PLOT: Pseudo-Labeling via Object Tracking for Monocular 3D Object Detection

Seokyeong Lee*[2], Sithu Aung*[3], Junyong Choi[4], Seungryong Kim[2], Ig-Jae Kim[1,5], and Junghyun Cho[1,5,6]

[1] Korea Institute of Science and Technology (KIST) [2] KAIST AI
[3] VRG, FEE, Czech Technical University in Prague [4] Samsung Research
[5] AI-Robotics, KIST School, University of Science and Technology (UST)
[6] Yonsei-KIST Convergence Research Institute, Yonsei University
{shapin94,seungryong}@kaist.ac.kr sithu.aung@fel.cvut.cz
{drjay,jhcho}@kist.re.kr

Abstract. Monocular 3D object detection is crucial for scalable perception across fields like autonomous driving, robotics, and surveillance. However, progress is hindered by limited 3D annotations and the inherent ambiguity of single-image geometry. Existing methods often rely on strong geometric assumptions or carefully curated datasets, which limit their applicability to real-world scenarios. In this paper, we present **PLOT** (**P**seudo-**L**abeling via **O**bject **T**racking), a framework that generates 3D annotations from monocular videos without auxiliary sensors or model retraining. PLOT tracks object and background trajectories to estimate camera motion and perform object association in pose-unknown settings. These trajectories provide point correspondences that align frame-wise pseudo-LiDARs, which are then fused via simple optimization into a unified object shape robust to occlusion and viewpoint shifts. Recognizing temporal coherence as a fundamental requirement for reliable shape fusion and video perception, we design a global object memory that preserves consistent object identities across frames. PLOT achieves robust annotation quality and strong generalization on both M3OD video benchmarks and in-the-wild videos, proving its effectiveness across diverse and unconstrained domains. Project page: https://plot-eccv.github.io.

Keywords: Pseudo-labeling, Monocular 3D Object Detection

1 Introduction

Monocular 3D object detection (M3OD) aims to recover the 3D geometry, pose, and extent of objects from a single RGB camera, offering a cost-effective alternative to LiDAR-based perception systems [14, 36, 46]. Despite substantial

* Equal contribution (Work done in KIST).

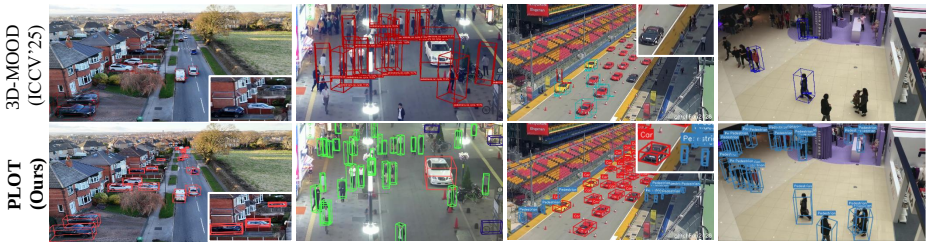


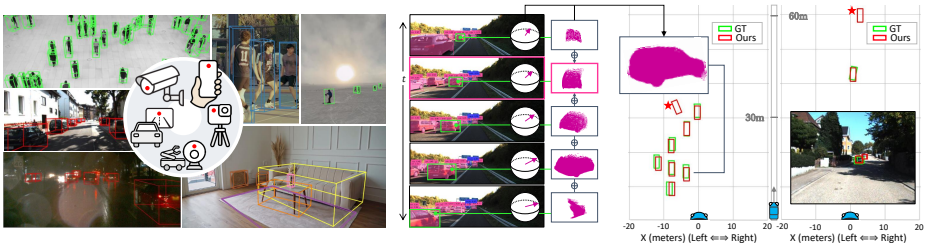
Fig. 1: Zero-shot estimation on MOT17 [34] and Pexels [8]: Predictions from (top) an open-set M3OD model [55] and (bottom) our pseudo-labeling method. Our method performs well on out-of-domain camera views (zoom in for a better view).

progress, M3OD remains fundamentally ill-posed due to depth–scale ambiguity and the lack of direct geometric supervision, which has led most methods to rely on curated benchmarks collected in sensor-rich environments, such as autonomous driving [6, 11, 49] and indoor scenes [2, 9]. Consequently, models developed under these settings often exhibit limited generalization when deployed in unconstrained scenarios with moving cameras, unknown poses, and diverse viewpoints, as evidenced by their poor zero-shot performance on in-the-wild videos [8, 34] (Fig. 1).

To alleviate the scarcity of high-quality 3D annotations, pseudo-labeling and weakly supervised approaches have been proposed to generate 3D supervision without explicit LiDAR or multi-view data [19, 29, 60]. However, most existing pipelines operate on single images, limiting their ability to resolve occlusions and disambiguate object scale and orientation. These single-frame formulations produce partial and noisy geometric observations, which propagate systematic errors into downstream training and evaluation. Moreover, many methods rely on sensor poses or hand-crafted geometric heuristics [19, 48], restricting their scalability and robustness across domains.

In this paper, we propose **PLOT** (**P**seudo-**L**abeling via **O**bject **T**racking), a framework that **generates reliable 3D annotations directly from monocular videos without requiring auxiliary sensors, camera poses, or model retraining.** Monocular videos are abundant and encode rich geometric cues over time, offering a natural source of supervision beyond single images. However, effectively leveraging such videos introduces two fundamental challenges that have remained underexplored in prior work.

The first challenge is maintaining long-term object identity consistency in pose-unknown videos. In realistic settings, clutter, occlusion, and intermittent detection failures frequently lead to fragmented trajectories and identity switches, degrading the quality of aggregated labels. PLOT addresses this challenge by aligning frame-wise object observations through dense point tracking [13] and introducing a *Global Object Memory* (GOM) that models object persistence over time, enabling robust association even under occlusion and detector noise.



(a) Labeling results in diverse domains (b) Multi-frame aggregation & BEV comparison with GT

Fig. 2: PLOT generates accurate 3D labels directly from monocular videos without requiring auxiliary sensors or training, as illustrated in (a) qualitative results across diverse scenarios. Furthermore, (b) our object tracking and aggregation pipeline produces shape-complete pseudo-LiDARs, yielding BEV maps comparable to ground truth and can identify miss-labeled objects (marked with a red star).

