This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Leveraging Temporal Cues for Semi-Supervised Multi-View 3D Object Detection

Jinhyung Park[1] Navyata Sanghvi[1] Hiroki Adachi[1] Yoshihisa Shibata[2] Shawn Hunt[3]
Shinya Tanaka[3] Hironobu Fujiyoshi[2] Kris Kitani[1]
[1]Carnegie Mellon University [2]DENSO Corporation [3]DENSO International America, Inc.

Abstract

While recent advancements in camera-based 3D object detection demonstrate remarkable performance, they require thousands or even millions of human-annotated frames. This requirement significantly inhibits their deployment in various locations and sensor configurations. To address this gap, we propose a performant semi-supervised framework that leverages unlabeled RGB-only driving sequences - data easily collected with cost-effective RGB cameras - to significantly improve temporal, camera-only 3D detectors. We observe that the standard semi-supervised pseudo-labeling paradigm underperforms in this temporal, camera-only setting due to poor 3D localization of pseudo-labels. To address this, we train a single 3D detector to handle RGB sequences both forward and backward in time, then ensemble both its forwards and backwards pseudo-labels for semi-supervised learning. We further improve the pseudo-label quality by leveraging 3D object tracking to infill missing detections and by eschewing simple confidence thresholding in favor of using the auxiliary 2D detection head to filter 3D predictions. Finally, to enable the backbone to learn directly from the unlabeled data itself, we introduce an object-query conditioned masked reconstruction objective. Our framework demonstrates remarkable performance improvement on large-scale autonomous driving datasets nuScenes and nuPlan.

1. Introduction

Camera-driven 3D object detection [17, 28, 46, 67] has seen remarkable improvement in recent years, even achieving close to LiDAR-based 3D detection [56, 78, 84] performance. Improvements in 2D backbones [15, 41], advancements in query-driven 3D detection [28, 34, 67], and emphasis on temporal modeling of objects [16, 30, 36, 46] have fueled these detectors, making camera-only pipelines an indispensable component of an autonomous driving stack due to their cost efficiency and semantically accurate predictions. However, these detectors heavily depend on labor-intensive human annotations of thousands or millions of frames [7, 8, 60], which hinders their widespread deployment to new cities and sensor configurations. For truly scalable development of autonomous driving, camera-driven 3D detectors must directly leverage unlabeled data for training in a semi-supervised pipeline.

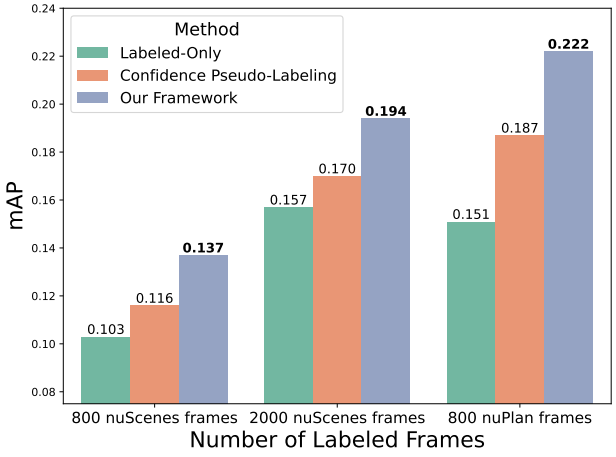

Figure 1. Our proposed semi-supervised learning framework for temporal camera-driven 3D detection substantially improves over labeled-only training and a strong pseudo-labeling baseline.

While prior work has explored semi-supervised learning for 2D and 3D detection in autonomous driving, the temporal aspect of camera-driven 3D object detection remains under-explored. Autonomous vehicles inherently capture RGB video sequences during operation, and since depth estimation is the main bottleneck for camera-based 3D detectors [43, 69], temporal RGB data offers valuable priors about the 3D world. To advance semi-supervised learning in this domain, it is crucial to fully leverage the temporal dimension of the problem. For instance, DPL [88] proposes a strong semi-supervised learning (SSL) pipeline for 3D object detection. However, they build on MonoFlex [91], a single-image 3D detector that does not use video, which places an upper bound on the performance the 3D detector can achieve. MVC-MonoDet [29] does leverage images from other timesteps during training for photometric con-

sistency loss, but it similarly only explores single-image 3D detectors. Other work [44, 48] use an additional LiDAR sensor to generate stronger pseudo-labels, but the requirement of an expensive LiDAR sensor hinders the scalability of this approach.

In this work, we propose a novel SSL framework that focuses on performant, temporal 3D detectors, fully exploits the video sequence for semi-supervised learning, and works solely with cost-effective RGB cameras. Our method incorporates both pseudo-labeling on unlabeled data as well as a self-supervised loss term directly on the RGB images. To derive accurate pseudo-labels, we focus squarely on the problem of 3D localization, the most significant challenge of camera-driven 3D detectors. Observing that pseudo-labeling errors are significantly higher for regions behind the ego-vehicle compared to regions ahead, we propose to train a single detector both on forward-running sequences as well as the reversed backward-running sequences. Without additional training costs, the detector, conditioned on consecutive timestamps, can effectively handle both directions. Using this detector for pseudo-labeling unlabeled data both forwards and backwards in time, we observe a remarkable performance improvement.

