UniLiPs: Unified LiDAR Pseudo-Labeling with Geometry-Grounded Dynamic Scene Decomposition

Filippo Ghilotti[1] Samuel Brucker[1] Nahku Saidy[1]
Matteo Matteucci[2] Mario Bijelic[1,3] Felix Heide[1,3]

[1]TORC Robotics [2]Politecnico of Milan [3]Princeton University

https://light.princeton.edu/unilips

Abstract

Unlabeled LiDAR logs, in autonomous driving applications, are inherently a gold mine of dense 3D geometry hiding in plain sight - yet they are almost useless without human labels, highlighting a dominant cost barrier for autonomous-perception research. In this work we tackle this bottleneck by leveraging temporal–geometric consistency across LiDAR sweeps to lift and fuse cues from text and 2D vision foundation models directly into 3D, without any manual input. We introduce an unsupervised multi-modal pseudo-labeling method relying on strong geometric priors learned from temporally accumulated LiDAR maps, alongside with a novel iterative update rule that enforces joint geometric-semantic consistency, and vice-versa detecting moving objects from inconsistencies. Our method simultaneously produces 3D semantic labels, 3D bounding boxes, and dense LiDAR scans, demonstrating robust generalization across three datasets. We experimentally validate that our method compares favorably to existing semantic segmentation and object detection pseudo-labeling methods, which often require additional manual supervision. We confirm that even a small fraction of our geometrically consistent, densified LiDAR improves depth prediction by 51.5% and 22.0% MAE in the 80–150 and 150–250 meters range, respectively.

1. Introduction

Large-scale annotated datasets and increased computing power have enabled the succes of learned vision methods. Datasets like ImageNet[16], PASCAL VOC[19], MSCOCO[36], Cityscapes[14], and ADE20K[67] have driven advances in classification, detection, and segmentation. In autonomous driving, annotating large-scale data, especially $3D$ LiDAR scans, is challenging and costly due to the need for precise multi-modal alignment. Multi-modal benchmarks such as KITTI[21], nuScenes[8], the Waymo Open Dataset[57], and Argoverse[11] reflect this effort. To tackle the annotation challenges for large datasets, a body of work explores automatic labeling, using pre-aligned $3D$ models to incorporate geometric and semantic constraints into the annotation pipeline, effectively reducing ambiguity and enhancing consistency across labels [10, 33, 39, 61, 62, 65]. Methods relying on synthetic data can generate fully annotated video sequences, providing detailed $2D$ and $3D$ multi-object tracking information, along with pixel-level labels for categories, instances, flow, and depth [7, 18, 51]. Other efforts aim to minimize annotation workload through offline perception [25], and semi-supervised approaches [2, 9, 23, 24, 42, 43, 52, 58, 66] leverage unlabeled data, although depending on specialized architectures to handle partial ground-truth labels. Specifically, we note that these existing annotation methods typically require separate methods for each task — be it depth

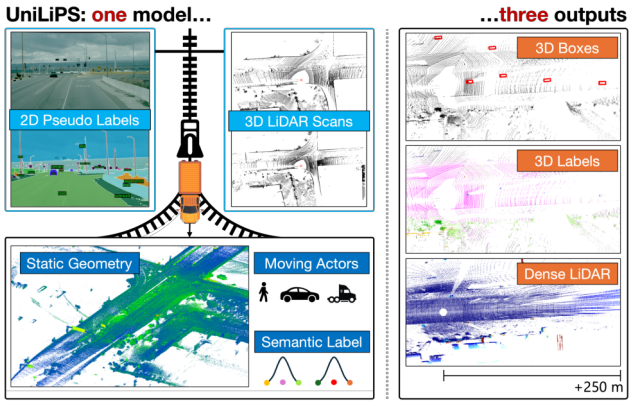

Figure 1. **Unified 3D Labeling.** Given a single driving trajectory, UniLiPs fuse consecutive LiDAR scans with our engine's 2D pseudo-labels to build a coherent 3D map. Within this consistent geometry, moving actors and semantic labels are optimized to *jointly* generate refined, temporally consistent 3D bounding boxes, semantic labels, and occlusion-aware, densified point clouds.

estimation, object detection, or semantic segmentation — and often rely on manually generated or pseudo labels that are hard to reproduce. In contrast, we rethink the annotation process in an unsupervised, *unified* $3D$ labeling framework, as presented in Figure 1, that concurrently tackles all these tasks by leveraging a consistent SLAM-based $3D$ map as a comprehensive semantic-geometric representation, ensuring frame invariance and enhanced reproducibility across modalities, to generate labels for tasks such as $3D$ bounding box detection, semantic segmentation, and depth estimation, with minimal parameters tuning. Our approach enriches a $3D$ map with semantic, geometric, and probabilistic information, and exploits sensor fusion and geometric consistency to automatically separate the static scene from dynamic objects. We cast the problem as a novel *Iterative Update Weighted Function* to distinguish moving objects — which break the static world assumption and are refined into trajectory-aware bounding boxes — from static regions, which are then converted into densified LiDAR-like scans via *Adaptive Spherical Occlusion Culling* and enhanced with rich semantic details. We evaluate the method against held-out manual labels and training state-of-the-art networks with our pseudo-labels, on semantic segmentation, object detection and depth estimation.

We make the following contributions:

- We introduce a novel method to obtain *jointly* pseudo-labels for 3 different downstream tasks (semantic segmentation, object detection and depth estimation), at scale, with no manual annotations needed and not tied to any specific dataset or sensor suite.
- We devise a method for static and dynamic object separation, exploiting points and labels temporal accumulation and an *Iterative Update Weighted Function*.
- We find that our semantic labels and bounding boxes achieve *SOTA* performances compared to standalone pseudo labeling methods and confirm they can grant *close to Oracle* performance on three different datasets.
- For depth estimation, we devise how a lightweight fine-tuning on a subset of our consistent pseudo ground-truth achieves *improvements of 51.5% in MAE between 80 and 150 meters and 22.0% between 150 and 250 meters.*

2. Related Work

Pseudo Depth. High-density LiDAR depth maps are traditionally produced using LiDAR Inertial Odometry algorithms [1, 13, 17, 55, 56, 59] that aggregate information across multiple frames. Conversely image-based depth foundation models [4–6, 34, 45, 46, 48, 49, 63, 64] have demonstrated significant potential for generating dense depth predictions from single images but still lack behind when delivering metric depth accuracy.

