

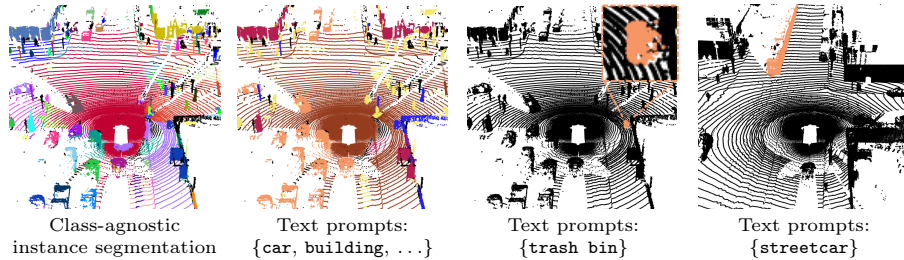
# Better Call SAL: Towards Learning to Segment Anything in Lidar

Aljoša Ošep<sup>1\*</sup> Tim Meinhardt<sup>1\*</sup> Francesco Ferroni<sup>1</sup> Neehar Peri<sup>2</sup>  
Deva Ramanan<sup>2</sup> Laura Leal-Taixé<sup>1</sup>

\*Equal contribution

<https://github.com/nv-dvl/segment-anything-lidar>

<sup>1</sup>NVIDIA <sup>2</sup>CMU



**Fig. 1:** The SAL (Segment Anything in Lidar) model performs class-agnostic instance segmentation (i) and zero-shot classification via text prompting. This allows us to not only predict panoptic segmentation (ii) for fixed class vocabularies but also segment any object (iii and iv) in a given Lidar scan.

**Abstract.** We propose the SAL (Segment Anything in Lidar) method consisting of a text-promptable zero-shot model for segmenting and classifying any object in Lidar, and a pseudo-labeling engine that facilitates model training without manual supervision. While the established paradigm for *Lidar Panoptic Segmentation* (LPS) relies on manual supervision for a handful of object classes defined a priori, we utilize 2D vision foundation models to generate 3D supervision “for free”. Our pseudo-labels consist of instance masks and corresponding CLIP tokens, which we lift to Lidar using calibrated multi-modal data. By training our model on these labels, we distill the 2D foundation models into our Lidar SAL model. Even without manual labels, our model achieves 91% in terms of class-agnostic segmentation and 54% in terms of zero-shot LPS of the fully supervised state-of-the-art. Furthermore, we outperform several baselines that do not distill but only lift image features to 3D. More importantly, we demonstrate that SAL supports arbitrary class prompts, can be easily extended to new datasets, and shows significant potential to improve with increasing amounts of self-labeled data. We release all models and the code.

## 1 Introduction

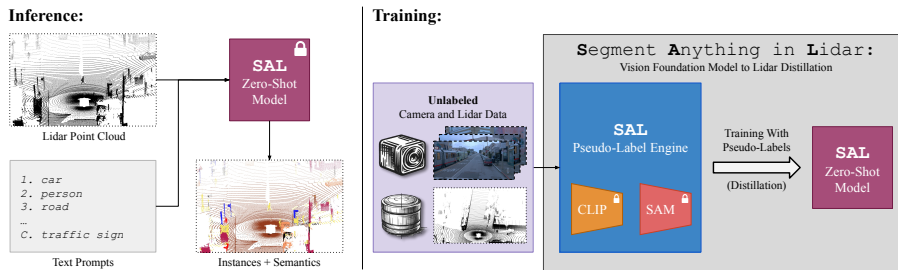
We tackle segmentation and recognition of objects in Lidar point clouds, a task commonly tackled via *Lidar Panoptic Segmentation* (LPS).

**Status quo.** LPS has been gaining significant attention in the community due to its role in scene understanding, which is vital for safe autonomous navigation. Thanks to the availability of labeled datasets [6, 18] and advances in learning representations from unordered point sets [12, 65, 66, 79], the Lidar community made significant progress in learning to segment and classify instances of pre-defined and manually-labeled classes. While this progress has been impressive, existing models [2, 26, 36, 45, 70, 74, 98] cannot adapt to continually evolving class ontologies [40] that may even vary in different geographic regions [82].