We further improve our temporal pseudo-labeling pipeline by incorporating a lightweight tracking mechanism that fills in missing detections for tracked objects. Since tracking and pseudo-labeling heavily depend on the chosen quality metric for thresholding, we eschew standard 3D detection confidence thresholding in favor of extending a 2D-3D consistency metric developed by DetMatch [46]. More specifically, DetMatch leverages separate 2D and 3D detectors trained on RGB and LiDAR, respectively, and pseudo-labels both modalities by matching 2D and 3D detections. Observing that camera-driven 3D detection is fundamentally a 2D detection task, with 3D localization and attribute prediction [68, 94], we leverage the auxiliary 2D detection head as our separate "2D detector" and match the temporal 3D predictions with the same model's 2D predictions. We find that such 2D-3D consistency, even within the *same* model, significantly improves pseudo-label quality.

Finally, to allow our model to learn directly from the unlabeled data itself, we include a masked reconstruction task on the RGB images. However, directly adding the masked autoencoder (MAE) head *worsens* 3D detection performance due to conflict between the 3D detection and reconstruction tasks. Instead, we condition the masked tokens on the 3D detection object queries. Intuitively, the 3D object queries not only encode information about the scene and objects from this timestep, but also from previous timesteps, facilitating the reconstruction task. Conversely, to solve the reconstruction task, the object queries are encouraged to focus on scene elements for both current and past timesteps, which is complementary to the temporal 3D object detection task. We find that this module further boosts SSL performance. By fully exploiting temporal priors for pseudo-labeling and self-supervision, our SSL pipeline demonstrates significant improvements using unlabeled data.

Our main contributions are as follows:

- We propose a novel SSL pipeline focusing on improving the performance of strong temporal 3D detectors.
- Our pipeline leverages forward-backward training and ensembling, 3D tracking, 2D-3D Hungarian Matching for pseudo-labeling, and an object-query conditioned masked reconstruction objective.
- We evaluate our method on nuScenes and nuPlan driving datasets, demonstrating improvement over a strong pseudo-labeling baseline as well as prior work.
- We extensively ablate each component of our pipeline.

2. Related Work

2.1. Camera-Driven 3D Object Detection

Earlier works in RGB-only 3D object detection focus on the single-view problem, predicting 3D boxes directly from CNN feature maps [5, 6, 12, 39, 45, 51, 68, 75, 94]. Later works leverage CAD models [40], formulate 3D box prediction first as keypoint regression [27, 90], leverage depth predictions before performing 3D detection on the point cloud [14, 70, 71, 74, 85], or focus on disentangling 3D box attributes [57, 68]. Many works, especially those focusing on the multi-camera setup, extract features and detect directly in 3D space. A line of work follows the Lift-Splat-Shoot (LSS) [49] framework of predicting a probability distribution over depths, then processing the depth probability-weighted, outprojected image features in BEV [17, 24, 53]. Follow-up works have improved the efficiency of BEV pooling [25, 42] and introduced depth supervision for more accurate out-projection. Another line of work follows DETR3D [72] in instantiating 3D queries that attend to each other and image features before making 3D box predictions [11, 18, 28, 35]. While most of these methods effectively handle the ill-posed problem of single-image 3D detection, their lack of temporal information upper bounds their absolute performance in large-scale driving datasets [7, 8, 60].

Recognizing that autonomous vehicles naturally capture video, follow-up works improve on the temporal aspect of camera-driven 3D detection. BEVFormer [28] and followup work [33, 79] use queries associated with BEV locations which attend to the image as well as past BEV features. BEVDet4D [16] proposes a BEV warping operation to fuse consecutive timesteps. BEVStereo [23] demonstrates the utility of stereo matching for depth prediction in 3D detection, and SOLOFusion [46] extends this work to incorporate long-term feature fusion. One line of