The second challenge concerns shape incompleteness and its impact on estimating key 3D attributes such as object size, orientation, and position. While recent work [48] attempts to leverage temporal aggregation from monocular videos, they often rely on simplified matching strategies or limited association mechanisms, resulting in incomplete and fragmented object geometry. Consequently, downstream geometric estimators operate on partial observations that do not adequately capture the full object extent. For instance, orientation estimation methods such as PCA-based inference implicitly assume sufficiently complete geometry; when applied to sparse or occluded point sets, these assumptions break down, leading to biased or unstable predictions. PLOT addresses this issue by aggregating object observations across time via tracking-based motion estimation and trajectory-guided shape fusion, progressively recovering more complete object geometry. Unlike global scene reconstruction methods [25,52,53,61], which primarily target static backgrounds and lack object-level reasoning, our approach enables object-centric shape completion tailored to dynamic instances.

We validate PLOT on multiple M3OD benchmarks [11,27,49] and unconfined monocular videos [8,10,34], demonstrating strong performance across diverse domains and camera motions (see Fig. 2). Overall, PLOT provides a practical and scalable solution for generating high-quality 3D supervision in pose-unknown, real-world video settings, expanding the applicability of M3OD beyond curated sensor environments.

2 Related Work

Supervised Monocular 3D Object Detection (M3OD). Early supervised methods [4, 20, 31, 38, 51] achieve progress in curated settings but remain restricted by scarce 3D annotations. To alleviate this, several approaches use external data such as video [5, 54], LiDAR [18], or CAD models [30, 37]. Recent works move toward open-set detection, exemplified by CubeR-CNN [3] and

successors [55, 56, 59], exploiting large-scale benchmarks [3] and 2D foundation models [35,43]. While these broaden domain coverage, they remain tied to 3D supervision and struggle to generalize consistently in unconstrained videos (Fig. 1).

Weakly-supervised M3OD. Given the challenges of cost and scarcity of 3D annotations, several works [12, 21, 41, 42] have pioneered weakly-supervised M3OD. WeakM3D [39] estimates 3D attributes given raw LiDAR points, while WeakMono3D [50] and subsequent work [12] utilize multi-view constraints. SKD-WM3D [21] proposes a self-teaching pipeline to lift 2D features into 3D space using depth completion, while VGW-3D [17] uses perspective-invariant error prediction and decoupled visual guidance. Importantly, MonoGRNet [42] introduces a general-purpose model that utilizes video priors. However, these methods require sophisticated priors, which limit their scalability. GGA [58] addresses this issue by introducing a generalizable weakly-supervised M3OD framework that leverages general geometric priors and 2D ground-truths. VSRD [29] proves the efficacy of pseudo-labeling and silhouette rendering in weakly-supervised M3OD.

Tracking for 3D Perception. Tracking has recently gained attention in video-based 3D perception for its potential to model scene dynamics over time. Seurat [7] and ViPE [16] utilize dense tracking across frames to enhance 3D perception in dynamic scenes. While Seurat focuses on depth map estimation, ViPE integrates tracking signals for broader scene-level understanding. However, neither method addresses object-level geometry or pose. Vid2CAD [33] targets CAD model alignment using multi-frame tracking, but focuses on aligning indoor objects to canonical CAD poses without addressing real-world clutter. In contrast, our approach uses off-the-shelf dense tracker [13] to associate object observations over time and estimate relative camera motion—enabling pose-free, training-free 3D labeling from monocular videos without bundle adjustment or calibration.

Pseudo-labeling for M3OD. To address the scarcity of 3D annotations, pseudo-labeling has emerged as a viable alternative for enabling M3OD without explicit 3D supervision. In particular, VSRD [29] introduced a multi-view-based auto-labeling approach, showing that reliable pseudo-labels can enhance weakly supervised training. OVM3D-Det [19], meanwhile, proposes a zero-shot labeling framework that utilizes open-vocabulary 2D detectors [15,45] and LLM priors [1] to estimate 3D box attributes without training. However, it operates on single images and thus remain sensitive to occlusions, noise, and scale ambiguities, limiting their applicability in more complex or dynamic scenes. MonoSOWA [48], a recent work, similarly leverages per-frame 2D mask association for pseudo-label generation, but the reliance on known camera poses with naive mask-based tracking limits its scalability, and does not address cluttered environments explicitly.

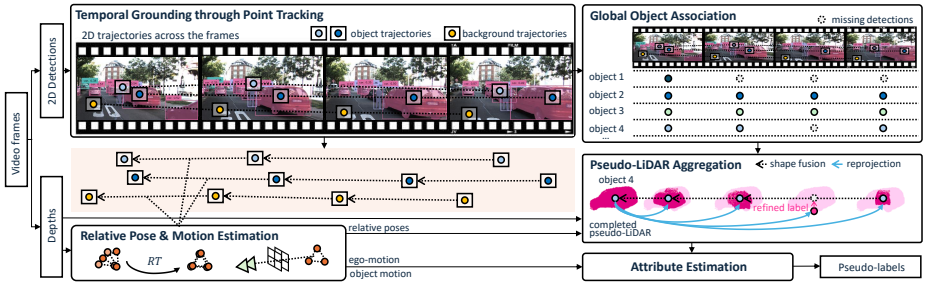


Fig. 3: Overall architecture of PLOT. Given monocular videos, we extract 2D detections and depth, and track points to obtain temporally grounded correspondences. These are used to estimate relative poses and camera motions across frames, enabling shape fusion and orientation estimation. A global object association module refines trajectories and recovers missing instances. Finally, completed pseudo-LiDARs are reprojected to each frame for consistent 3D attribute annotation.

3 Method

We propose PLOT, a framework for generating 3D annotations from monocular videos without auxiliary sensors or model retraining. As illustrated in Fig. 3, PLOT leverages off-the-shelf detectors and depth estimators to extract 2D masks and metric depth maps, which are combined via dense point tracking to form temporally grounded correspondences (Sec. 3.1). These are used to estimate relative poses through point-based registration (Sec. 3.2), enabling both shape fusion and motion analysis. To maintain identity consistency and recover missed instances, we introduce a global object memory (GOM) that refines labels over time (Sec. 3.3). Finally, object pseudo-LiDARs are constructed by aggregating registered points across frames and projecting them back to each time frame for consistent attribute estimation (Sec. 3.4).

3.1 Temporal Grounding through Point Tracking

To enable robust object association and temporally grounded geometric reasoning, we distinguish between two types of 2D object mask instances. *Predicted masks* are obtained at each frame via an open-vocabulary detector (GSAM) [45], while *tracked masks* are generated by propagating previous-frame instances using a point-based 2D tracker [13]. The former reflects per-frame shape variations, whereas the latter preserves point-level correspondences necessary for relative pose estimation and shape fusion.