**Stirring the pot.** We challenge this well-established approach and investigate how we can train general LPS models that can be prompted to segment point clouds according to *any* object class vocabulary. Towards such a promptable Lidar segmentation approach, we first need to be able to segment any object, a capability that remains elusive in the Lidar domain.

**Towards segmenting anything in Lidar.** To address these challenges, we propose SAL, which consists of a text-promptable zero-shot model (Fig. 2, *left*) for panoptic segmentation of arbitrary objects (Fig. 1), and pseudo-labeling engine (Fig. 2, *right*) that facilitates model training directly from raw sensory data. The SAL pseudo-label engine automatically labels Lidar sequences using image segmentation [31] and vision-language models [67]. We utilize SAM [31] to generate class-agnostic masks in images, and CLIP [67] to generate per-mask tokens that connect visual features to language, and finally, transfer both to Lidar using a calibrated sensory setup. Even though generated pseudo-labels only partially cover Lidar scans and are inherently noisy due to errors in the image-level generation process and imperfect sensory calibration, we demonstrate their effectiveness as self-supervised training data for the SAL zero-shot model. In contrast to prior work in LPS, SAL can be prompted with any semantic vocabulary during test-time without model re-training or tuning. Different from recent advances in zero-shot semantic segmentation [58] SAL *does not* require image features during inference and can segment full 360° point clouds.

**Talk to your SAL.** Our model predicts Lidar segmentation masks and their corresponding CLIP tokens, which allow us to perform zero-shot classification of segmented objects using arbitrary text prompts (Fig. 1). We evaluate SAL by prompting our zero-shot model on standard benchmarks for LPS. For class-agnostic segmentation, we reach 91% of the performance of manually supervised baselines. By classifying segmented objects in a zero-shot manner, we report the *first* (and very encouraging) results for zero-shot LPS and reach 42% and 54% of the supervised baselines trained on SemanticKITTI and nuScenes, respectively. Beyond that, as shown in Fig. 1, we can prompt SAL model to segment objects outside of existing Lidar dataset class vocabularies.



**Fig. 2: SAL overview:** Given a Lidar scan and a class vocabulary prompt, specified as a list of per-class free-form text descriptions (*left*), SAL segments and classifies objects (**things** and **stuff** classes). As labeled data for training such a model does not exist, we supervise SAL by distilling off-the-shelf vision foundation models to Lidar (*right*).

**Contributions.** As our main contribution, we re-think the established approach to *Lidar Panoptic Segmentation* and (i) present SAL for segmentation and classification of any object in a Lidar scan. As we do not utilize labeled Lidar data, we propose (ii) a pseudo-label engine that distills vision foundation models to the Lidar domain. The resulting pseudo-labels allow us to train (iii) a zero-shot LPS model without any human supervision. We demonstrate (iv) encouraging results on standard LPS benchmarks and outline a clear path towards universal, promptable segmentation foundation models for Lidar data.

## 2 Related Work

Lidar has played a pivotal role since the dawn of embodied navigation [62,80,81]. Recent data-driven efforts in Lidar perception have been pushing boundaries in semantic segmentation [3, 12, 37, 49, 71, 77, 84, 85, 87, 100], object detection [33, 42, 44, 59–61, 90, 92, 97], and tracking [25, 51, 78].

**Lidar panoptic segmentation.** Recently, *Lidar Panoptic Segmentation* [4, 6, 7, 18] has emerged as a holistic approach to Lidar-based dynamic scene understanding. Prior works [2, 4, 19, 26, 28, 32, 36, 45–47, 70, 91, 98, 99] learn to group and classify points specified in the training data according to the target class vocabularies, while methods for open-set instance [24, 51, 78] and panoptic segmentation [83] rely on bottom-up grouping based on Euclidean distance between points. While these developments have been impressive, end-to-end methods remain limited to specific target classes that appear in the training data, while bottom-up, clustering-based approaches [24, 27, 51, 78, 83] are sensitive to a particular choice of clustering parameters and are unable to improve their performance in a data-driven fashion. In contrast, the SAL zero-shot model not only *learns* to segment objects but is also capable of zero-shot classification.