Pseudo Segmentation for LiDAR data. Recent advances in LiDAR pseudo-labeling leverage motion and appearance cues to generate robust labels, such as unsupervised instance segmentation methods to exploits these cues [53] and methods extending 2D vision proposals into 3D space using grouping and voxelization techniques [20, 44]. Additionally flow estimation through motion segmentation [37] can achieve real-time accuracy, but faces challenges with pose estimation in longer sequences.

Outputs PL	Det. From Motion [2, 37, 54]	Pseudo Seg. [20, 28]	Depth Pred. [59, 64]	**Ours**
Bounding Boxes	✓	✗	✗	✓
Semantic Labels	✗	✓	✗	✓
Dense Depth	✗	✗	✓	✓
Moving Objects	✓	(✓)	✗	✓
Requirements and Specifications				
Long-Range	✗	(✓)	✓	✓
Dataset Invariant	(✓)	✗	✗	✓
Time Consistent	(✓)	(✓)	✗	✓
Unsupervised	(✓)	(✓)	✗	✓

Table 1. **Unified Labeling.** Our approach jointly generates (✓) all Pseudo Labels (PL) types, at long range, without any ground-truth supervision. By contrast, state-of-the-art methods often rely on ground-truth data, only partially ((✓)) satisfy consistency and invariance requirements and not deliver (✗) all the outputs.

3D Pseudo Bounding Boxes for LiDAR Data. Pseudo labeling has emerged as a pivotal technique in LiDAR object detection, addressing the reliance on extensive labeled datasets by generating pseudo labels for point clouds. 3DIoUMatch [60] employed a semi-supervised framework to filter high-quality pseudo labels object detection. More recently, [54] leveraged motion cues to group coherently moving points into objects, though tracking across numerous frames remains computationally demanding. Similarly, [2] exploited self-supervised flow estimation and trajectory consistency to mine 3D bounding boxes.

Proposed Unified Labeling. While addressing the same tasks tackled in isolation in prior work, we introduce a unified 3D labeling framework to concurrently deliver consistent depth estimation, object detection, and semantic segmentation pseudo labels, at longer range and without any form of supervision, as detailed in Table 1. Despite handling three tasks together, our method still matches and surpasses dedicated models on each task.

3. Geometry-Grounded Pseudo-Labeling

We introduce a pseudo-labeling method for LiDAR point clouds, agnostic to datasets and sensor setups, combin-

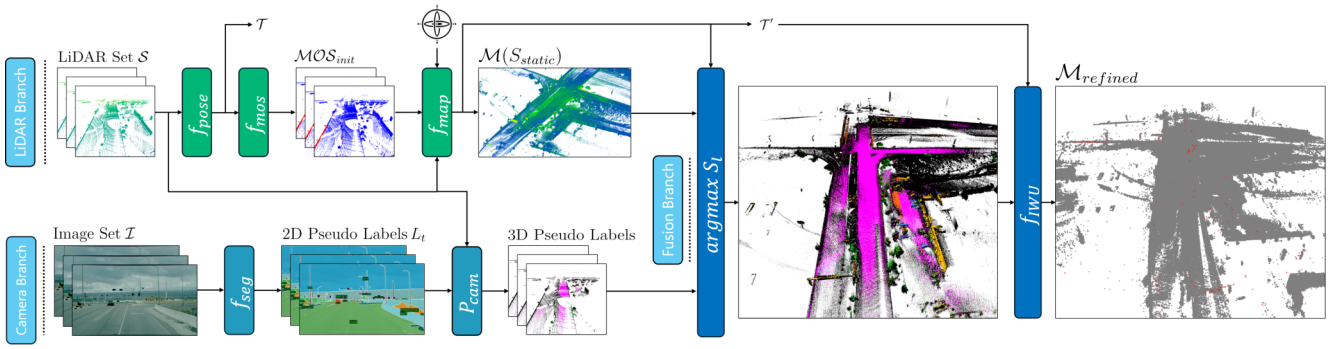

Figure 2. **Overview of Geometry-Grounded Dynamic Scene Decomposition.** Starting from a set of raw images, a set of LiDAR scan and IMU measurements, we first produce 3D semantic labels. Therefore, the 2D semantic masks produced by f_{seg} are integrated into a map generated by the SLAM method f_{map}, by projecting them through P_{cam}, while simultaneously removing moving points identified by f_{mos} from the map. To obtain a refined static scene map $\mathcal{M}_{refined}$ we first propagate the labels through geometric and temporal constraints and later on exploit them to remove remaining floaters and outliers (in red) through a our *Iterative Weighted Update Function* f_{IWU}.

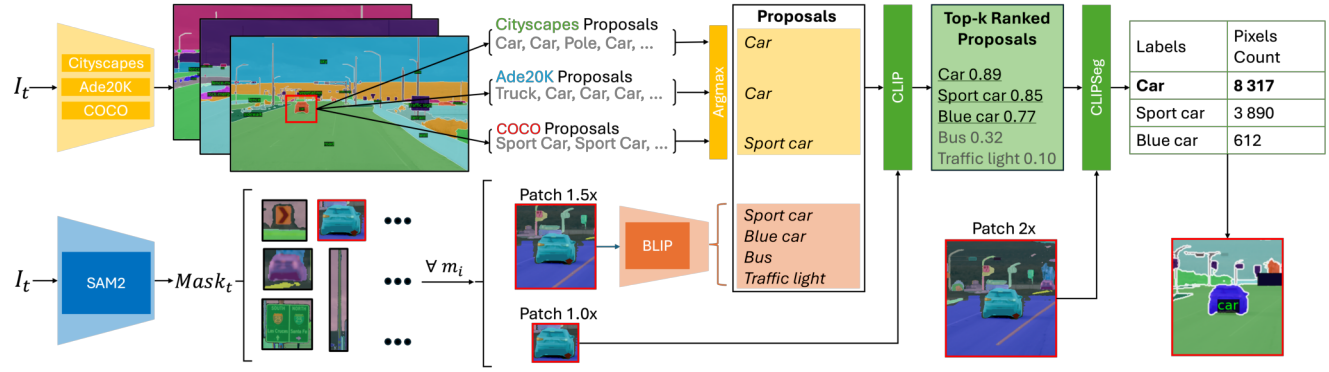

Figure 3. **Overview of Pseudo Labeling Function.** Our proposed pseudo labeling method f_{seg} robustly segments each $2D$ image I_t by combining the predictions from an ensamble of three OneFormer [27] models with weights from three different datasets (COCO [36], ADE20K [68, 69] and Cityscapes [15]) and a SAM2 [50] instance prediction set $Mask_t$. For each mask m_i, BLIP [31] enriches the class proposals and through modal alignment with CLIP [47] and CLIPSeg [38], we ensure high quality domain-specific annotations.