Specifically, given the k-th object mask M_k^t from an open-vocabulary 2D detector [45] in each frame image I_t where $t \in \mathcal{T}$ (the full set of video time steps), we generate tracked masks $M_k^{\acute{t}\leftarrow t}$ at other time steps $\acute{t}$ by tracking M_k^t to $\acute{t}$ using a dense point tracker [13]:

$$M_k^{\acute{t}\leftarrow t} = T_{\acute{t}\leftarrow t}(M_k^t), \quad \acute{t} \in \mathcal{T} \setminus \{t\}, \quad k \in \mathcal{K}^t. \tag{1}$$

Here, $\mathcal{K}^t$ represents the total number of objects predicted at time t, and $T_{\acute{t}\leftarrow t}(\cdot)$ denotes dense tracking from t to $\acute{t}$. Note that $\mathcal{K}^t$ can vary between frames due to missing labels caused by distance- or occlusion-related challenges, or simply detector errors.

Thus, the predicted and tracked masks may diverge, despite originating from the same stem. We therefore perform explicit matching to associate them and maintain consistent identity over time. Using frames from two distinct time steps s and t, we apply bipartite matching using the Hungarian algorithm [23] between the set of tracked masks and the predicted masks, as follows:

$$\hat{\sigma} = \arg\max_{\sigma \in \mathcal{S}_n} \sum_i \text{IoU}(M_i^{s\leftarrow t}, M_{\sigma(i)}^s), \tag{2}$$

where $\hat{\sigma}$ represents the optimal assignment of predicted masks to tracked masks, determined by maximizing the sum of the Intersection of Union (IoU) scores, while $\mathcal{S}_n$ denotes the set of all possible permutations of assignments. Tracked masks are used to estimate rigid motion, while matched predicted masks serve as the basis for shape fusion.

In parallel, we track background points to infer camera motion and disentangle objects from static scene structure. Specifically, we sample background masks M_{bg}^t from regions outside object masks and track them over time as follows:

$$M_{bg}^{\acute{t}\leftarrow t} = T_{\acute{t}\leftarrow t}(M_{bg}^t), \quad \acute{t} \in \mathcal{T} \setminus \{t\}, \quad M_{bg}^t \subset I_t \setminus M^t, \quad M^t = \bigcup_{k \in \mathcal{K}^t} M_k^t. \tag{3}$$

These object and background correspondences are used to determine the motion and dynamic states of both the camera and objects, as demonstrated below.

3.2 Pose and Motion Estimation via Point Trajectories

In addition to object shape fusion, which relies on relative pose estimation across frames, it is equally important to recover the motion states of the camera. Accurate camera motion estimation enables downstream tasks such as object orientation reasoning, trajectory prediction, and relative positioning. To achieve this, we derive the trajectories of both background points and object points across consecutive frames. Background point trajectories provide constraints for robust ego-motion estimation, while object point trajectories additionally capture individual object dynamics.

Camera Motion Estimation via Background Trajectories. To estimate the camera motion, we leverage the trajectories of the background points, which are lifted to 3D using monocular depth predictions [40]. This yields a set of 3D motion trajectories $\{\mathbf{p}_i^t | i \in M_{bg}^t, t \in \mathcal{T}\}$. The camera motion between a reference frame r (typically the first frame) and each frame t is parameterized by rotation $\mathbf{R}_c^t \in SO(3)$ and translation $\mathbf{t}_c^t \in \mathbb{R}^3$. We recover this motion via Procrustes

alignment [32]:

$$\underset{s_c^t \mathbf{R}_c^t, \mathbf{t}_c^t}{\arg\min} \sum_{i \in M_{bg}^t} \mathcal{M}\, \mathcal{V}\, ||\mathbf{p}_i^r - (s_c^t \mathbf{R}_c^t \mathbf{p}_i^{r \leftarrow t} + \mathbf{t}_c^t)||^2, \tag{4}$$

where $\mathcal{M}$ is a binary mask that suppresses background points with unreliable depth estimates (e.g., beyond 50 m), and $\mathcal{V}$ indicates mutual visibility of point i in both frames t and r. The scale term s_c^t is optional: it can be activated only when monocular depth predictions exhibit noticeable temporal scale drift or flicker; otherwise, we fix $s_c^t = 1$ and recover a purely rigid transformation. This flexibility allows PLOT to remain robust across different depth estimators while avoiding unnecessary degrees of freedom when depth is already stable. The estimated camera motion provides a temporally coherent reference frame, which is later used to transform object poses by compensating for ego-motion.

Object Pose and Motion Estimation via Object Trajectories. Similarly to camera motion, we estimate object motion by computing frame-to-frame registration using reliable point correspondences within each object mask M_k^t and its tracked counterpart $M_k^{r \leftarrow t}$:

$$\underset{s^t \mathbf{R}^t, \mathbf{t}^t}{\arg\min} \sum_{i \in M_k^t} \mathcal{V}\, ||\mathbf{p}_i^r - (s^t \mathbf{R}^t \mathbf{p}_i^{r \leftarrow t} + \mathbf{t}^t)||^2, \tag{5}$$

where $\mathbf{R}^t$, $\mathbf{t}^t$ and s^t denote the object's rotation, translation, and scale (optional). The resulting rigid or similarity transformation is later used for temporal aggregation and shape fusion Sec. 3.4.

Using the camera motion from Eq. 4, each object's local trajectory is transformed into the reference frame's coordinate to obtain a world-aligned trajectory:

$$\hat{\mathbf{p}}_i^t = s_c^t \mathbf{R}_c^t \mathbf{p}_i^t + \mathbf{t}_c^t. \tag{6}$$

The motion status of an object is then determined by its displacement between adjacent frames, i.e., $||\hat{\mathbf{p}}_i^s - \hat{\mathbf{p}}_i^t||$. If the displacement is negligible, the object is treated as static; otherwise, it is considered dynamic. For dynamic objects, orientation is inferred from the direction of motion in world space. Since driving scenes (e.g., KITTI [11]) predominantly exhibit motion along the ground plane, we compute the azimuth angle from trajectory changes:

$$\theta_{k,\text{dyna}}^t = \arctan2\left(\frac{\hat{\mathbf{p}}_i^s(z) - \hat{\mathbf{p}}_i^t(z)}{\hat{\mathbf{p}}_i^s(x) - \hat{\mathbf{p}}_i^t(x)}\right). \tag{7}$$

For static objects, we estimate orientation using principal component analysis (PCA) on the fused pseudo-LiDAR from Sec. 3.4, assuming the major axis (highest variance) aligns with the object's longitudinal axis. Unlike prior work [19], our method benefits from denser pseudo-LiDAR from multiple observations, enabling PCA to yield more precise orientation estimates. Per-frame orientation estimates are derived by correcting orientation based on ego-motion as follows:

$$\theta_{k,\text{static}}^t = \Phi^{-1}\Big((\mathbf{R}_c^t)^T \Phi(\theta_k^r)\Big), \tag{8}$$

where Φ denotes the mapping from a yaw angle to a rotation matrix, and Φ^{-1} is its inverse.