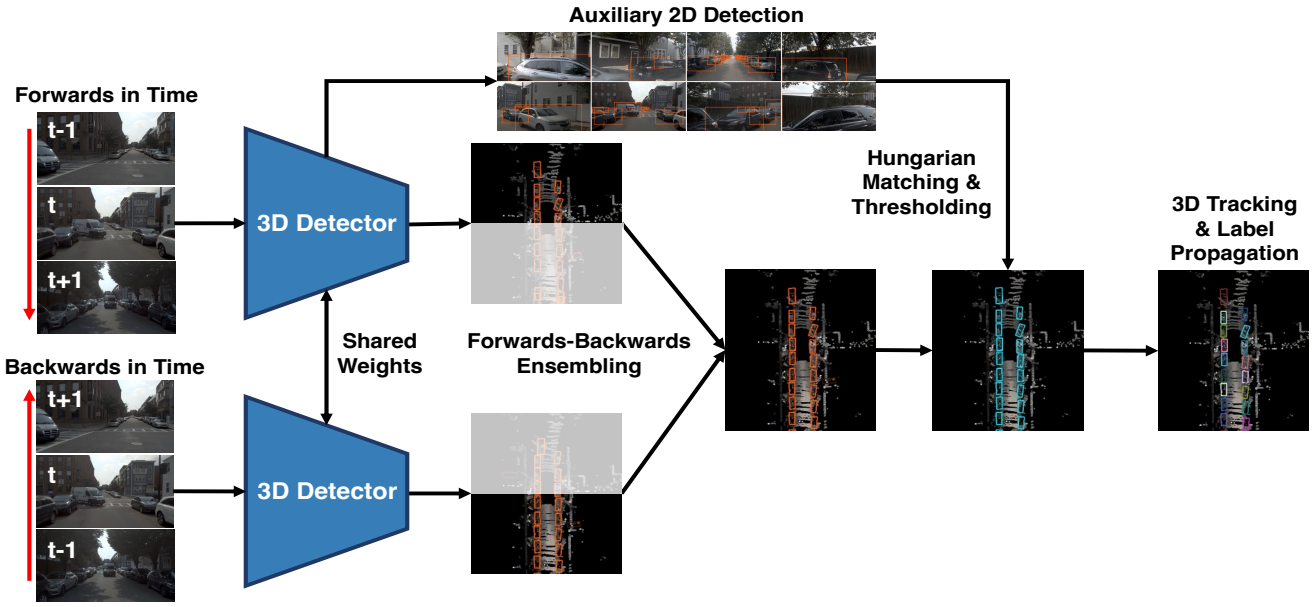

Figure 2. **Our Proposed Semi-Supervised Learning Framework.** Observing that temporal 3D object detectors consistently perform worse for objects ahead of it compared to objects behind it, we propose training a temporal 3D detector jointly on forwards-running and reversed backwards-running RGB sequences. To pseudo-label, we ensemble forwards & backwards predictions on unlabeled data, leverage 2D-3D hungarian matching to derive an accurate and 2D-consistent set of predictions, and perform 3D object tracking and label propagation. More details are in the text.

work focuses specifically on temporal processing with object queries. PETRv2 [36] and Sparse4D [30] fuses multi-timestep features into object queries through attention. Follow up works StreamPETR [73], Sparse4Dv2 [31], and Sparse4Dv3 [32] focus on recurrent feature propagation over time, dramatically accelerating inference speed. In this work, we use StreamPETR as our base model for its efficiency and strong performance, reducing the significant annotation costs of training a camera-based 3D detector.

2.2. Semi-Supervised Learning

To derive training signals from unlabeled data, SSL methods either enforce consistency between augmented variants of a single input [1, 19, 52, 55, 63] or leverage a network trained on a small amount of labeled data to pseudo-label the unlabeled data [2, 3, 20, 58, 62, 86]. While most semi-supervised methods work with image classification, many recent methods have examined this problem for 2D detection and 3D LiDAR-based detection.

In 2D detection, STAC [59] pre-trains a model on a small amount of labeled data, pseudo-labels offline, then trains the detector on the entire dataset. Instant-Teacher [93] examines augmentation strategies [4, 87], other works [22, 76, 77] modify the quality score of predictions to threshold on for pseudo labels, and a line of work focuses on exponential moving average (EMA) for predictions or the pseudo-labeler [37, 61, 82].

Many works have extended SSL to 3D detection on point clouds. SESS [92] focuses on consistency over strong-weak augmentation, 3DIoUMatch [65] improves the thresholding metric for pseudo-labeling with an IoU score, ProficientTeachers [83] improves pseudo-labels by ensembling predictions from multiple LiDAR views, and PseudoAugment [21] proposes pseudo-labeling driven augmentation. Leveraging the temporal, 3D nature of the LiDAR-based detection setting, many works focus on offline labeling and detection [9, 26, 50, 66]. Notably [50] uses extensive test-time augmentation and LiDAR point cloud aggregation to derive pseudo-labels that approach human annotation quality. While our work is similar in leveraging multiple timesteps over a driving sequence, effective pseudo-labeling in camera-driven 3D detection is a more difficult challenge due to the lack of reliable 3D signals.

2.3. SSL for Camera-Driven 3D Detection

Semi-supervised learning in camera-driven 3D detection is comparatively less explored. Mix-Teaching [81] uses a teacher network to assemble a database of background images and high-confidence object instances and trains a student network on mixed images. Other works seek additional sources of supervision such as LiDAR [44, 48, 89]. While demonstrating strong performance, these works rely on LiDAR to provide precise 3D cues that, in many cases, obviate the 3D location of an object. DPL [88] explicitly decouples the 3D and 2D attributes of pseudo-labels, and devises a homography estimation method for determining the quality of the 3D attributes. While DPL additionally identifies incorrect depth as a major issue in pseudo-labeling, their depth

gradient projection method necessitates that the 3D detector separates 3D box prediction into attribute prediction and depth estimation, which is not the case for most query and BEV based methods. Furthermore, DPL is a single-view method, under-utilizing video frames naturally present in autonomous driving. MVC-MonoDet [29] proposes to use a photometric loss between consecutive frames as a depth supervision signal for objects. While demonstrating performance improvement, the base detector is a single-frame model. To the best of our knowledge, this is the first work to examine SSL in the context of temporal, camera-driven 3D object detection.

3. Method

Figure 2 provides an overview of our pseudo-labeling pipeline. In Section 3.1 we first describe the SSL problem setting and also provide an overview of our base 3D detector StreamPETR [67]. We then detail our forward-backward pseudo-labeling algorithm in Section 3.2, provide details on the 2D-3D Hungarian Matching in Section 3.3, and propose pseudo-label refinement via tracking in Section 3.4. Finally, we introduce our object query-conditioned masked reconstruction pipeline in Section 3.5

3.1. Preliminaries

Task Overview. In semi-supervised learning, we have a small dataset of labeled samples $\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and a larger dataset of unlabeled samples $\mathcal{D}_u = \{x_i^u)\}_{i=1}^{N_u}$, where $N_l << N_u$. In the temporal, camera-driven 3D detection task, the input x consists of RGB images at a single timestep $I \in \mathbb{R}^{K \times H \times W \times 3}$, where K is the number of cameras mounted on the ego-vehicle, the camera intrinsics, and the extrinsics e_t from an IMU or GPU sensor. The label y is a set of bounding boxes $b \in \mathbb{R}^{M \times 9+C}$, where each box has 3D location, dimensions, rotation, BEV velocity, and a class label between 1 and C. The goal is to leverage a small amount of labeled data and a large amount of unlabeled data to scalably improve 3D detection performance.