ing foundational vision models with geometry-aware probabilistic constraints on an accumulated scene map. As illustrated in Figure 2, images, LiDAR point clouds and IMU data are fused to initialize low-confidence semantic and motion labels. Iterating over the map we update labels probabilities, refining static structures and removing outliers through our *Iterative Weighted Update*, yielding a dense $3D$ point map with reliable semantic labels and a sparse set of high-confidence moving objects.

3.1. LiDAR Branch Processing

We process a set of LiDAR scans $\mathcal{S} = \{s_t \mid t = 1, \dots, N\}$ and IMU to obtain an accumulated map $\mathcal{M}(S_{static})$ free of the identified, low-confidence moving objects in MOS_{init}.

Pose Estimation and Motion Cues. We first apply a LiDAR-odometry method f_{pose} to estimate frame transformations $\mathcal{T} = T_t \in \text{SE}(3)$, that a motion-segmentation network f_{mos} exploits to identify an initial set of low-confidence moving objects MOS_{init}.

SLAM Mapping. All points marked static are fused by a LiDAR-inertial SLAM module f_{map} to obtain the initial map $(\mathcal{M}, \mathcal{T}') = f_{\text{map}}(\mathcal{S}_{\text{static}}, IMU)$. Here $\mathcal{S}_{\text{static}} = \{s_t^{\text{static}} \mid s_t^{\text{static}} = s_t \setminus \text{mos}_t,\ s_t \in \mathcal{S},\ \text{mos}_t \in \text{MOS}_{\text{init}}\}$. This early pruning of moving objects mitigates ghost artifacts and allows for better geometric optimization in the SLAM.

3.2. Image Branch Processing

We process a set of images $\mathcal{I} = \{I_t \mid t = 1, \dots, M\}$, to generate and lift $2D$ semantic pseudo labels in $3D$.

Our pseudo labeling function is presented in details in Figure 3: given the set $\mathcal{I}$ of RGB images, for each $I_t \in \mathbb{R}^{H \times W \times 3}$ it predicts a per-pixel label image $L_t = f_{\text{seg}}(I_t) \in \mathcal{L}^{H \times W}$. Each frame is down-sampled and processed with SAM2 [50], producing a set of object masks $Mask_t$. The input image is segmented by three separate OneFormer [27] models, individually trained on COCO [36], ADE20K [68, 69], and Cityscapes [15]. For each $m_i \in Mask_t$, the initial proposal list is generated by stacking the most recurrent ($Argmax$) class from each OneFormer model per-pixel logit maps. We then ex-

tract three image crops centered on the mask's bounding box: original size (1.0×), large (1.5×), and huge (2.0×). Subsequently, open-vocabulary classification is performed by non-prompted BLIP [31] on the large patch, proposing class candidates that augment the OneFormer proposal list. CLIP [47] re-ranks the candidates list on a tighter crop, producing a shortlist of the top-k keywords: we select specifically $k = 3$. CLIPSeg [38] processes the full-resolution crop together with this shortlist and outputs per-pixel scores. A majority vote assigns the final class to all pixels of m_i, reducing boundary noise. If multiple classes remain, we keep the one with the highest pixel count. Iterating through every $m_i \in Mask_t$ we obtain a refined label map L_t which serves as initial guess.

Occlusion Aware Semantic Lifting. Each LiDAR pointcloud s_t is projected into the correspondent label map $L_t(u, v)$ with calibration matrix P_{cam}. Let $D(u, v)$ be the depth at pixel (u, v), the visibility mask $M(u, v)$ is determined by comparing the depth with the minimum depth in its neighborhood $\mathcal{N}(u, v)$. A point is marked as visible if

$$D(u, v) \leq \min(D(\mathcal{N}(u, v))) + 0.5. \quad (1)$$

Only visible points inherit the semantic label $l_t(u, v)$ of $L_t(u, v)$, ensuring noisy labels are reduced in sparser region or in common penetration cases.

3.3. Geometry-Consistent Fusion Branch

After differentiating the world into static world ($\mathcal{M}$) and dynamic objects (MOS_{init}) we propose a geometry-grounded method to iteratively refine the static world representation.

Semantic Multimodal Propagation. By sequentially associating each point label of a LiDAR scan to the correspondent point in the map, we project the semantics from each camera into the world map. As a result, each point in the map then is be represented as $p_i = (x_i, y_i, z_i, \{(l_{i1}, n_{i1}), (l_{i2}, n_{i2}), \ldots, (l_{im_i}, n_{im_i})\})$, where x_i, y_i, z_i are the spatial coordinates and $\{(l_{ij}, n_{ij})\}$ is a set of label-count pairs associated with point i, and

- l_{ij} is the j-th label assigned to point p_i.
- n_{ij} is the number of times label l_{ij} was assigned p_i.

Here, m_i is the total number of unique labels assigned to point i. We propagate labels probabilistically in order to enhance segmented areas and fill gaps in the map following Algorithm 1. Here, $w_{ij} = \exp\big(-\|p_i - p_j\|^2/(2\sigma^2)\big)$, and $\delta(l_j = l)$ equal to 1 if $l_j = l$, and 0 otherwise.