3.3 Global Object Memory

Although 2D point trackers [13, 22] and detectors [45] are effective per frame, occlusions and clutter still cause missed detections and visibility gaps, leading to fragmented trajectories. These issues manifest as identity switches and noisy pseudo-LiDARs from incorrect point aggregation. To address this, we analyze common failure modes and introduce a Global Object Memory (GOM) that maintains object hypotheses and enforces consistent associations across frames.

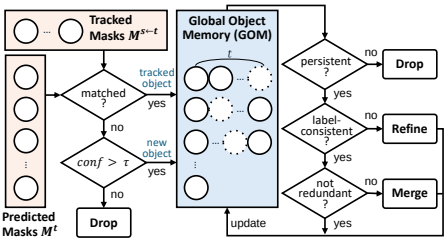


Fig. 4: Global object association pipeline. Our memory-based object association consolidates noisy frame-level predictions into globally consistent tracks.

Failure Modes in Video Object Tracking. To examine the limitations of naive tracking-based video understanding, we analyze two dominant failure modes in real-world videos. First, detection dropouts and label noise arise under occlusion, distance, or challenging conditions, where detectors may drop masks or assign inconsistent labels to the same object over time. Second, identity switches in cluttered scenes occur when multiple objects enter, exit, or overlap within a frame, confusing local matching (Eq. 2). In addition, temporary occlusions often reduce the visibility of tracked points below the tracker's confidence threshold, causing the pipeline to prematurely terminate the trajectory-even though the object remains present-resulting in fragmented tracks. Under these conditions, naive mask-level associations tend to assign new identities to previously observed objects or, conversely, merge distinct objects into a single track. Such failure cases highlight the limits of local heuristics and motivate global, temporally structured association; qualitative examples and further analysis are provided in the supplementary material.

Label Refinement via Global Object Memory. We introduce a Global Object Memory (GOM) that consolidates frame-level predictions into coherent object tracks across frames. As illustrated in Fig. 4, GOM maintains persistent entries, each representing a unique object observed over time. Predicted masks M^t are matched to memory entries based on spatial overlap with the tracked masks $M^{s\leftarrow t}$; if no match is found, a new entry is provisionally created only if the prediction confidence is greater than the detection threshold τ.

GOM updates object entries by checking 1) whether the object persists across frames through point tracking, 2) whether the class labels and geometric attributes remain consistent, and 3) whether multiple memory entries correspond

to the same object. Based on these checks, GOM either 1) discards unstable or incoherent instances, 2) refines entries by updating their attributes with dominant evidence, or 3) merges redundant entries. This refinement enforces long-term identity consistency and improves the reliability of object-level 3D annotations.

In summary, GOM addresses two complementary sources of discontinuity: 1) breaks in temporal continuity caused by missed or duplicated masks from the detector, and 2) fragmented trajectory segments that arise when occlusion reduces point visibility and interrupts tracker propagation.

3.4 Trajectory-guided Object Shape Fusion

Using the globally associated object tracks constructed in Sec. 3.3, we build complete pseudo-LiDARs by aggregating object observations over time. Rather than requiring absolute camera poses, we perform shape fusion using relative poses derived from point trajectories in Sec. 3.2.

For each object, we first identify the reference frame image I_r that minimizes the total registration error between matched frames, using Eq. 5. We then align all matched object masks to this target frame using their estimated relative rotation and translation:

$$\hat{\mathbf{P}}_k = \bigcup_{t \in \mathcal{T}} s^{t \to r} \mathbf{R}^{t \to r} \mathbf{P}_k^t + \mathbf{t}^{t \to r}, \tag{9}$$

where $\mathbf{P}_k^t$ denotes the 3D points extracted from the k-th object mask at time t, and $t \to r$ indicates the transformation from t to the reference frame r. The resulting point cloud $\hat{\mathbf{P}}_k$ serves as a completed pseudo-LiDAR representation, fusing multiple partial views into a consistent shape.

To ensure reliable 3D annotations across all frames, we further project the completed pseudo-LiDAR $\hat{\mathbf{P}}_k$ back into each frame's coordinate system using the inverse of the same relative transformation.

4 Experiments

In this section, we evaluate PLOT on standard monocular 3D object detection (M3OD) video benchmarks—KITTI [11], KITTI-360 [27], and Waymo [49]—and compare it against recent open-set detectors, weakly-supervised and pseudo-labeling methods that rely on driving-specific priors or pose estimates. While our labeler is designed for open-world settings, existing benchmarks with video inputs are constrained to driving scenes, limiting the scope of evaluation. Thus, we provide qualitative comparisons on in-the-wild videos [8, 10, 34] to assess the generalization of PLOT beyond vehicle-centric environments. Finally, we present ablations and runtime comparisons to validate each component. Further details about experimental setup and evaluation metrics are reported in the supplementary material.

Table 1: Detection results for the 'Car' class on the KITTI [11] validation set. 'Attr-Net': an additional attribute estimation network. 'Depth': ground-truth depth map used in training. 'Depth†': KITTI-trained depth completion model. 'VPS': Video, Pose and Shape prior. Both supervised and open-set M3OD models use 3D ground-truth data of KITTI in their training.

