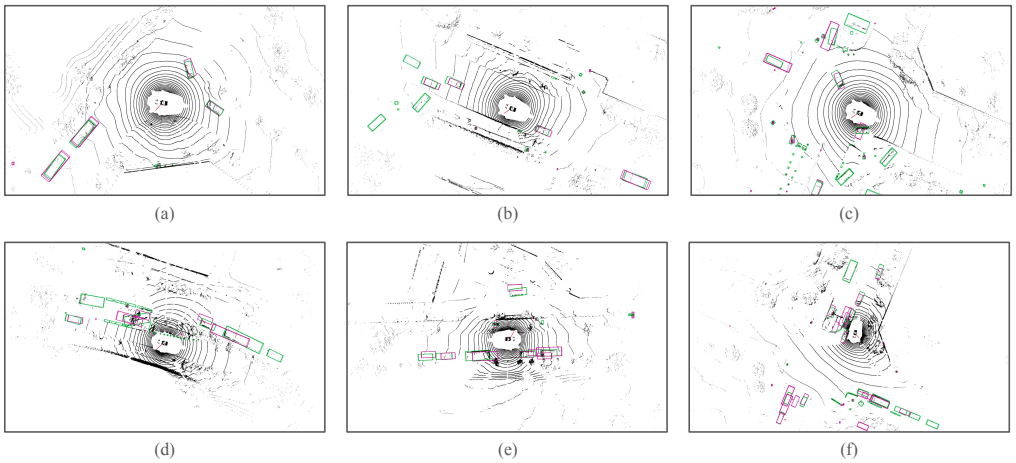

Figure 4: **Qualitative Results of Pseudo-Labels**. We visualize pseudo-labels (pink) and ground-truth labels (green) across all 10 object classes on the nuScenes val-set. In (**a**), our pseudo-labels accurately estimate location, cuboid size, and orientation, demonstrating the general effectiveness of medoid compensation and map-based orientation estimation. In (**b**), we find that CM3D often misses heavily-occluded objects. This is unsurprising because our method relies on accurate RGB-based detections, which often fail with heavy occlusions. In (**c**), our map-based orientation estimation fails when the predicted object is not oriented in the direction of any lane. For example, the incorrect orientation of the `car` turning into the intersection (not aligned to any nearby lanes) illustrates the limitations of our approach. In both (**d**) and (**f**), we are unable to label several barriers. We attribute these missed detections to the ambiguity of the class name `barrier`. Notably, a `barrier` in nuScenes may not be the same as `barrier` as defined in internet pre-training data [18]. In (**d**), (**e**), and (**f**), we produce duplicate boxes for the same instances, indicating a failure of NMS.

tuning tasks can yield better performance. See Appendix C for additional results with 20% and 50% ground truth labels. Notably, we find that the impact of our proposed pre-training approach diminishes when training with additional ground truth data.

Qualitative Results. We visualize the output of our pseudo-label generator in Fig. 4. Although our method generates reasonable predictions in many cases, we find that our method fails in cases of occlusions (where there is no 3D information) and in cases where the VLM predicts a false positive detection with high confidence. See Fig 4 and Appendix A for detailed analysis.

5 Conclusion

In this paper, we propose Cross-Modal 3D Detection Distillation (CM3D), a zero-shot approach for generating 3D bounding boxes using vision-language models. We demonstrate that pre-training detectors with zero-shot 3D bounding box pseudo-labels achieves higher downstream semi-supervised detection than prior contrastive learning objectives. We conduct comprehensive experiments to ablate our design choices and demonstrate that our simple method significantly improves over prior work in limited data settings. However, we find that our proposed pre-training strategy limits generalization to arbitrary downstream tasks. Moreover, our current pseudo-labeling pipeline requires LiDAR for precise box localization and HD maps for orientation estimation, which may not be readily available for broader robotics domains (Appendix J).

6 Acknowledgements

We would like to thank Jenny Seidenschwarz for her feedback on early drafts. This work was supported in part by the NSF GRFP (Grant No. DGE2140739).

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[2] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. 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.