Map Refinement. To refine the map from remaining floaters we propose our *Iterative Weighted Update Function* f_{IWU}: by iteratively comparing the sparse LiDAR with the map, points belonging to moving objects but mistakenly registered in the map are likely to be observed only once or twice by subsequent scans. Consequently, we update the probability that each map point is static by considering the frequency of its observations, incorporating a distance based influence factor. For each point $p_j \in s_t, s_t \in \mathcal{S}, t = 1, \ldots, N$, we calculate the Euclidean distance d_{ij} to all map points $m_i \in \mathcal{M}$ and from the sensor origin r_j, $dij = \|p_j - m_i\|, \quad r_j = \|p_j\|$. We locate the nearest map point $\tilde{m}_i$ for each scan point p_j and compute

$$r^*{}_j = \min\left(1, \frac{r_{\max}}{r_j}\right), \quad C(\tilde{m}_i) = \frac{\max_j n_{ij}}{\sum_j n_{ij}}, \quad (2)$$

with n_{ij} counting how often $\tilde{m}_i$ was labeled as class $l \in \{\text{movable}, \text{non-movable}\}$ and r_{max} defining a full-credibility radius of 200 meters. The static probability update rule, if $\tilde{m}_i$ is found in a 30 centimeters radius, is:

$$P^t(\tilde{m}_i) = \alpha \cdot P^{t-1}(\tilde{m}_i) + (1 - \alpha) \cdot r_j^* \cdot \big(1 + C(\tilde{m}_i)\big), \quad (3)$$

otherwise

$$P^t(\tilde{m}_i) = \alpha \cdot P^{t-1}(\tilde{m}_i) + (1 - \alpha) \cdot (1 - r_j^*) \cdot \big(1 - C(\tilde{m}_i)\big). \quad (4)$$

Points with probabilities exceeding a predefined threshold τ_s are classified as static, while those below τ_s are marked as moving and discarded from the map.

Algorithm 1 Probabilistic Label Propagation

Require: Point cloud $\{(p_i, l_i)\}_{i=1}^N$, neighborhood radius r
Ensure: Refined labels $\{l_i\}_{i=1}^N$

```
Build a KD-Tree from points {p_i}
Set σ = r/2
for all points p_i do
    N_i = {(p_j, l_j) | ||p_i − p_j|| ≤ r, l_j ∉ {0, −1}}
    if N_i ≠ ∅ then
        S_l = Σ_{(p_j,l_j)∈N_i} w_ij · δ(l_j = l)
        l_i = arg max_l S_l
    end if
end for
```

3.4. Pseudo Label Outputs

After the refinement stage, our pipeline can generate different pseudo ground-truth supervision signals like densified LiDAR scans, 360° semantic labels and $3D$ bounding boxes from moving-object segmentation masks, as in Figure 4.

3D Semantics. We extract semantic labels for each LiDAR scan from the semantically propagated map, preserving the initial guess l_{ij} for points without correspondence.

Moving Objects. We detect moving objects by aligning each LiDAR scans in $\mathcal{S}$ to the refined consistent static map $\mathcal{M}_{\text{refined}}$, and segment as moving those points without correspondence in the map, requiring the existence of at least 2 other moving candidates in a neighborhood of 1 meter.

3D Bounding Boxes. To transform the moving object detections into bounding boxes, we first exploit the pose $\mathcal{T}'$ to