Overview of StreamPETR. A strong but efficient 3D detector, StreamPETR [67] propagates a set of object queries through the driving sequence, aggregating image features from each timestep to them before predicting 3D boxes. More specifically, StreamPETR consists of a CNN backbone and a custom DETR [10] head. At each timestep, given images I_t, the backbone extracts features $f_t = \text{CNN}(I_t)$. Then, the custom DETR head processes both image features from the current timestep f_t, object queries from the previous timestep q_{t-1}, and motion attributes between the timesteps (ego-movement, predicted object velocity, time difference) and yields refined object queries for the current timestep $q_t = \text{DETR}(f_t, q_{t-1})$. Finally, an MLP predicts box parameters from the object queries $\hat{y}_t = \text{MLP}(y_t)$. StreamPETR achieves efficiency by propagating temporal information exclusively through object queries rather than through past feature maps.

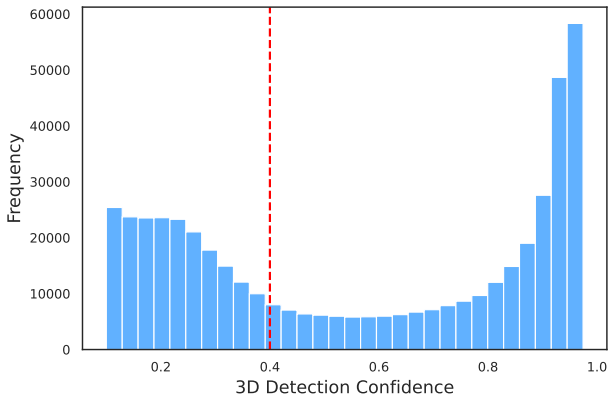

Figure 3. **Histogram of 3D Detection Confidence Preserved After Hungarian Matching.** The 2D-3D hungarian matching preserves a large number of low-confidence, yet accurate 3D detections that would have otherwise been discarded by a 0.4 pseudo-labeling confidence threshold.

3.2. Forward-Backward Pseudo-Labeling

The dominant approach in semi-supervised learning is pseudo-labeling. However, for temporal 3D detection, we find that pseudo-labels for objects in front of the vehicle are consistently worse than those behind. This is because as the car moves forward and passes objects, the temporal detector has observed objects that are now behind it for a longer period of time and can draw from a wider multi-view stereo baseline for more precise 3D localization. Furthermore, this observation is symmetric over time - if the driving sequence is seen backwards in time, the detector would have observed objects originally in front of it for longer, potentially allowing it to better pseudo-label objects in front of it as well.

Leveraging this insight, we propose to jointly train a detector on both forward-running and reversed backward-running sequences. As the temporal DETR head in StreamPETR takes as input the timestamp difference between consecutive timesteps, the detector can adaptively handle both temporal directions. After training the detector on the labeled data, we pseudo-label the unlabeled set by ensembling predictions from both a forward and backwards pass of the driving sequences. More specifically, we directly combine backward detections ahead of the ego-car and forward detections behind it. We find this intuitive method is enough to substantially improve performance — as shown in Figure 4, this achieves accurate pseudo-labels both in front and behind the vehicle.

The detector is able to achieve this level of performance by effectively leveraging privileged information. The detector's performance on objects ahead of it — which it

# Labeled	Method	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
800	Labeled Only	0.151	0.234	0.924	0.177	1.132	1.263	0.317
	Pseudo-Labeling	0.187	0.249	0.916	0.171	1.108	1.192	0.354
	Ours	**0.222**	**0.280**	0.851	0.160	1.115	1.078	0.294
2000	Labeled Only	0.219	0.293	0.849	0.156	0.936	0.936	0.284
	Pseudo-Labeling	0.269	0.314	0.803	0.151	0.947	1.051	0.300
	Ours	**0.271**	**0.330**	0.810	0.150	0.934	0.868	0.292
4000	Labeled Only	0.291	0.384	0.807	0.140	0.765	0.693	0.210
	Pseudo-Labeling	0.304	0.405	0.806	0.137	0.726	0.600	0.201
	Ours	**0.331**	**0.410**	0.778	0.137	0.782	0.672	0.187

Table 1. **Comparison of methods on the nuPlan dataset [8] with different labeled sample sizes.** Our framework achieves better 3D localization due to more accurate pseudo-labels.

achieved by moving backwards in time — will always be ahead of the detector's forward performance on those regions. This low-cost addition substantially improves the performance of the final trained model.

3.3. 2D-3D Hungarian Matching

While generating pseudo labels by thresholding on detection confidence can achieve good performance [38, 47], we draw inspiration from DetMatch's Hungarian Matching-based pseudo-labeling to improve pseudo-label quality for free. DetMatch proposes to match 2D and 3D detections from separate RGB and LiDAR models, respectively, and threshold on their matching cost instead. While DetMatch had cited agreement between disparate models and modalities as the cause of this method's success, we find that using even 2D and 3D detections from *the same model and modality* can improve performance. More specifically, StreamPETR is equipped with an auxiliary 2D detection head for additional training-time supervision. We find that in terms of 2D detection performance, the 2D head's predictions achieve higher mAP compared to the projections of the 3D head (53.8 AP@50 vs 40.5 AP@50). Motivated by this discrepancy, we perform Hungarian Matching between these 2D and 3D detections, minimizing focal loss, GIoU [54], and 2D box parameter difference. Thresholding on the matching cost, we see a boost in pseudo-label quality, and this method retains low-confidence but accurate 3D detections that would otherwise have been discarded, as shown in Figure 3. We emphasize that this addition imposes no additional cost to training or inference, and it adds negligible runtime to the pseudo-labeling procedure.