**Self-supervised learning for Lidar.** As labeled data in the Lidar domain is scarce, several works investigate how contrastive learning, proven effective in the image domain [10, 23], can reduce the need for labeled data. SegContrast [54] learns to align representations of point clouds and their augmented views, utilizing density-based clustering (DBSCAN [17]) to pool and contrast

features. Rather than using DBSCAN, [41, 72] leverage image-based methods [1] and foundation models [31] to perform contrastive alignment between image and corresponding Lidar features. The aforementioned methods only provide pre-training recipes for potential segmentation downstream tasks. Our SAL method goes a step further by training a full zero-shot *Lidar Panoptic Segmentation* model that could benefit from such pre-training recipes.

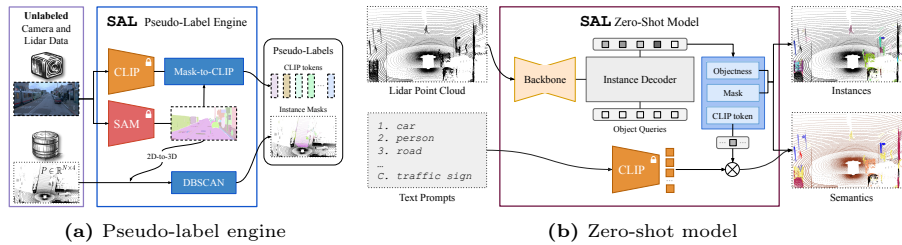
**Self-supervised (3D) object detection.** Several methods (self-)supervise object detectors with (RGB-D) videos [22, 55–57, 63, 64], or Lidar sequences [52, 73, 94] to discover objects in sensory streams. However, such motion-based supervision provides no semantic information. Therefore, [53] distills per-point (PCA quantized) CLIP [67] features to Lidar for zero-shot classification of detected objects. Significantly different from [53], we tackle the more general and challenging task of LPS, which segments and classifies both moving and stationary **things** and **stuff** classes. To this end, we distill CLIP features not per point but per object instance. This provides a more holistic object-centric semantic distillation and allows us to supervise our model with non-quantized CLIP features.

**Zero-shot recognition.** Zero-shot recognition (ZSR) [86] methods (recently also referred to as “open-vocabulary recognition”) tackle the recognition of distinct, yet related, semantic concepts that are either not labeled in the training data (transductive) or not observed at all (inductive). Such methods assume a dataset with labels for *some* classes, whereas unseen instances are supervised via attributes or class names. The latter is often used in conjunction with word embeddings [48]. By learning to align image-language features, ZSR methods can infer class labels for unseen objects [67]. ZSR has also been explored in the context of object detection, semantic segmentation, and panoptic segmentation [16, 88] using word embedding models [5, 8, 50, 68], or CLIP vision-language embeddings [20, 21, 35, 38, 69, 89, 93, 95, 96]. In the context of semantic segmentation in RGB-D and Lidar data, [43, 58] augment image and language embeddings with features extracted from point clouds. Both rely on lifting dense image features from images to 3D during inference, which limits their applicability to setups with dense camera coverage. Similarly, [76] segments instances in accumulated RGB-D point clouds (ScanNet [14]) and relies on manual supervision in conjunction with image-based dense CLIP features for zero-shot classification. Moreover, datasets such as ScanNet [14] were recorded with RGB-D sensors in *static* environments and fused into 3D reconstructions based on scans, taken from multiple viewpoints. In contrast, SAL directly applies to *full* Lidar point clouds while requiring no image features at the inference time.

### 3 SAL: Segment Anything in Lidar

This section outlines the key challenges and components towards a Lidar model that segments and classifies any object. We first define the underlying problem formation in Sec. 3.1 and then present our (Segmenting Anything in Lidar) SAL model in Sec. 3.2.