[3] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[4] K. Vedder, N. Peri, N. Chodosh, I. Khatri, E. Eaton, D. Jayaraman, Y. Liu, D. Ramanan, and J. Hays. Zeroflow: Fast zero label scene flow via distillation. *arXiv preprint arXiv:2305.10424*, 2023.

[5] W. Chen, A. Edgley, R. Hota, J. Liu, E. Schwartz, A. Yizar, N. Peri, and J. Purtilo. Rebound: An open-source 3d bounding box annotation tool for active learning. *arXiv preprint arXiv:2303.06250*, 2023.

[6] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020.

[7] J. Yin, D. Zhou, L. Zhang, J. Fang, C.-Z. Xu, J. Shen, and W. Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022.

[8] Z. Li, Z. Chen, A. Li, L. Fang, Q. Jiang, X. Liu, J. Jiang, B. Zhou, and H. Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1500–1508, 2022.

[9] J. Sun, H. Zheng, Q. Zhang, A. Prakash, Z. M. Mao, and C. Xiao. Calico: Self-supervised camera-lidar contrastive pre-training for bev perception. *arXiv preprint arXiv:2306.00349*, 2023.

[10] B. Wilson, Z. Kira, and J. Hays. 3d for free: Crossmodal transfer learning using hd maps. *arXiv preprint arXiv:2008.10592*, 2020.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[14] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021.

[15] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

[16] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[17] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[18] A. Madan, N. Peri, S. Kong, and D. Ramanan. Revisiting few-shot object detection with vision-language models. *arXiv preprint arXiv:2312.14494*, 2023.

[19] A. Ošep, T. Meinhardt, F. Ferroni, N. Peri, D. Ramanan, and L. Leal-Taixé. Better call sal: Towards learning to segment anything in lidar. *arXiv preprint arXiv:2403.13129*, 2024.

[20] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, and D. Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8602–8612, 2023.

[21] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] A. Dewan, T. Caselitz, G. D. Tipaldi, and W. Burgard. Motion-based detection and tracking in 3d lidar scans. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 4508–4513. IEEE, 2016.

[23] K. Wong, S. Wang, M. Ren, M. Liang, and R. Urtasun. Identifying unknown instances for autonomous driving. In *Conference on Robot Learning*, pages 384–393. PMLR, 2020.

[24] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017.

[25] J. Cen, P. Yun, J. Cai, M. Y. Wang, and M. Liu. Open-set 3d object detection. In *2021 International Conference on 3D Vision (3DV)*, pages 869–878. IEEE, 2021.

[26] H. Tian, Y. Chen, J. Dai, Z. Zhang, and X. Zhu. Unsupervised object detection with lidar clues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5972, 2021.

[27] A. Collet, S. S. Srinivasay, and M. Hebert. Structure discovery in multi-modal data: a region-based approach. In *2011 IEEE International Conference on Robotics and Automation*, pages 5695–5702. IEEE, 2011.

[28] G. M. García, E. Potapova, T. Werner, M. Zillich, M. Vincze, and S. Frintrop. Saliency-based object discovery on rgb-d data with a late-fusion approach. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1866–1873. IEEE, 2015.

[29] J. Shin, R. Triebel, and R. Siegwart. Unsupervised discovery of repetitive objects. In *2010 IEEE International Conference on Robotics and Automation*, pages 5041–5046. IEEE, 2010.

[30] L. Ma and G. Sibley. Unsupervised dense object discovery, detection, tracking and reconstruction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 80–95. Springer, 2014.

[31] L. Ma, M. Ghafarianzadeh, D. Coleman, N. Correll, and G. Sibley. Simultaneous localization, mapping, and manipulation for unsupervised object discovery. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1344–1351. IEEE, 2015.

[32] D. Kochanov, A. Ošep, J. Stückler, and B. Leibe. Scene flow propagation for semantic mapping and object discovery in dynamic street scenes. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1785–1792. IEEE, 2016.

[33] S. Choudhary, A. J. Trevor, H. I. Christensen, and F. Dellaert. Slam with object discovery, modeling and mapping. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1018–1025. IEEE, 2014.

[34] Y. You, K. Luo, C. P. Phoo, W.-L. Chao, W. Sun, B. Hariharan, M. Campbell, and K. Q. Weinberger. Learning to detect mobile objects from lidar scans without labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1130–1140, 2022.