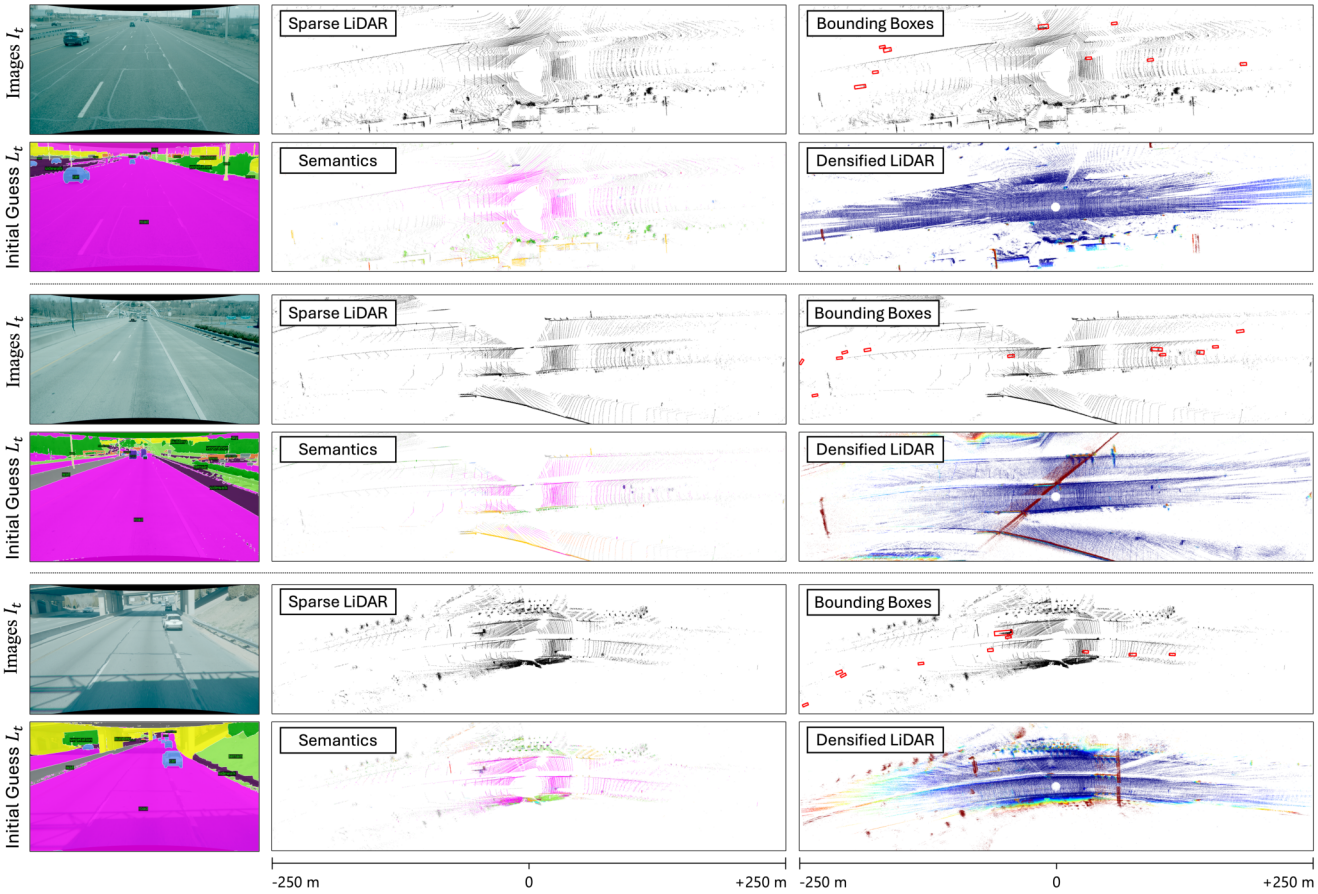

Figure 4. **UniLiPs Unified Labeling Outputs.** Coupling geometric point-cloud aggregation with image segmentation cues from our f_{seg}, UniLiPs rivals standalone methods by jointly producing temporally consistent semantic labels, trajectory-smoothed bounding boxes, and densified LiDAR sweeps that are denser and offer finer angular resolution, especially at long range. In the Figure, densified LiDARs are z-colored between -2 (blue) and $+5$ (red) meters, while semantics are class-coloured based on SemanticKITTI mapping.

align three consecutive scans, considering only the points labeled as moving, and then use HDBSCAN [41] to cluster them. We fit the minimum enclosing cuboid to each cluster and assign a minimum size; we use PCA to get an initial estimate of the yaw and use a Kalman Filter based tracker, with constant velocity model including yaw dynamics. We then refine each object trajectory using a spline optimization method. We represent the yaw ψ as a combination of basis functions, modeling both the sine $f_s(t)$ and cosine $f_c(t)$ components, and minimize the following cost function,

$$e_{\text{yaw}} = \tfrac{1}{2}\sum_j \left((f_c(t_j) - \cos\psi_j)^2 + (f_s(t_j) - \sin\psi_j)^2\right) \quad (5)$$

We then employ the x and y positions and minimize

$$e_{\text{position}} = \frac{1}{2}\sum_j \left((f_x(t_j) - x_j)^2 + (f_y(t_j) - y_j)^2\right) \quad (6)$$

where x_j and y_j represent the measured positions at time t_j. More details in the Supplementary Material.

High-quality Accumulated Depth. Further we provide high resolution LiDAR frames from the accumulated map, exploiting the pose $\mathcal{T}'$ to reintroduce moving objects corresponding to that specific pose and time, and to transform the coordinate system. To compensate for occlusions our *Adaptive Spherical Occlusion Culling* converts each point to spherical coordinates (r, θ, ϕ), define angular resolutions $\Delta\theta$ and $\Delta\phi$, and create bins

$$\theta_{\text{bins}} = \{\theta_{\min}, \theta_{\min} + \Delta\theta, \theta_{\min} + 2\Delta\theta, \ldots, \theta_{\max}\} \quad (7)$$

$$\phi_{\text{bins}} = \{\phi_{\min}, \phi_{\min} + \Delta\phi, \phi_{\min} + 2\Delta\phi, \ldots, \phi_{\max}\} \quad (8)$$

In contrast to existing methods, for each bin (i, j), we find the minimum range $r_{\min}^{(i,j)}$, that is

$$r_{\min}^{(i,j)} = \min\{r_k \mid k \in \text{bin } (i,j)\}. \quad (9)$$

and define a threshold function $T(r)$ that increases with range $T(r) = 1 + \alpha r$, where α is a small positive constant. A point k in bin (i, j) is considered visible if $r_k \leq r_{\min}^{(i,j)} + T(r_k)$ and otherwise, the point is considered occluded. These high-resolution LiDAR frames, with three to five times the density across all ranges, serve as reference data (ground truth) for depth learning from images.

4. Experiments

To validate our approach, which is unique in its ability to generate pseudo-labels simultaneously for 3 tasks, we benchmark it against state-of-the-art methods that tackle each task in isolation. We conduct experiments on the short-range datasets KITTI [21] and nuScenes [8], and an experimental long-range highway dataset that captures beyond the $80m$ LiDAR range limit of public datasets. Adhering to the common protocol adopted in recent works [20, 28, 37, 54], we first compare our pseudo-labels with human annotations and then train task-specific models on a mix of ground-truth and pseudo-labels. This evaluation highlights both the accuracy of our pseudo annotations and the extent to which models can absorb the noise introduced by pseudo-labeling.

4.1. Common Settings

Across all datasets used, we kept the parameters necessary for our method constant: we set the label propagation radius $r = 0.2m$, the probability threshold for moving points detection to $\tau_s = 0.5$ and for the probability update $\alpha = 0.7$.

4.2. Accumulated Pseudo-Depth Evaluation

We evaluate pseudo-depth generation by finetuning an NMRF [22] model, using both the short-range-LiDAR, small-baseline stereo pairs on KITTI [21] and our long-range-LiDAR, and wide-baseline cameras. We supervise NMRF with projected pseudo LiDAR and reverse Huber loss [29, 71], and validate the improvements computing Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) on the pixels where ground truth is available.

Baselines. To compare our pseudo depth generation with existing methods, we produce pseudo ground-truth depth using three distinct baselines. First, a LiDAR-based, dense pseudo depth obtained through LIO-SAM [56], to show the importance of our Adaptive Spherical Occlusion Culling and floaters refinement. Second, a monocular foundation model [64] to predict metric depth from single images. Third, a robust stereo prediction network [32] to generate depth maps. The resulting pseudo-labeled frames are used for depth supervision following a consistent train-test split.

Short Range Dataset (KITTI). We randomly sample the sequences from the KITTI training dataset to obtain training and evaluation sets. Then, we evaluate the NMRF model, by using the weights pretrained on synthetic data [40] and fine-tuned on a subset of the naively sparse LiDAR (denoted as *Oracle*). For all other methods we sub select a set of 400 stereo pairs and train for $30,000$ steps on our accumulated depth as well as on the introduced alternative sources of gt-depth data. The performance is evaluated against the pseudo ground truth generated from our accumulated LiDAR data as this allows us to evaluate ranges up to 250m. Evaluations of standard KITTI ranges are shown in the Supplementary Material. Final results are reported in Table 2, where we observe an improvement of 26.8%, 56.6% and 43.3% in MAE and 25.7%, 47.4%, 38.9% in RMSE for 0-80m, 80-150m, 150-250m ranges respectively. Moreover, we achieve an average improvement over all baselines of 46.3%, 37.2%, and 17.5%in MAE and 36.4%, 31.4%, and 16.3% in RMSE for 0-80m, 80-150m, 150-250m ranges, respectively.