**Fig. 3:** Our **pseudo-label engine** (Fig. 3a) utilizes SAM [31] to estimate segmentation masks in images, MaskCLIP [16] to estimate corresponding per-mask CLIP features, and a calibrated sensory setup to transfer them to the Lidar domain. We distill these pseudo-labels to our **zero-shot model** (Fig. 3b), which segments and classifies Lidar point clouds. The SAL zero-shot model employs a sparse-convolutional backbone [12], followed by a Transformer decoder that predicts objectness scores, segmentation masks, and CLIP tokens for each query. To (optionally) perform zero-shot classification, we forward the dataset class vocabulary through the CLIP text encoder and match the encoded vocabulary with predicted CLIP tokens. Our model requires no retraining for different vocabularies and no image features at inference time.

### 3.1 Zero-Shot Lidar Panoptic Segmentation

**The task.** We discuss and evaluate SAL in the context of *Lidar Panoptic Segmentation* (LPS), which tackles both instance segmentation and object recognition. LPS methods take as input an unordered point cloud  $P \in \mathbb{R}^{N \times 4}$  encoding spatial coordinates and their sensor intensity. The currently established problem setting [7, 18] assumes supervision and task outputs as per-point semantic class and instance identity labels. The class labels are confined to a pre-defined and fixed class vocabulary  $\mathcal{V}$ . **In contrast**, we tackle LPS in a generalized, zero-shot setting, where the class vocabulary is only provided at inference, not training time. Such a vocabulary  $\mathcal{V}$  specifies target classes as a list of  $C$  class prompts. Each prompt  $c_i \in \mathcal{V}$  is specified via free-form text, *e.g.*, with a class name and optional class description. To perform zero-shot LPS a model must be designed to segment and classify any object in a Lidar scan.

**The challenge.** *Where do we get the data to train such a model?* Image-based models [16, 21, 31, 93, 95] rely on large and diverse datasets for pre-training [15], dense prediction [13, 39], and foundation models that align images with textual descriptions [67] – a commodity not available in the Lidar domain.

### 3.2 SAL Overview

Our 2D-to-Lidar distillation method consists of two core components: (i) The **pseudo-label engine** transfers 2D vision foundation models into Lidar pseudo-labels using multi-modal inputs from a calibrated sensory setup, shown in Fig. 3a. (ii) Our **zero-shot model** is trained on the generated pseudo-labels and is able to perform class-agnostic segmentation and zero-shot classification via text prompts. Given a semantic dataset vocabulary, this allows us to tackle zero-shot LPS as illustrated in Fig. 3b.

### 3.3 SAL Pseudo-Label Engine

The SAL pseudo-label engine relies on a calibrated multi-modal sensory setup with  $k \geq 1$  RGB cameras. Furthermore, a sufficient overlap between the Lidar sensing area and each camera view is paramount.

**Mask generation:** To pseudo-label object instances, we utilize the *Segment Anything* (SAM) model [31], which generates an overlapping set of segmentation masks. We flatten SAM’s output mask hierarchy by non-maxima suppression (NMS) with a minimal overlap threshold to ensure mutually exclusive masks, suppressing object parts and subparts in favor of objects. This way, we obtain a set of non-overlapping binary masks  $m_i^k \in \{0, 1\}^{W \times H}$  for each camera view  $k$  with image plane of size  $W \times H$ .

**CLIP image token generation:** As shown by CLIP [67], aligning image and text features allows vision-language foundation models to perform zero-shot image classification based on arbitrary text prompts. To transfer this capability to the Lidar domain, we generate a localized CLIP image feature token  $f_i^k \in \mathbb{R}^{C_t}$  for each binary SAM mask  $m_i^k$ . To obtain image tokens for a masked region of the input image, we utilize [16] and their relative mask attention in the CLIP image encoder feature space. Note that we never classify segments in the image domain but merely distill CLIP image feature tokens to Lidar to facilitate zero-shot classification during inference where we do not use any image features.