3.4. 3D Tracking and Label Propagation

While the detector is able to maintain most objects through time, in the presence of uncertainty, distance, or significant occlusion, we see object detections disappear between frames. Such missing detections negatively influence the final trained model and exacerbates the problem of object impermanence. To address this issue, we propose a straightforward but effective object tracking-based fix. More specifically, we use a predicted velocity-based tracking pipeline [84], where each tracklet from the previous timestep is matched with the current prediction moved backwards its predicted velocity. Notably, unlike LiDAR where object locations, and hence velocities, are more 3D accurate, camera-based predictions have more significant per-frame velocity errors. To alleviate this, we maintain a velocity of each tracklet which we set as the exponential moving average (EMA) of its associated predictions' center location changes over time. Then, to match past tracklets and current detections, we use velocity estimates to move tracklets forward and detections backwards a half-timestep. This robustly handles static initialization and uncertain velocity predictions. We use a greedy, multi-object, distance-based tracker [84] for its good 3D performance.

3.5. Object Query-Conditioned Masked Reconstruction

In addition to our pseudo-labeling framework, we also enable the network to learn directly from unlabeled samples. More specifically, we formulate an object query-conditioned masked reconstruction loss. First, we uniformly randomly mask patches in the input images and process it with a CNN once more:

$$f_t^{mask} = \text{CNN}(M \odot I_t) \tag{1}$$

At this stage, f_t^{mask} encodes information about the visible part of the input image. While prior work directly input f_t^{mask} into a masked decoder for pre-training, we find that doing so significantly *hurts* the performance of our 3D detector. The network focuses on optimizing the auxiliary loss at the expense of the main task loss. To more explicitly tie the detector to the reconstruction, we update the masked features f_t^{mask} by attending onto the object queries output from the temporal DETR head q_t. More specifically:

$$\tilde{f}_t^{mask} = \text{TransformerDecoder}(f_t^{mask}, q_t, q_t) \tag{2}$$

# Labeled	Method	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
800	Labeled Only	0.103	0.191	0.953	0.352	1.299	1.359	0.299
	UniPAD + Pseudo.	0.116	0.198	0.954	0.359	1.216	1.075	0.286
	Ours	**0.137**	**0.221**	0.940	0.340	1.339	0.940	0.258
2000	Labeled Only	0.157	0.242	0.894	0.318	0.988	0.916	0.250
	UniPAD + Pseudo.	0.170	0.282	0.896	0.325	0.894	0.685	0.230
	Ours	**0.194**	**0.293**	0.902	0.315	0.932	0.664	0.226

Table 2. **Comparison of methods on the nuScenes dataset [7].** Our framework significantly improves over both the Labeled-Only baseline as well as UniPAD [80].

# Labeled	Method	mAP ↑	Car	Tru.	Bus	Trai.	Con.	Bic.	Mot.	Ped.	Traf.	Bar.
800	Labeled Only	0.103	0.247	0.054	0.048	0.018	0.003	0.035	0.019	0.116	0.278	0.208
	UniPAD + Pseudo-Lab	0.116	0.267	**0.061**	0.026	0.016	0.007	0.040	0.013	0.138	0.318	0.270
	Ours	**0.137**	**0.284**	0.060	**0.075**	**0.023**	**0.008**	**0.074**	**0.041**	**0.182**	**0.351**	**0.271**
2000	Labeled Only	0.157	0.346	0.091	0.124	0.020	**0.007**	0.085	0.080	0.184	0.336	0.299
	UniPAD + Pseudo-Lab	0.170	0.359	0.098	0.089	**0.030**	0.005	0.102	0.079	0.188	0.382	0.367
	Ours	**0.194**	**0.384**	**0.101**	**0.134**	**0.030**	0.006	**0.137**	**0.122**	**0.239**	**0.420**	**0.371**

Table 3. **mAP of methods across 10 classes on the nuScenes dataset [7].** Our method consistently outperforms prior work over both easy and difficult classes.

where TransformerDecoder(f_t^{mask}, q_t, q_t) performs self-attention between the masked features f_t^{mask} and then performs cross-attention, pulling information from q_t to refine f_t^{mask}. By conditioning the masked reconstruction on the predicted object queries, this enables gradient flow from the reconstruction loss to influence the temporal DETR head directly. To minimize the reconstruction loss, the object queries need to retain information about scene elements from both current and past timesteps, which is complementary to the main detection task. We find that this self-supervised objective improves performance in conjunction with our pseudo-labeling method.

4. Experiments

4.1. Datasets and Evaluation Metrics

The nuScenes Dataset. nuScenes [7] includes 1000 scenes split into 700, 150, and 150 sequence for training, validation, and online testing. Each sequence is 20s long and is annotated at 2Hz, resulting in 28130 total training frames. The ego-vehicle has 6 high-resolution 900 x 1600 cameras which cover the entire 360° field-of-view. The annotation covers 10 classes: car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle, bicycle, barrier, and traffic cone, and methods are evaluated with mean Average Precision (mAP), Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), Average Attribute Error (AAE), and a combined metric Nuscenes Detection Score (NDS). For semi-supervised learning, we split the training set into either 800 or 2000 labeled frames, and take the remainder as the unlabeled set. Note that as we work in the temporal 3D detection setting, we sample frames at the sequence level in chunks of 40 consecutive frames.