	Pseudo	MAE ↓ [m]			RMSE ↓ [m]		
		0-80	80-150	150-250	0-80	80-150	150-250
KITTI	Oracle	4.48	22.03	30.76	7.62	25.66	35.83
	LIO-SAM	4.71	13.00	18.16	8.25	16.67	22.55
	CREStereo	8.19	17.72	23.90	10.99	21.62	27.05
	DA-V2	6.38	15.73	22.17	8.00	21.61	29.98
	Proposed	**3.28**	**9.57**	**17.43**	**5.66**	**13.49**	**21.89**
Long Range	Oracle	5.44	20.79	31.83	7.98	25.70	38.96
	LIO-SAM	5.53	11.89	32.33	10.51	20.11	46.22
	CREStereo	8.33	22.34	37.04	10.90	27.13	45.86
	DA-V2	8.31	23.55	32.67	11.83	28.75	39.96
	Proposed	**2.27**	**6.14**	**21.07**	**4.21**	**9.81**	**25.16**

Table 2. **Depth Estimation Evaluation** of NMRF [22], supervised with pseudo depth frames from LiDAR and Image methods, on KITTI and our Long Range Dataset. Excluding the Oracle (in gray), best results are **bold**; second bests are underlined.

Long-Range Dataset. We generate 400 ground truth samples for training extracted from diverse highway scenes. For a fair comparison, we first fine-tune the pre-trained model on our sparse LiDAR recordings, as our sensor can capture points at longer ranges compared to the Velodyne HDL-64E deployed in KITTI [21], with the same number of iterations used later for the dense ground truths. Then, we fine-tune on each of the aforementioned pseudo-ground truths. The results are reported in Table 2, where we observe an improvement of 58.7%, 70.2% and 33.2% in MAE and 47.6%, 61.3%, 35.7% in RMSE for 0-80m, 80-150m, 150-250m ranges respectively. Moreover, we achieve an average improvement against all baselines of 68.1%, 64.9% and 37.8% in MAE and 61.9%, 60.3%, and 42.6% in RMSE for 0-80m, 80-150m, 150-250m ranges respectively. Especially on our dataset, rich in highway scenarios, reference SLAM system often encounter numerous dynamic objects that leave residual traces, or "floaters" (showed qualitatively in the Supplement), which degrade the accuracy of depth predictions and *confirm that our refinement method significantly enhances performance by effectively reducing these inaccuracies*.

4.3. Semantic Pseudo-Labels Evaluation

We evaluate our semantic pseudo-labels on SemanticKITTI [3] *val* sequence 08 and on more than 40k samples of NuScenes [8], generating them using only the front-left camera for the former and all six cameras for the latter.

Semantic Pseudo Labels Comparison. We evaluate and compare our pseudo labels against LeAP [20] and Seman-

	Method	KITTI mIoU	KITTI cat.mIoU	nuScenes mIoU	nuScenes cat.mIoU
POINT	SSAM	10.7	19.7	13.4	23.2
	LeAP (points)	46.8	68.6	–	–
	Our f_{seg}	**59.4**	**69.6**	**54.9**	**62.6**
	Our (Propagated)	64.9	76.2	58.0	65.2
VOXEL	SSAM (Propagated)	11.5	23.9	25.3	29.2
	LeAP + 3D-CN (2)	58.1	81.6	–	–
	Our (Propagated)	**68.3**	**86.6**	**59.1**	**76.3**

Table 3. **Pseudo Labels SOTA Comparison.** We evaluate pseudo labels generated by our f_{seg} and the refined ones on Semantic KITTI and NuScenes, on the [35] benchmark reduced sets of classes, per-point and voxelizing, according to LeAP [20]. Best results are **bold**; second bests are underlined.

tic SAM [12]: for a fair comparison we as well consider only labeled points and reduce the number of classes to a set of 11 classes (*car, bicycle, motorcycle, other-vehicle, person,road, sidewalk, other-ground, manmade, vegetation, terrain*) as well as to the 6 coarse *category* classes (*flat, construction, object, nature, human, vehicle*), both well defined in the benchmark paper of KITTI 360 [35]. Results shown in Table 3 highlight how our pseudo labeling function f_{seg} (§ 3.2) is the most accurate in labeling LiDAR data. To further compare with LeAP, which propagates its initial labels on a $0.2m$ voxel grid, we voxelize our propagated labels at the same $0.2m$ resolution and compare with the reported best result. Thanks to the higher point-level accuracy of our propagation technique (in Table 3 point propagated), our voxelized predictions outperform all competing methods without the need for additional voxel refinement, achieving state of the art in semantic pseudo-labeling.

Quality vs. Oracle. We select PVKD [26] as fixed off-the shelf model to be trained: we keep all hyper-parameters fixed, and vary only the supervision source. In the *Oracle* case we use 100 % Semantic KITTI ground-truth labels; for *Limited GT* a randomly chosen 10% ground-truth subset; for *Our* we feed that identical 10 % subset plus 90 % pseudo labels generated by our pipeline. Each regime is repeated five times with different 10 % splits, and mIoU on *val* sequence 08 is reported in Table 4. Our pseudo labels recover near-oracle performance, with a small average difference of 1.09% mIoU and of 0.30% when classes not predicted by our method (parking, bicyclist, motorcyclist, other-ground, other-objects, trunk) are excluded. To compare these results, we train the same PVKD network with pseudo labels supervision from three alternative sources: 3D projections of Semantic SAM predictions with temporal propagation (*SSAM*) [12], the self-supervised LaserMix training scheme [28] and inference pseudo-labels from a Cylinder3D [70] model pre-trained on nuScenes [8] and lightly fine-tuned for two epochs on 2000 SemanticKITTI