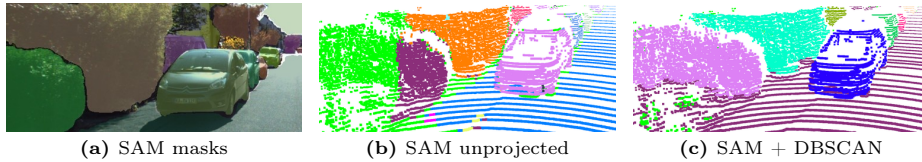
**Image-to-Lidar unprojection:** From the Lidar perspective, we unproject each image mask  $m_i^k \in \{0, 1\}^{W \times H}$  to a binary Lidar segmentation mask  $\tilde{m}_i \in \{0, 1\}^N$  by transforming the respective camera coordinate frame to the Lidar space. For datasets with multiple cameras, such as nuScenes [18], we process each image independently, followed by a cross-camera fusion of masks with a sufficient IoU overlap and averaging of their CLIP image tokens. The unprojection yields pairs  $\{\tilde{m}_i^k, f_i^k\}$  of Lidar masks and their corresponding CLIP features.

**Refinement via density-based clustering:** Image-to-Lidar transformation (Fig. 4a→Fig. 4b) is inherently noisy due to imperfect calibration and issues with synchronization and Lidar rolling shutter. We improve our pseudo-label quality by creating an ensemble of DBSCAN [17] clusters  $\tilde{m}_l \in \{0, 1\}^N$ , obtained by varying the density threshold to compensate for varying density in Lidar point clouds (details in Appendix A.1). We replace each  $\tilde{m}_i$  with its best-matching  $\tilde{m}_l$  in case their IoU exceeds a minimal overlap threshold and retain the original mask otherwise to obtain a refined set of pseudo-labels (Fig. 4c) that retain their original cardinality and associated CLIP features.

### 3.4 SAL Zero-Shot Model

The universal SAL zero-shot model deconstructs LPS into class-agnostic segmentation and zero-shot classification via text prompts.

**Universal architecture.** Instead of relying on highly specialized and engineered LPS models, we base the SAL model on a universal Transformer decoder



**Fig. 4: Refinement via clustering.** After transferring image masks (Fig. 4a) to Lidar (Fig. 4b), we obtain pseudo-labels that suffer from sensory misalignment-related issues. Our geometric refinement (Fig. 4c) improves localization.

architecture (Fig. 3b), similar to [45]. Its 2D counterparts [9, 11] provide top-performing results across semantic, instance, and panoptic segmentation for images. With increasing amounts of (pseudo-)labeled data, Transformer decoders have the potential to achieve a similar impact in the Lidar domain. Following [45], we deploy a Minkowski U-Net [12] backbone for feature extraction followed by a Transformer decoder architecture with object query to point/voxel feature cross-attention. Training our model to segment and classify *any* object and perform the evaluation in a zero-shot setting requires a unique design of the final task heads.

**Class-agnostic segmentation.** To localize segments in a point cloud, we rely on two model heads that predict segmentation masks and objectness scores for each query. The architecture of the former follows [45] and predicts binary output masks  $\hat{m}_i \in \{0, 1\}^N$  by computing the dot product between queries and point features. The objectness head reduces the multi-class problem to a binary object or no-object decision.

**Zero-shot classification.** To equip our model with zero-shot classification capabilities without relying on any image input, we learn to predict CLIP [67] Lidar tokens, *i.e.*, features in the CLIP space obtained from Lidar inputs, for each query. The token head consists of a three-layer MLP that directly regresses tokens  $\hat{f}_i^j$ . At inference, the predicted tokens are matched to text prompts via the CLIP text encoder. The matching yields a probability distribution over the prompts. To perform zero-shot LPS on a pre-defined dataset, we use its class vocabulary as input text prompts (Fig. 3b).

We train our model jointly for both class-agnostic segmentation and zero-shot classification with the following loss:

$$\mathcal{L}_{\text{SAL}} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{token}}, \quad (1)$$

with binary cross-entropy loss  $\mathcal{L}_{\text{obj}}$  and a cosine distance loss  $\mathcal{L}_{\text{token}}$ . The segmentation loss  $\mathcal{L}_{\text{seg}}$  follows [45] and consists of a binary mask cross-entropy and dice loss. The  $\mathcal{L}_{\text{SAL}}$  loss is supervised by pseudo-label pairs  $\{\tilde{m}_i^k, f_i^k\}$  obtained from our label engine and thereby distills both 2D foundation models (SAM and CLIP) into our LPS model.