The nuPlan Dataset. nuPlan [8] is a large-scale planning dataset that releases sensor data (RGB videos) and 3D box annotations for 120 hours of driving. This dataset contains 1065 driving logs, which totals 590,499 frames across four regions — Las Vegas with 455k frames, Boston with 64k, Pittsburgh with 51k, and Singapore with 19k. Each frame contains 8 high-resolution RGB images at 1080 x 1920 resolution, as well as 3D bounding box labels. We use the most common car and pedestrian classes on this dataset. We choose this dataset to demonstrate the scalability of our method leveraging a large amount of unlabeled data.

In our work, we leverage 800, 2000, or 4000 labeled frames, keeping the rest as unlabeled data to learn from. Considering the uneven distribution of frames from each city, we evenly split the labeled data frames across the four cities. Notably, we select sequences at the *log* level, which, on average, has 600 frames (5 minutes) of driving data. We do this to mimic realistic data collection scenarios where annotators label 3D boxes continuously through time. We use the same metrics as in nuScenes.

4.2. Implementation Details

We develop our SSL pipeline on the performant temporal 3D detector StreamPETR [67]. For the backbone, we use the ConvNeXt-S [41] backbone with 50M parameters pretrained by SparK [64]. As SparK pretrains the backbone on ImageNet-1k [13] for masked reconstruction, we adopt its pretrained decoder as our masked reconstruction head as well, affording the backbone a continuous transition in supervision. We split our training two stages. First, the detector is trained on the small amount of labeled data until convergence. Then, we leverage our pseudo-labeling pipeline

	Method	mAP ↑	NDS ↑	mATE ↓
(a)	Labeled-Only	0.155	0.232	0.938
(b)	+ Pseudo-Labeling	0.182	0.247	0.924
(c)	+ Fw-Bw	0.208	0.267	0.855
(d)	+ Masked Recon	0.209	0.265	0.862
(e)	+ Masked Recon & Query Cond.	0.209	0.271	0.847
(f)	+ 3D Tracking	0.215	0.275	0.848
(g)	+ Hung. Matching	0.219	0.278	0.850

Table 4. **Ablation study of the major components of our framework.** Results are on nuPlan with 800 labeled and 60k unlabeled samples.

to annotate the unlabeled data offline. Finally, we continue training our network on both the labeled and unlabeled data, evenly sampling between them. For our final models in the second stage, we disable the forwards-backwards flipping to allow the model to focus solely on the deployment setting. We use a masking ratio of 0.3. Additional implementation details can be found in the supplementary.

4.3. Results on nuPlan

We present results on the large-scale nuPlan dataset in Table 1. As we are the first to address the problem of SSL for temporal 3D object detection, we develop a straightforward but strong pseudo-labeling baseline based on confidence thresholding. We note that this baseline already significantly improves performance over the labeled-only baseline. Our proposed pipeline, by generating higher-quality pseudo-labels and leveraging a self-supervised objective, further improves results in for all # of labeled samples. The performance gain is especially notable for the low-resource setting of just 800 labeled samples, where we see a substantial +3.5 mAP and +3.1 NDS.

4.4. Results on nuScenes

We also evaluate our pipeline on the 10 detection classes in the nuScenes dataset. For comparison, we include UniPAD [80], a *self*-supervised pre-training method designed for nuScenes. Specifically, UniPAD builds on the camera-based 3D detector [24], pre-training its 2D backbone using an RGB and depth rendering loss. Although this approach leverages privileged information (e.g., depth from LiDAR), it does not rely on human annotations during the pre-training phase. To enhance this baseline, we further train UniPAD on the limited labeled dataset, use it to pseudo-label the unlabeled data, and then continue training. We note that UniPAD also uses the ConvNeXt-S [41] initially pre-trained by SparK [64].

The results are in Table 2. While both UniPAD and our method significantly improve over the labeled-only baseline, our pipeline consistently demonstrates stronger performance. We further examine mAP over each of the 10 classes in nuScenes in Table 3 and observe a similar trend. While all methods struggle on rare classes like trailer and construction vehicle, our method significantly improves on common classes (car and pedestrian) as well as on classes that are difficult to localize in 3D (bicycle, motorcycle).

4.5. Ablations and Analysis

4.5.1. Ablation of Components

We carefully ablate the major components of our framework in Table 4. In this table, we focus on three metrics - mAP, which captures overall detection performance, NDS, which balances detection with attribute prediction, and mATE, which examines localization of the object center. Starting with the Labeled-Only baseline (a), adding pseudo-labeling (b) improves mAP substantially, but mATE does not improve significantly. This is due to pseudo-labeling expanding the semantic boundary of the classes and enabling the model to detect previously uncertain objects. Despite this, the 3D localization of such objects remains poor due to the suboptimal 3D localization of the pseudo-labels themselves. Adding our forwards-backwards ensembling approach (c), however, substantially boosts mATE in addition to mAP and NDS. By allowing the model to adopt a more diverse multi-view stereo baseline, the resulting pseudo-labels have more precise 3D center estimation.

We then add the self-supervised objective masked reconstruction loss (d) as an independent decoder head as is done in SparK [64]. We find that this hurts performance due to the backbone alone handling reconstruction in addition to 3D detection. However, by conditioning the masked reconstruction on object queries predicted from the detection head (e), mATE improves further. This demonstrates that by looping in the detection head, the reconstruction task is now *complementary* to 3D detection. Adding 3D tracking (f) further improves performance by ensuring pseudo-labels remain more consistent over time, and replacing confidence thresholding with the cost-free Hungarian Matching (g) further improves both mAP and NDS. We hypothesize that tracking and Hungarian Matching improve mAP as opposed to mATE because these modules mainly focus on reducing false positives and false negatives as opposed to refining the 3D location of pseudo-labels.