	Method	Labels	Extra	AP_{3D}@IoU = 0.3↑			AP_{BEV}@IoU = 0.3↑			AP_{3D}@IoU = 0.5↑			AP_{BEV}@IoU = 0.5↑		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Sup.	DEVIANT [24]	GT-3D	✗	-	-	-	-	-	-	61.00	46.00	40.18	65.28	49.63	43.50
	MonoDETR [62]	GT-3D	✗	79.72	65.87	58.83	80.30	67.16	59.54	68.05	48.42	43.48	72.34	51.97	46.94
Opn.	OVMono3D [56]	GT-3D	✗	74.01	51.25	42.40	76.93	53.85	44.89	51.23	33.09	27.24	55.56	36.47	30.25
	3D-MOOD [55]	GT-3D	Depth	81.97	64.16	54.36	83.04	66.69	56.79	60.74	43.81	36.95	64.93	47.22	40.00
Weak	MonoGRNet [42]	GT-2D	Attr-Net	56.16	42.61	35.36	58.61	48.75	41.49	25.66	21.57	17.40	32.23	26.88	22.47
	WeakM3D [39]	GT-2D	LiDAR	78.44	56.42	45.81	81.17	59.87	48.98	50.16	29.94	23.11	58.20	38.02	30.17
	WeakMono3D [50]	GT-2D	Stereo	-	-	-	-	-	-	49.37	39.01	36.34	54.32	42.83	40.07
	SKD-WM3D [21]	GT-2D	Depth†	-	-	-	-	-	-	50.21	**41.57**	**36.92**	55.47	**44.35**	**41.86**
Pseudo	OVM3D-Det [19]	GSAM	GPT-4	44.48	33.29	26.69	47.42	35.96	27.99	25.63	18.85	15.67	34.52	24.46	20.40
	MonoSOWA [48]	MViT2	VPS	72.70	56.30	47.70	73.38	57.23	48.59	51.55	37.09	33.15	59.76	44.08	36.99
	PLOT (Ours)	GSAM	Video	**80.48**	**60.83**	**51.49**	**83.06**	**63.59**	**54.15**	**52.02**	36.78	30.40	**60.25**	42.80	35.77

Table 2: Detection results for the 'Car' class on the KITTI-360 [27] test set.

	Method	Labels	Extra	AP_{3D}@0.3↑		AP_{BEV}@0.3↑		AP_{3D}@0.5↑		AP_{BEV}@0.5↑	
				Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Sup.	MonoDETR [62]	GT-3D	✗	64.32	57.83	66.42	60.31	47.69	38.76	52.76	43.56
Open.	OVMono3D [56]	Zero-shot	✗	41.82	31.94	49.45	39.08	4.49	3.39	29.18	22.07
	3D-MOOD [55]	Zero-shot	✗	21.15	13.12	32.68	21.41	0.06	0.04	7.62	4.59
Weak.	WeakM3D [39]	GT-2D	LiDAR	21.25	15.34	29.89	24.01	2.96	2.01	8.10	2.96
	Autolabels [57]	GT-2D	LiDAR, Shape	12.92	9.94	48.16	37.34	4.69	2.79	20.18	14.33
Pseudo.	VSRD [29]	GT-2D	Pose	50.86	43.45	58.40	50.61	21.77	16.46	29.07	22.83
	MonoSOWA [48]	MViT2	VPS	42.72	46.59	50.84	49.22	**29.98**	**27.56**	**38.41**	**35.26**
	PLOT (Ours)	GSAM	Video	**54.78**	**48.75**	**60.75**	**54.58**	27.34	21.98	36.88	30.45

4.1 Experiments on Driving Benchmarks

In this subsection, we compare our method with various M3OD works (listed on the left of each table) on multiple benchmark datasets. The methods are grouped as **Sup.** (fully-supervised), **Weak.** (weakly-supervised), **Open.** (open-set), and **Pseudo.** (pseudo-labeling). For pseudo-labeling methods, including ours, results are obtained by training MonoDETR [62] on the generated labels with the same default configuration across all methods to ensure fairness; additional details are provided in the supplementary material. As most M3OD methods focus on the 'Car' and 'Pedestrian' classes, our main paper reports quantitative results primarily for these categories. Additional qualitative results and further analyses are provided in the supplementary material.

KITTI & KITTI-360 (Car). Tab. 1 and Tab. 2 provide quantitative results on the KITTI and KITTI-360 datasets for the 'Car' class. For clarity, the label column lists both 2D ground-truth annotations (GT-2D) and predictions from 2D detectors [26, 45], used in place of ground truth. On KITTI, PLOT

Table 3: Detection results for the 'Pedestrian' class on KITTI [11] validation set.

	Method	Labels	Extra	AP_{3D}@IoU = 0.3↑			AP_{BEV}@IoU = 0.3↑			AP_{3D}@IoU = 0.5↑			AP_{BEV}@IoU = 0.5↑		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Sup.	MonoDIS [47]	GT-3D	✗	-	-	-	-	-	-	3.20	2.28	1.71	4.04	3.19	2.45
	MonoXiver [28]	GT-3D	✗	-	-	-	-	-	-	7.95	5.49	4.62	-	-	-
	GUP-Net [31]	GT-3D	✗	-	-	-	-	-	-	9.37	6.84	5.73	-	-	-
	DEVIANT [24]	GT-3D	✗	-	-	-	-	-	-	9.85	7.18	5.42	-	-	-
W	WeakM3D [39]	GT-2D	LiDAR	-	-	-	3.79	3.21	3.12	-	-	-	-	-	-
Psd.	OVM3D-Det [19]	GSAM	GPT-4	10.63	8.96	7.32	11.25	9.36	7.85	8.19	6.88	5.56	9.11	7.71	6.27
	PLOT (Ours)	GSAM	Video	**16.64**	**14.39**	**12.27**	**17.66**	**15.01**	**12.76**	**13.97**	**11.71**	**9.84**	**14.52**	**12.48**	**10.59**

Table 4: Detection results for the 'Vehicle' class on the Waymo [49] validation set.

	Method	Labels	Extra	AP_{3D}@0.5↑				AP_{BEV}@0.5↑			
				All	0-30m	30-50m	50-∞m	All	0-30m	30-50m	50-∞m
Sup.	DEVIANT [24]	GT-3D	✗	10.29	26.75	4.95	0.16	-	-	-	-
	CaDDN [44]	GT-3D	LiDAR	16.51	44.87	8.99	0.58	-	-	-	-
	MonoDETR [62]	GT-3D	✗	21.41	37.89	18.61	3.69	23.63	39.10	20.95	4.70
Open.	OVMono3D [56]	Zero-shot	✗	5.46	14.68	5.61	0.65	6.11	15.24	6.54	0.99
	3D-MOOD [55]	Zero-shot	✗	1.13	2.44	1.68	0.05	4.60	9.80	6.71	0.30
Weak.	WeakM3D [39]	GT-2D	LiDAR	4.50	12.16	3.67	0.40	8.18	20.31	7.50	0.99
Pseudo.	MonoSOWA [48]	MViT2	VPS	13.46	24.65	5.87	4.55	18.98	33.51	12.02	4.55
	PLOT (Ours)	GSAM	Video	**14.04**	**26.97**	**17.71**	**4.87**	**23.32**	**41.74**	**31.39**	**8.22**

outperforms OVM3D-Det [19] by an average of +19 AP, and also surpasses weakly-supervised methods that rely on auxiliary sensors such as stereo [50] and LiDAR [39]. PLOT also achieves results comparable to fully supervised baselines and open-set detectors. Compared to the previous state-of-the-art method [48], which relies on IMU sensors and simplified shape assumptions, PLOT achieves superior AP@0.3 (up to +9.68) and, as illustrated in Fig. 5, yields more accurate 3D boxes by leveraging trajectory-guided shape fusion and global object association. On KITTI-360, PLOT also improves over recent pseudo-labeling methods [29, 48], achieving the best AP@0.3 in both Easy and Hard regimes.