Supervision	GT Pseudo	All Classes mIoU%	All Classes Oracle %	Mapped Classes mIoU%	Mapped Classes Oracle %
Limited GT	10-0	43.41	70.3	48.25	70.3
Oracle	100-0	61.73	-	68.63	-
SSAM [12]	10-90	33.70	54.6	43.17	63.0
Pre-Trained [70]	10-90	44.87	72.7	52.07	75.9
Lasermix Vx [28]	10-90	59.38	96.3	67.49	98.4
UniLiPS Full (ours)	0-100	51.48	83.4	55.10	80.3
UniLiPS 95% (ours)	5-95	59.46	96.3	66.71	97.2
UniLiPS 90% (ours)	10-90	**60.63**	**98.2**	**68.33**	**99.6**

Table 4. **Semantic Segmentation.** We evaluate pseudo labels quality supervising a PVKD [26] model with pseudo labels produced by different methods. Our results demonstrate that incorporating additional pseudo-labels is crucial for regaining oracle-level performance, as evidenced by the differences between the $10-0$ and $10-90$ configurations. Furthermore, our approach benefits from label re-weighting and accumulation, yielding significant improvements over the Semantic SAM baseline. Excluding Oracle (in gray), best results are **bold**; second bests are underlined.

frames with $1/10$ the learning rate. Across all the comparisons, the PVKD model trained on our labels delivers consistently higher mIoU than when trained on baselines pseudo labels, requiring *no* extra manual annotation, underscoring their effectiveness for semantic oracle recovery. Moreover, aside from Semantic SAM, which achieves rather weak performance, our approach is the only one that can function (0-100), entirely without ground-truth supervision. The strongest competing baseline, LaserMix, can work with small proportions of ground-truth in the GT-pseudo mix, yet it still needs some labeled data and *cannot handle the* 0% *ground-truth regime* that our method successfully addresses.

4.4. Object Detection Evaluation

We evaluate our pseudo bounding boxes performance on our highway long range dataset.

Pseudo Bounding Boxes are evaluated in Table 5 using mAP and ND-Score, with 6 meters threshold, on a maximum range of 250 meters. We compare with LISO [2], due to the similar detection-trajectory-refinement methodology. We train their model on our data and produce inference bounding box on the same validation split. Additionally, we compare against pseudo bounding boxes from ICP-Flow [37], an effective annotation-free pseudo-labeling method: we threshold its flow estimates at 1 m/s to segment movers, then derive boxes using our procedure.

Quality Vs Oracle. Secondly, we train an off-the-shelf $3D$ detector following the architecture of PointPillars[30] on full ground truth (*Oracle*) and on 20% ground truth and 80% pseudo labels, generated by our method (*Proposed*) and by using ICP-Flow, as described before. We report the results in Table 6, where we find our pseudo labels can

	ICP–Flow [37]	LISO [2]	**Ours**
mAP [%]	7.2	21.1	**31.0**
NDS [%]	11.4	40.9	**45.2**

Table 5. **Pseudo Bounding Boxes** evaluation on the highway-driving dataset. We achieve state-of-the-art compared to other pseudo-labeling and detection-from-motion approaches.

Method	**-25 - 25m**		**-50 - 50m**		**-70 - 70m**	
	bev AP	3d AP	bev AP	3d AP	bev AP	3d AP
Oracle	35.55	34.45	33.54	33.08	32.44	30.13
ICP-Flow [37]	11.01	3.60	9.45	3.41	9.44	3.20
Proposed	**31.02**	**26.53**	**29.43**	**25.44**	**29.19**	**25.33**

Table 6. **Object Detection Evaluation Results** on the challenging experimental highway dataset: the model trained on our pseudo labels achieves near-*Oracle* performances compared to baseline methods. Excluding Oracle (greyed out), best results are **bold**.

	Ablated	**None**	**SAM2**	**OneF**	**BLIP**	**CLIP**
KI	**mIoU [%]**	**59.4**	58.4	10.3	31.5	59.3
	Cat-mIoU [%]	**69.6**	68.3	19.1	51.0	69.3
NU	**mIoU [%]**	**54.9**	50.2	21.4	19.6	50.5
	Cat-mIoU [%]	**62.6**	59.0	26.0	28.7	60.5

Table 7. **Pseudo Labeling Engine** mIoU and category mIoU degradation ablating each engine module, evaluating on Semantic KITTI (KI) and NuScenes (NU).

achieve near-oracle performances compared to other effective pseudo labeling methods.

5. Ablations

Figure 5 shows qualitatively the importance of our geometry grounded label propagation for temporal consistency and reweighting of mislabeled points. Table 7 reports point-wise mIoU after ablating each f_{seg} sub-model, highlighting their individual impact. Additionally, we note that increasing the number of top-k CLIP proposals ($k > 3$) doesn´t impact the mIoU score on the evaluated datasets. In Table 8a we analyze performance drop of our pseudo bounding boxes after ablating the spline optimization, which helps pose and orientation score, and our f_{iwu}, which increases detection probability. In Table 8b we complement Table 3 point-wise evaluation ablating sequentially Algorithm 1, the accumulation and the occlusion mask in the lifting module (§3.2): the former effectively re-weights labels, especially in dense regions, for more accurate prediction, while the latter removes noise from penetration and misaligned projections. More ablations are presented in the Supplement.

6. Conclusion

We propose an unsupervised pseudo-labeling method that generates semantic labels, bounding boxes, and precise

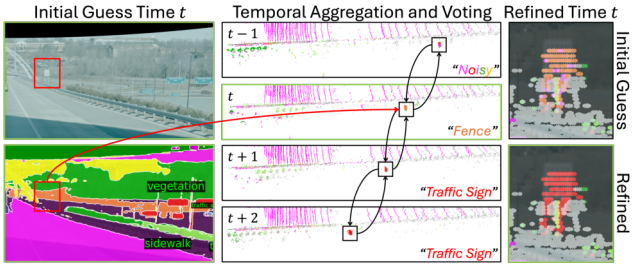

Figure 5. **Effect of Semantic Multimodal Propagation.** Leveraging our refined, geometry-grounded map as a reference, mislabeled points in each LiDAR scan are systematically corrected, ensuring label consistency across all timestamps.

	Full	**w/o spline opt.**	**w/o f_{iwu}**
mAP [%]	**31.0**	23.9	11.7
NDS [%]	**45.2**	33.2	30.8

(a) Pseudo Bounding Boxes Ablations.