4.5.2. Qualitative Analysis of Pseudo-Labels

We also show qualitative results of our pseudo-labeling algorithm with 800 labeled samples on nuPlan in Figure 4. Our framework more consistently captures difficult objects — those ahead of the vehicle and objects at a far depth. The more precise and consistent pseudo-labels in turn improve performance of the final trained model. Additional visualizations on pseudo-labels and reconstruction can be found in the supplementary.

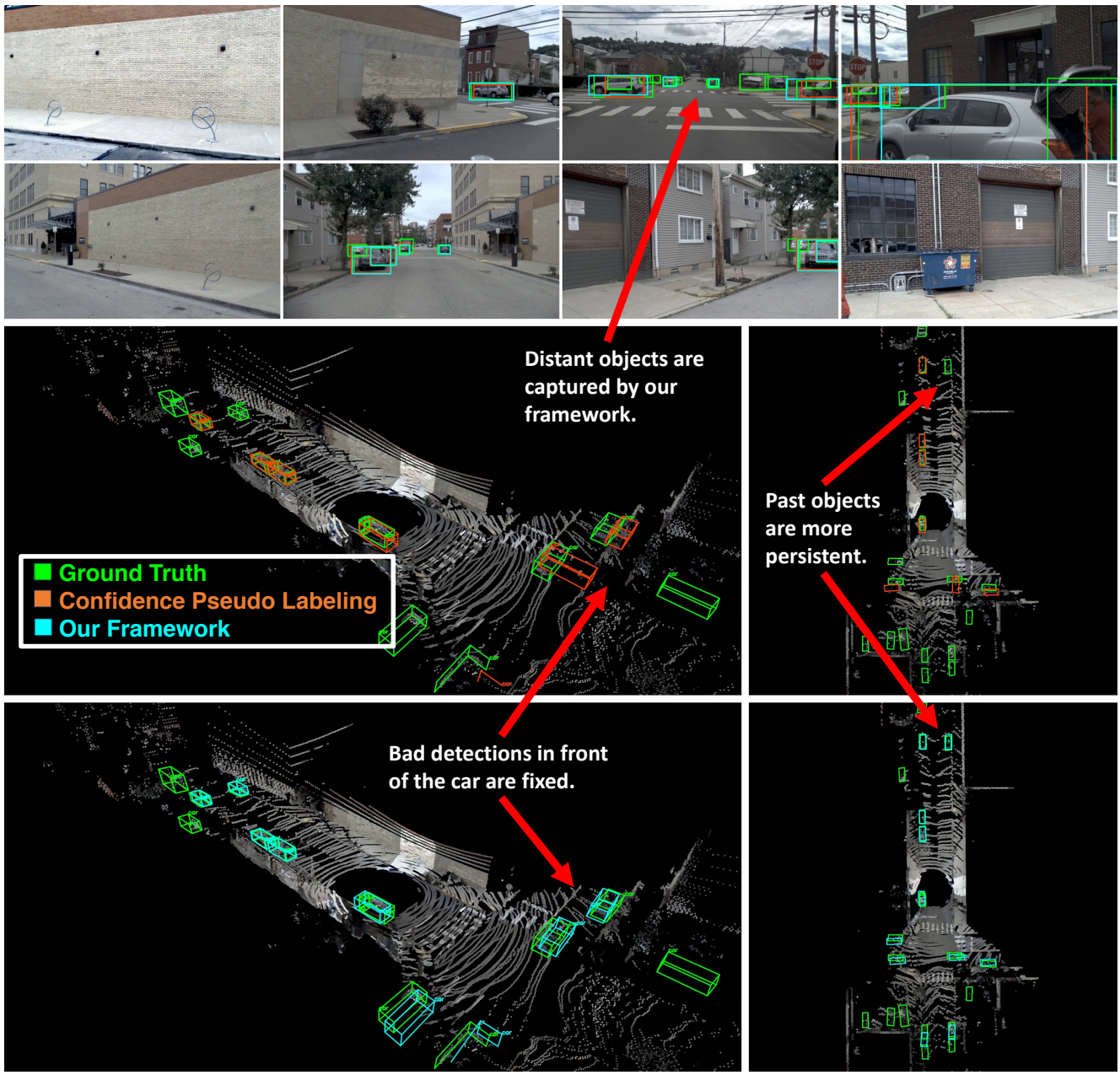

Figure 4. **Qualitative results on pseudo-labeling.** We visualize pseudo labels from confidence thresholding in orange and our framework in blue. Our pseudo labels capture distant objects with 2D-3D hungarian matching, achieve better performance in front of the vehicle due to forward-backward ensembling, and maintains persistency of past objects through tracking.

5. Conclusion

In this paper, we present a novel semi-supervised learning pipeline for temporal 3D object detection, designed to fully leverage the temporal dynamics of RGB video data. Our approach addresses the challenges of 3D localization in camera-driven detectors by introducing forward-backward ensembling for pseudo-labeling, leveraging 2D-3D Hungarian Matching for consistency, and enhancing self-supervised learning with an object-query conditioned masked reconstruction task. Through extensive evaluation on nuScenes and nuPlan datasets, we demonstrate that our method achieves significant performance gains over a strong pseudo-labeling baseline, highlighting the value of temporal modeling in semi-supervised 3D detection. Our framework paves the way for scalable, cost-effective deployment of camera-based 3D detection systems, leveraging easy-to-collect unlabeled data to develop accurate 3D detectors.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 3

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020. 3

[5] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 2

[6] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 2

[7] 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*, pages 11621–11631, 2020. 1, 2, 6