KITTI (Pedestrian). Beyond the commonly evaluated 'Car' category, we further assess performance on the more challenging 'Pedestrian' class in the KITTI validation split (Tab. 3). Pedestrians introduce additional challenges due to their small size, frequent occlusions, and non-rigid motion, which often violate shape assumptions adopted by existing methods [19, 48]. Despite these difficulties—and although pedestrian evaluation is less commonly reported in prior pseudo-labeling works—PLOT achieves strong performance and surpasses existing approaches. This result indicates that trajectory-guided shape fusion remains effective even when rigid shape assumptions break down, enabling robust localization and orientation estimation for small and dynamic objects.

Fig. 5: Qualitative comparisons on KITTI. Only the predicted boxes are displayed for clarity.

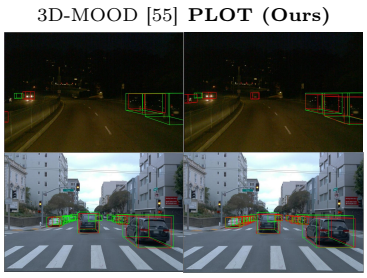


Fig. 6: Qualitative comparison on Waymo-Open.

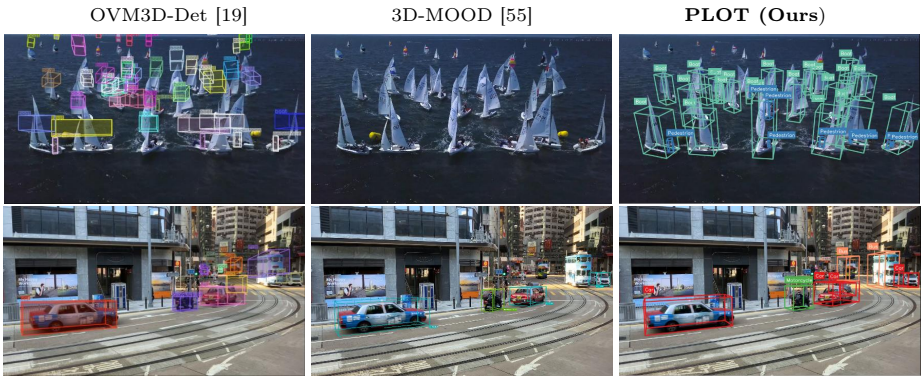


Fig. 7: Qualitative comparisons in the wild.

Waymo-Open (Vehicle). We report results on the Waymo-Open dataset in Tab. 4. Unlike KITTI benchmarks, Waymo features long, diverse sequences under adverse conditions such as night and rain, making it a strong proxy for real-world complexity. PLOT achieves the best performance in distant regions (>30m), surpassing fully supervised baselines, and shows superior AP_{BEV} across nearly all ranges. As shown in Fig. 6, PLOT remains effective under challenging conditions such as night scenes, even recovering valid instances missing from the ground truth (red: predictions, green: ground truth), underscoring the scalability and reliability of our framework in challenging environments.

4.2 Experiments on In-the-wild Videos

To demonstrate the open-set capability of our method, we present qualitative comparisons on in-the-wild videos using identical text prompts (e.g., "boat") as inputs for PLOT, a state-of-the-art open-set method, 3D-MOOD [55], and a single-image pseudo-labeling method OVM3D-Det [19]; MonoSOWA [48], which depends on auxiliary sensors and object shape priors, is not applicable here. As shown in Fig. 1 and Fig. 7, PLOT generalizes well under diverse and challenging

Table 5: Ablation on trajectory-guided shape fusion and number of adj. frames.

#Fr.	AP_{3D}@IoU = 0.3↑ Easy	Mod.	Hard	AP_{BEV}@IoU = 0.3↑ Easy	Mod.	Hard	Time (10/20 obj.)
1	51.75	37.48	30.63	57.41	41.96	34.00	**0.46** / **0.47**
+2	62.41	48.06	40.31	67.57	51.39	43.35	0.66 / 0.70
+8	77.23	56.16	48.43	80.27	59.13	51.29	1.56 / 1.73
+20	**80.48**	**60.83**	**51.49**	**83.06**	**63.59**	**54.15**	2.45 / 2.83

Table 6: Ablation on global object memory and PLOT as standalone detector.

	Train	GOM	AP_{3D}@IoU = 0.3↑ Easy	Mod.	Hard	AP_{BEV}@IoU = 0.3↑ Easy	Mod.	Hard
KIT.	✗	✓	59.82	51.00	45.40	62.41	54.18	48.44
	✓	✓	**80.48**	**60.83**	**51.49**	**83.06**	**63.59**	**54.15**
K360	✓	✗	48.14	-	40.28	54.86	-	46.61
	✓	✓	**54.78**	-	**48.75**	**60.75**	-	**54.58**

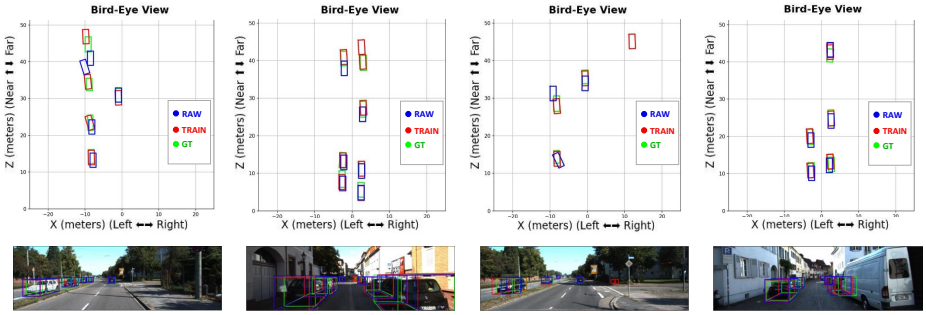


Fig. 8: BEV comparison between the detection results of raw pseudo-labels and the detector trained with it, on KITTI [11].

conditions (e.g., variations in scene layouts, camera setups) while maintaining accurate object attribute estimates. In contrast, 3D-MOOD does not detect any objects in the 'boat' scene (top) in Fig. 7, and OVM3D-Det often produces noisy estimates with most falls back to priors of LLM-driven fixed objects, due to single image observations. Such in-the-wild evaluations are particularly important, as they go beyond structured driving benchmarks and validate the scalability of our method to unconstrained real-world settings. Additional qualitative results along with video demonstrations are provided in the supplementary material.