[35] E. Herbst, X. Ren, and D. Fox. Rgb-d object discovery via multi-scene analysis. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4850–4856. IEEE, 2011.

[36] E. Herbst, P. Henry, X. Ren, and D. Fox. Toward object discovery and modeling via 3-d scene comparison. In *2011 IEEE international conference on robotics and automation*, pages 2623–2629. IEEE, 2011.

[37] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, and D. Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *European Conference on Computer Vision*, pages 424–443. Springer, 2022.

[38] J. Seidenschwarz, A. Ošep, F. Ferroni, S. Lucey, and L. Leal-Taixé. Semoli: What moves together belongs together. *arXiv preprint arXiv:2402.19463*, 2024.

[39] S. Baur, F. Moosmann, and A. Geiger. Liso: Lidar-only self-supervised 3d object detection. *arXiv preprint arXiv:2403.07071*, 2024.

[40] A. Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 626–642. Springer, 2020.

[41] J. Sauder and B. Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019.

[42] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, and V. G. Kim. Self-supervised learning of point clouds via orientation estimation. In *2020 International Conference on 3D Vision (3DV)*, pages 1018–1028. IEEE, 2020.

[43] J. Hou, B. Graham, M. Nießner, and S. Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021.

[44] Z. Zhang, Y. Dong, Y. Liu, and L. Yi. Complete-to-partial 4d distillation for self-supervised point cloud sequence representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17661–17670, 2023.

[45] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, and W. Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[46] J. Hou, S. Xie, B. Graham, A. Dai, and M. Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021.

[47] S. Lal, M. Prabhudesai, I. Mediratta, A. W. Harley, and K. Fragkiadaki. Coconets: Continuous contrastive 3d scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2021.

[48] L. Li and M. Heizmann. A closer look at invariances in self-supervised pre-training for 3d vision. In *European Conference on Computer Vision*, pages 656–673. Springer, 2022.

[49] X. Tian, H. Ran, Y. Wang, and H. Zhao. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13570–13580, 2023.

[50] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[51] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[52] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021.

[53] C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022.

[54] A. Mahmoud, J. S. Hu, T. Kuai, A. Harakeh, L. Paull, and S. L. Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7102–7110, 2023.

[55] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu. Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347*, 2023.

[56] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[57] J. Cen, Z. Zhou, J. Fang, W. Shen, L. Xie, D. Jiang, X. Zhang, Q. Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36, 2024.

[58] Q. Shen, X. Yang, and X. Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023.

[59] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023.

[60] D. Zhang, D. Liang, H. Yang, Z. Zou, X. Ye, Z. Liu, and X. Bai. Sam3d: Zero-shot 3d object detection via segment anything model. *arXiv preprint arXiv:2306.02245*, 2023.

[61] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[62] T. Yin, X. Zhou, and P. 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.

[63] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[64] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[65] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[66] D. Deng and A. Zakhor. Rsf: Optimizing rigid scene flow from 3d point clouds without labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1277–1286, January 2023.

[67] L. Zhang, A. J. Yang, Y. Xiong, S. Casas, B. Yang, M. Ren, and R. Urtasun. Towards unsupervised object detection from lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9317–9328, 2023.

[68] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022.

[69] T. Yin, X. Zhou, and P. Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

[70] T. Wang, X. Zhu, J. Pang, and D. Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021.

[71] N. Peri, A. Dave, D. Ramanan, and S. Kong. Towards long-tailed 3d detection. In *Conference on Robot Learning*, pages 1904–1915. PMLR, 2023.

[72] Y. Ma, N. Peri, S. Wei, W. Hua, D. Ramanan, Y. Li, and S. Kong. Long-tailed 3d detection via 2d late fusion. *arXiv preprint arXiv:2312.10986*, 2023.

[73] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. ISSN 1424-8220. doi:10.3390/s18103337. URL https://www.mdpi.com/1424-8220/18/10/3337.

[74] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[75] J. Yin, J. Fang, D. Zhou, L. Zhang, C.-Z. Xu, J. Shen, and W. Wang. Semi-supervised 3d object detection with proficient teachers. In *European Conference on Computer Vision*, pages 727–743. Springer, 2022.

[76] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.