	Ablated	**None**	**Agorithm 1**	**Accumulation**	**Occ. Mask**
KI	**mIoU [%]**	**64.9**	60.7	59.4	57.0
	cat.-mIoU [%]	**76.2**	70.4	69.6	68.5
NU	**mIoU [%]**	**58.0**	55.2	54.9	50.1
	cat.-mIoU [%]	**65.2**	64.9	62.6	55.3

(b) Pseudo Semantic Labels Ablations.

Table 8. **Ablation Experiments.** Pseudo bounding boxes results (a) ablating the spline optimizer and f_{iwu}, and semantic labels results (b) ablating sequentially the label propagation algorithm, the accumulation (camera-only) and the occlusion mask in the lifting.

long-range depth from LiDAR, camera and IMU datas recorded in a single driving trajectory. Our approach is based on a geometry-grounded dynamic scene decomposition: we first reconstruct a LiDAR map of the environment, then propagate semantic labels from vision foundation models across each observed point. By detecting and reconciling inconsistencies, we remove moving objects and correct label errors, enabling a truly automatic annotation pipeline that achieves near-oracle performance compared to manual labeling. Our method is not tailored to any specific sensor configuration and generalizes successfully across KITTI, NuScenes and our Long Range datasets. We validate that the generated pseudo-labels achieve state-of-the-art in semantic segmentation and object detection and consistently enhance depth estimation up to $250m$, with improvement of 51.5% in MAE between 80 and 150 meters and 22.0% between 150 and 250 meters.

7. Acknowledgments

Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, a Amazon Science Research Award, and a Bosch Research Award.

References

[1] Chunge Bai, Tao Xiao, Yajie Chen, Haoqian Wang, Fang Zhang, and Xiang Gao. Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels. *IEEE Robotics and Automation Letters*, 7(2):4861–4868, 2022. 2

[2] Stefan Baur, Frank Moosmann, and Andreas Geiger. Liso: Lidar-only self-supervised 3d object detection. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 7, 8

[3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 6

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 2

[5] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

[6] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv*, 2024. 2

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 1

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 6, 7

[9] Mu Cai, Chenxu Luo, Yong Jae Lee, and Xiaodong Yang. Cross-modal self-supervised learning with effective contrastive units for lidar point clouds. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9468–9475, 2024. 1

[10] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image, 2017. 1

[11] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8740–8749, 2019. 1

[12] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023. 7

[13] Kenny Chen, Ryan Nemiroff, and Brett T Lopez. Direct lidar-inertial odometry: Lightweight lio with continuous-time motion correction. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3983–3989, 2023. 2

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 1

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[17] Zihao Dong, Jeff Pflueger, Leonard Jung, David Thorne, Philip R. Osteen, Christa S. Robison, Brett T. Lopez, and Michael Everett. Lidar inertial odometry and mapping using learned registration-relevant features. *ArXiv*, abs/2410.02961, 2024. 2

[18] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator, 2017. 1

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 1

[20] Simon Gebraad, Andras Palffy, and Holger Caesar. Leap: Consistent multi-domain 3d labeling using foundation models, 2025. 2, 6, 7

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1, 6

[22] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2024. 6

[23] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[24] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[25] K. Tan et al. H. Caesar, J. Kabzan. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. In *CVPR ADP3 workshop*, 2021. 1

[26] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022. 7

[27] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation, 2022. 3

[28] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023. 2, 6, 7

[29] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 6

[30] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 7

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3, 4

[32] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 6

[33] Jianhao Li, Tianyu Sun, Zhongdao Wang, Enze Xie, Bailan Feng, Hongbo Zhang, Ze Yuan, Ke Xu, Jiaheng Liu, and Ping Luo. Segment, lift and fit: Automatic 3d shape labeling from 2d prompts, 2024. 1

[34] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. 2024. 2

[35] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023. 7

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1, 3

[37] Yancong Lin and Holger Caesar. Icp-flow: Lidar scene flow estimation with icp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15501–15511, 2024. 2, 6, 7, 8

[38] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 3, 4

[39] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. Cad-estate: Large-scale cad model annotation in rgb videos. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20132–20142, 2023. 1

[40] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 6

[41] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017. 5

[42] Siva Karthik Mustikovela, Shalini De Mello, Aayush Prakash, Umar Iqbal, Sifei Liu, Thu Nguyen-Phuoc, Carsten Rother, and Jan Kautz. Self-supervised object detection via generative image synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[43] Lucas Nunes, Louis Wiesmann, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Temporal Consistent 3D LiDAR Representation Learning for Semantic Perception in Autonomous Driving. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[44] Aljoša Ošep, Tim Meinhardt, Francesco Ferroni, Neehar Peri, Deva Ramanan, and Laura Leal-Taixé. Better call sal: Towards learning to segment anything in lidar, 2024. 2

[45] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[46] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025. 2

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3, 4

[48] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2

[49] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 2

[50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[51] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 1

[52] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9891–9901, 2022. 1

[53] Corentin Sautier, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Vincent Lepetit. UNIT: Unsupervised online instance segmentation through time. In *3DV*, 2025. 2

[54] Jenny Seidenschwarz, Aljosa Osep, Francesco Ferroni, Simon Lucey, and Laura Leal-Taixe. Semoli: What moves together belongs together. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14685–14694, 2024. 2, 6

[55] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018. 2

[56] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Rus Daniela. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5135–5142. IEEE, 2020. 2, 6

[57] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[58] Xavier Timoneda, Markus Herb, Fabian Duerr, Daniel Goehring, and Fisher Yu. Multi-modal nerf self-supervision for lidar semantic segmentation, 2024. 1

[59] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns, 2017. 2

[60] 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, 2020. 2

[61] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, 2014. 1

[62] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In *European Conference Computer Vision (ECCV)*. 2016. 1

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2, 6

[65] Sergey Zakharov, Wadim Kehl, Arjun Bhargava, and Adrien Gaidon. Autolabeling 3d objects with differentiable rendering of sdf shape priors, 2020. 1

[66] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. 2021. 1

[67] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[69] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 3

[70] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception, 2021. 7

[71] Laurent Zwald and Sophie Lambert-Lacroix. The berhu penalty and the grouped effect, 2012. 6