4.3 Ablation Studies

Efficacy of Trajectory-guided Shape Fusion. Tab. 5 compares trajectory-guided shape fusion with varying frame counts, reporting detection performance and labeling time per frame on KITTI. Performance improves with more frames, achieving up to 1.7× higher average AP compared to single-frame labels. The rightmost column of the table shows the computation time of the whole pipeline, which grows proportionally with the number of objects, yet remains practical (running at 2.83s per frame on a single CPU with a RTX 3090 GPU) with 20 objects per scene, while the KITTI dataset on average contains about 10–12 objects per frame. Considering both accuracy and computation time, we adopt a 20-frame window for the experiments.

Table 7: Quantitative analysis of the pseudo-label accuracy for the 'Car' class on the KITTI [11] validation set.

Method	Object Type	AP_{2D}↑ (%)			AOE↓ (rad)			ASE↓ (1-IoU)		
		Near	Mid.	Far	Near	Mid.	Far	Near	Mid.	Far
OVM3D-Det [19]	static	64.43	49.87	40.92	1.245	1.200	1.380	0.338	0.312	0.412
PLOT (Ours)	static	**80.40**	**73.03**	**64.18**	**0.716**	**0.695**	**0.794**	**0.149**	**0.140**	**0.249**
OVM3D-Det [19]	All	65.57	59.68	48.27	1.360	1.127	1.413	0.329	0.297	0.318
PLOT (Ours)	All	**83.98**	**79.40**	**71.53**	**0.536**	**0.621**	**0.712**	**0.136**	**0.131**	**0.168**

PLOT as Standalone 3D Detector. In the top 2 rows of Tab. 6, we examine the feasibility of using PLOT as a standalone 3D detector without training an additional M3OD model, MonoDETR [62] on KITTI. Remarkably, the raw pseudo-labels already surpass the previous pseudo-labeling method (OVM3D-Det [19]) and the weakly-supervised method (MonoGRNet [42]) (see Tab. 1), suggesting the utility of PLOT in scenarios where training is infeasible, such as in-the-wild videos or cases with only a single sequence available. However, the integration of uncertainty-aware depth estimation and 2D–3D geometric consistency in MonoDETR mitigates depth errors and label inconsistencies, yielding more stable and accurate 3D detections. Fig. 8 provides qualitative evidence supporting this observation by comparing ground truth, raw pseudo-labels, and the predictions obtained after training MonoDETR on the generated labels in BEV space. While the raw pseudo-labels already exhibit reasonable localization quality, depth noise occasionally shifts object centers (noticeable in far objects); training reduces these deviations through regularization, resulting in more consistent geometric alignment.

Necessity of Global Object Memory. We evaluated the impact of global object memory on KITTI-360 (K360) in the bottom 2 rows of Tab. 6, as KITTI's 3D detection benchmark (single-frame) is not suitable for assessing GOM, whereas KITTI-360 provides continuous sequences. The capability of GOM to reduce 2D detector-induced noise and maintain temporal consistency results in more stable and accurate pseudo-labels. Such video-consistent label generation is directly relevant for practical annotation pipelines and downstream 3D perception tasks. In the supplementary material, we also provide video-consistent label results on the Waymo dataset as well as the breakdown of temporal consistent 3D labels on in-the-wild scene in the supplementary material.

Analysis of Pseudo-Labels Accuracy. Tab. 7 compares raw pseudo-labels from PLOT and OVM3D-Det [19] using nuScenes [6] metrics without any training. PLOT yields a higher detection rate due to its label improvement scheme, which minimizes missing and duplicate detections. The AOE (Average Orientation Error) difference shows that PLOT improves over OVM3D-Det (which also uses PCA) on both all and static objects only, indicating that our pseudo-LiDAR

completion provides more reliable geometry, even when only limited visible surfaces are available, as also qualitatively shown for parked cars in Fig. 6 of the supplementary material. The improved geometry also benefits the object size estimation, evidenced by a significant drop in ASE (Average Scale Error).

5 Conclusion

In this paper, we introduced PLOT, a framework for generating reliable 3D annotations from monocular videos through tracking-driven object association and label refinement. Beyond monocular 3D detection, this video-based formulation holds potential for broader 3D tasks that suffer from missing camera information or incomplete shapes, such as CAD model retrieval and object-level reconstruction. We hope that this paradigm provides a scalable alternative to data-intensive monocular 3D detection pipelines and opens new directions for video-based 3D understanding without explicit supervision.

Limitations. Like most open-set 3D perception frameworks, PLOT relies on pre-trained depth estimators. While this dependency is not unique to our approach, depth errors under noise or at long ranges remain a key bottleneck. In practice, however, training downstream M3OD models on the generated pseudo-labels helps reduce these errors by regularizing noisy depth estimates. In addition, orientation for static objects remains inherently ambiguous under partial-view observations. While we adopt PCA-based estimation, the trajectory-guided shape completion in PLOT provides more complete object geometry over time, making such estimators substantially more robust in practice. Further analysis and additional failure cases are provided in the supplementary materials.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(RS-2023-00227592, Development of 3D Object Identification Technology Robust to Viewpoint Changes), the Korea Institute of Police Technology (KIPoT) funded by the Korean National Police Agency & Ministry of the Interior and Safety (RS-2024-00405100), and the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 62E0062). This work was also partly funded by the Czech Science Foundation (GACR) JUNIOR STAR Grant No. 22-23183M (supporting S.Aung).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

2. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021)
3. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13154–13164 (2023)
4. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9287–9296 (2019)
5. Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. pp. 135–152. Springer (2020)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
7. Cho, S., Huang, J., Kim, S., Lee, J.Y.: Seurat: From moving points to depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2025)
8. Contributors, P.: Pexels - free stock photos and videos (2025), `https://www.pexels.com/`, accessed: 2025-03-04
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
10. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20224–20234 (2023)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
12. Han, W., Tao, R., Ling, H., Shen, J.: Weakly supervised monocular 3d object detection by spatial-temporal view consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
13. Harley, A.W., You, Y., Sun, X., Zheng, Y., Raghuraman, N., Gu, Y., Liang, S., Chu, W.H., Dave, A., Tokmakov, P., You, S., Ambrus, R., Fragkiadaki, K., Guibas, L.J.: AllTracker: Efficient dense point tracking at high resolution. In: ICCV (2025)
14. Hu, J.S.K., Kuai, T., Waslander, S.L.: Point density-aware voxels for lidar 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8469–8478 (June 2022)
15. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. In: arXiv preprint (2024)
16. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. In: NVIDIA Research Whitepapers (2025)
17. Huang, K.C., Tsai, Y.H., Yang, M.H.: Weakly supervised 3d object detection via multi-level visual guidance. In: ECCV (2024)

18. Huang, K.C., Wu, T.H., Su, H.T., Hsu, W.H.: Monodtr: Monocular 3d object detection with depth-aware transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4012–4021 (2022)
19. Huang, R., Zheng, H., Wang, Y., Xia, Z., Pavone, M., Huang, G.: Training an open-vocabulary monocular 3d detection model without 3d data. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
20. Jia, J., Li, Z., Shi, Y.: Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
21. Jiang, X., Jin, S., Lu, L., Zhang, X., Lu, S.: Weakly supervised monocular 3d detection with a single-view image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10508–10518 (2024)
22. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: European Conference on Computer Vision. pp. 18–35. Springer (2025)
23. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
24. Kumar, A., Brazil, G., Corona, E., Parchami, A., Liu, X.: Deviant: Depth equivariant network for monocular 3d object detection. In: European Conference on Computer Vision. pp. 664–683. Springer (2022)
25. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024)
26. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4804–4814 (June 2022)
27. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. Pattern Analysis and Machine Intelligence (PAMI) (2022)
28. Liu, X., Zheng, C., Cheng, K.B., Xue, N., Qi, G.J., Wu, T.: Monocular 3d object detection with bounding box denoising in 3d by perceiver. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6436–6446 (2023)
29. Liu, Z., Sakuma, H., Okutomi, M.: Vsrd: Instance-aware volumetric silhouette rendering for weakly supervised 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17354–17363 (2024)
30. Liu, Z., Zhou, D., Lu, F., Fang, J., Zhang, L.: Autoshape: Real-time shape-aware monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15641–15650 (2021)
31. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3111–3121 (2021)
32. Luo, B., Hancock, E.R.: Procrustes alignment with the em algorithm. In: International Conference on Computer Analysis of Images and Patterns. pp. 623–631. Springer (1999)
33. Maninis, K.K., Popov, S., Nießner, M., Ferrari, V.: Vid2cad: Cad model alignment using multi-view constraints from videos. IEEE transactions on pattern analysis and machine intelligence **45**(1), 1320–1327 (2022)
34. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (Mar 2016), `http://arxiv.org/abs/1603.00831`, arXiv: 1603.00831

35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
36. Pan, T.Y., Ma, C., Chen, T., Phoo, C.P., Luo, K.Z., You, Y., Campbell, M., Weinberger, K.Q., Hariharan, B., Chao, W.L.: Pre-training liDAR-based 3d object detectors through colorization. In: The Twelfth International Conference on Learning Representations (2024)
37. Parihar, R., Sarkar, S., Vora, S., Kundu, J.N., Babu, R.V.: Monoplace3d: Learning 3d-aware object placement for 3d monocular detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6531–6541 (June 2025)
38. Peng, L., Wu, X., Yang, Z., Liu, H., Cai, D.: Did-m3d: Decoupling instance depth for monocular 3d object detection. In: European Conference on Computer Vision. pp. 71–88. Springer (2022)
39. Peng, L., Yan, S., Wu, B., Yang, Z., He, X., Cai, D.: Weakm3d: Towards weakly supervised monocular 3d object detection. arXiv preprint arXiv:2203.08332 (2022)
40. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: CVPR. pp. 10106–10116 (2024)
41. Qin, Z., Wang, J., Lu, Y.: Weakly supervised 3d object detection from point clouds. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4144–4152 (2020)
42. Qin, Z., Wang, J., Lu, Y.: Monogrnet: A general framework for monocular 3d object detection. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5170–5184 (2021)
43. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
44. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8555–8564 (2021)
45. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
46. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019)
47. Simonelli, A., Bulo, S.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1991–1999 (2019)
48. Skvrna, J., Neumann, L.: Monosowa: Scalable monocular 3d object detector without human annotations. arXiv preprint arXiv:2501.09481 (2025)
49. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: 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)
50. Tao, R., Han, W., Qiu, Z., Xu, C.Z., Shen, J.: Weakly supervised monocular 3d object detection using multi-view projection and direction consistency. In: Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17482–17492 (2023)
51. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021)
52. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
53. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
54. Wang, T., Pang, J., Lin, D.: Monocular 3d object detection with depth from motion. In: European Conference on Computer Vision. pp. 386–403. Springer (2022)
55. Yang, Y.H., Piccinelli, L., Segu, M., Li, S., Huang, R., Fu, Y., Pollefeys, M., Blum, H., Bauer, Z.: 3d-mood: Lifting 2d to 3d for monocular open-set object detection. arXiv preprint arXiv:2507.23567 (2025)
56. Yao, J., Gu, H., Chen, X., Wang, J., Cheng, Z.: Open vocabulary monocular 3d object detection. arXiv preprint arXiv:2411.16833 (2024)
57. Zakharov, S., Kehl, W., Bhargava, A., Gaidon, A.: Autolabeling 3d objects with differentiable rendering of sdf shape priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12224–12233 (2020)
58. Zhang, G., Fan, J., Chen, L., Zhang, Z., Lei, Z., Zhang, L.: General geometry-aware weakly supervised 3d object detection. In: European Conference on Computer Vision. pp. 290–309. Springer (2024)
59. Zhang, H., Jiang, H., Yao, Q., Sun, Y., Zhang, R., Zhao, H., Li, H., Zhu, H., Yang, Z.: Detect anything 3d in the wild. arXiv preprint arXiv:2504.07958 (2025)
60. Zhang, J., Li, J., Lin, X., Zhang, W., Tan, X., Han, J., Ding, E., Wang, J., Li, G.: Decoupled pseudo-labeling for semi-supervised monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16923–16932 (June 2024)
61. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. In: International Conference on Learning Representations (2025)
62. Zhang, R., Qiu, H., Wang, T., Guo, Z., Cui, Z., Qiao, Y., Li, H., Gao, P.: Monodetr: Depth-guided transformer for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9155–9166 (2023)