[8] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 1, 2, 5, 6

[9] Benjamin Caine, Rebecca Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Z. Chen, and Jonathon Shlens. Pseudo-labeling for scalable 3d object detection. *ArXiv*, abs/2103.02093, 2021. 3

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4

[11] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar parametrization for vision-based surround-view 3d detection. *arXiv preprint arXiv:2206.10965*, 2022. 2

[12] Xiaozhi Chen, Huimin Ma, Jixiang Wan, B. Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6526–6534, 2017. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[14] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, pages 1000–1001, 2020. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[16] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:/2203.17054*, 2021. 1, 2

[17] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2

[18] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*, 2022. 2

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 3

[21] Zhaoqi Leng, Shuyang Cheng, Benjamin Caine, Weiyue Wang, Xiao Zhang, Jonathon Shlens, Mingxing Tan, and Dragomir Anguelov. Pseudoaugment: Learning to use unlabeled data for data augmentation in point clouds. In *European conference on computer vision*, pages 555–572. Springer, 2022. 3

[22] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry Davis. Rethinking pseudo labels for semi-supervised object detection. *ArXiv*, abs/2106.00168, 2021. 3

[23] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*, 2022. 2

[24] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *arXiv preprint arXiv:2206.00630*, 2022. 2, 7

[25] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 2

[26] Yingwei Li, Charles R Qi, Yin Zhou, Chenxi Liu, and Dragomir Anguelov. Modar: Using motion forecasting for 3d object detection in point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9329–9339, 2023. 3

[27] Zhuoling Li, Z. Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploit-

ing depth clues for reliable monocular 3d object detection. *ArXiv*, abs/2205.09373, 2022. 2

[28] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2

[29] Qing Lian, Yanbo Xu, Weilong Yao, Yingcong Chen, and Tong Zhang. Semi-supervised monocular 3d object detection by multi-view consistency. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 715–731. Springer, 2022. 1, 4

[30] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 1, 3

[31] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023. 3

[32] Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4d v3: Advancing end-to-end 3d detection and tracking. *arXiv preprint arXiv:2311.11722*, 2023. 3

[33] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 2

[34] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 1

[35] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 531–548. Springer, 2022. 2

[36] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 1, 3

[37] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Péter Vajda. Unbiased teacher for semi-supervised object detection. *ICLR*, 2021. 3

[38] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 5

[39] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020. 2

[40] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2

[41] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1, 6, 7

[42] 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. IEEE, 2023. 2

[43] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 1

[44] Xinzhu Ma, Yuan Meng, Yinmin Zhang, Lei Bai, Jun Hou, Shuai Yi, and Wanli Ouyang. An empirical study of pseudo-labeling for image-based 3d object detection. *arXiv preprint arXiv:2208.07137*, 2022. 2, 3

[45] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 2

[46] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 1, 2

[47] Jinhyung D. Park, Chenfeng Xu, Yiyang Zhou, Masayoshi Tomizuka, and Wei Zhan. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *European Conference on Computer Vision*, 2022. 5

[48] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *European Conference on Computer Vision*, pages 123–139. Springer, 2022. 2, 3

[49] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 2

[50] C. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa T. Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6130–6140, 2021. 3

[51] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8851–8858, 2019. 2

[52] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with

ladder networks. *Advances in neural information processing systems*, 28, 2015. 3

[53] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2

[54] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[55] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 3

[56] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10526–10535, 2020. 1

[57] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 2

[58] 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. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 3

[59] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *ArXiv*, abs/2005.04757, 2020. 3

[60] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 2

[61] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, 2021. 3

[62] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[63] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[64] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv:2301.03580*, 2023. 6, 7

[65] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14610–14619, 2021. 3

[66] Jianren Wang, Haiming Gang, Siddarth Ancha, Yi-Ting Chen, and David Held. Semi-supervised 3d object detection via temporal graph neural networks. *2021 International Conference on 3D Vision (3DV)*, pages 413–422, 2021. 3

[67] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 1, 4, 6

[68] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua 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. 2

[69] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 1

[70] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 2

[71] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 2

[72] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. *arXiv preprint arXiv:2110.06922*, 2020. 2

[73] Zitian Wang, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Object as query: Equipping any 2d object detector with 3d detection ability. *arXiv preprint arXiv:2301.02364*, 2023. 3

[74] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *ICCV Workshops*, 2019. 2

[75] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, 2018. 2

[76] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3040–3049, 2021. 3

[77] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3060–3069, 2021. 3

[78] Yan Yan, Yuxing Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors (Basel, Switzerland)*, 18, 2018. 1

[79] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. *arXiv preprint arXiv:2211.10439*, 2022. 2

[80] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15238–15250, 2024. 6, 7

[81] Lei Yang, Xinyu Zhang, Li Wang, Minghan Zhu, Chuang Zhang, and Jun Li. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. *arXiv preprint arXiv:2207.04448*, 2022. 3

[82] Qize Yang, Xihan Wei, Biao Wang, Xia Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5937–5946, 2021. 3

[83] 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. 3

[84] 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. 1, 5

[85] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019. 2

[86] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *ArXiv*, abs/2110.08263, 2021. 3

[87] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018. 3

[88] Jiacheng Zhang, Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Decoupled pseudo-labeling for semi-supervised monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16923–16932, 2024. 1, 3

[89] Weijia Zhang, Dongnan Liu, Chao Ma, and Weidong Cai. Odm3d: Alleviating foreground sparsity for enhanced semi-supervised monocular 3d object detection. *ArXiv*, abs/2310.18620, 2023. 3

[90] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3288–3297, 2021. 2

[91] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 1

[92] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11076–11084, 2020. 3

[93] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 3

[94] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2