**How to train with partial labels?** Partially (pseudo-)labeled point clouds present a challenge for SAL model training. For example, in SemanticKITTI, only 14% of points are pseudo-labeled due to low camera coverage (see Tab. B.1). If

we naively train SAL on partially (pseudo-)labeled point clouds, the objectness loss  $\mathcal{L}_{obj}$  would penalize any segmentation in these regions as a false positive, thereby teaching the model to ignore these points entirely. Empirically, we determined that the most effective training strategy is to (i) remove unlabeled points from point clouds, (ii) utilize standard data augmentations (rotations, flipping, scaling, and translations), in conjunction with proposed *FrankenFrustum* augmentation (detailed in Appendix A.4) that mimics fully-labeled point clouds during training by randomly removing unlabeled points and replicating labeled frustum regions around the vertical axis. This augmentation does not increase the overall label coverage, but our ablations in Sec. 4.2 show its effectiveness in reducing the domain gap between training and inference input clouds.

## 4 Experiments

In the following section, we outline our experimental setup (Sec. 4.1) used to ablate the SAL model and its training strategies (Sec. 4.2). Since SAL presents the first *Zero-Shot Lidar Panoptic Segmentation* (ZS-LPS) model, we design and compare with multiple hand-crafted baselines that implement concepts from related works. We conclude our analysis by comparing our zero-shot performance (Sec. 4.4) to fully-supervised LPS methods.

### 4.1 Experimental Setup

**Datasets.** We evaluate SAL on two public Lidar Panoptic Segmentation (LPS) datasets, SemanticKITTI [6, 7] and nuScenes [18]. We utilize provided, human-annotated ground-truth (GT) panoptic segmentation labels solely for evaluation purposes, linear probing ablations, and baselines. We train SAL model *only* our pseudo-labels (Sec. 3.3). To demonstrate the general applicability of SAL to other domains and datasets, we further show qualitative results of SAL trained on Waymo Open [75] dataset, which contains *no* GT LPS labels.

**Metrics.** For the evaluation, we utilize the standard Panoptic Segmentation [30] metrics. The Panoptic Quality  $PQ = RQ \times SQ$  combines Segmentation Quality ( $SQ$ ) and Recognition Quality ( $RQ$ ), *i.e.*, F-1 score. True positives (TPs) are mask predictions with sufficient intersection-over-union (IoU) overlap with GT masks of the same class.

**Semantic Oracle (SO).** The aforementioned metrics assume GT labels and predictions with instance IDs and semantic classes for each point. Therefore, they are not suitable for evaluating class-agnostic segmentation. To assess the class-agnostic segmentation independently of semantics, we apply SO during the evaluation by assigning ground truth semantic classes  $c_i$  to predicted masks  $m_i$  via majority voting. Evaluation with SO is only needed for evaluation of class-agnostic segmentation – our zero-shot models provide both instance *and* semantic class predictions.

**Stuff Merging (SM).** SO allows us to assess class-agnostic *instance* segmentation for **things** classes. However, this approach is unsuitable for assessing

the performance of **stuff** classes, as existing Lidar datasets merge instances of **stuff** classes into a single instance. The **things-stuff** separation can be unintuitive and inconsistent across datasets (*e.g.*, SemanticKITTI merges instances of traffic signs into a single **stuff** class). In contrast, we fully embrace the philosophy that *all* classes can be segmented into individual instances (*e.g.*, segmentation models should localize individual trees/bushes in the **vegetation** class or distinguish individual **buildings**). To evaluate our models on the target datasets in the SO regime on Lidar datasets that do not provide instance-level annotations for all classes, we additionally report results using a *merge* strategy (SM), which merges all instances of a particular **stuff** class into a single mask to ensure outputs of our models are consistent with the format of target datasets.

## 4.2 SAL Ablations

In this section, we discuss *how* to train SAL zero-shot segmentation model using self-generated pseudo-labels (Tab. 1) and discuss design decisions behind our instance (Tab. 2) and semantic distillation (Tab. 3).

**Learning with partial labels.** For the single-camera setup [6,7], our pseudo-labels only cover 14% of the input point clouds (see Tab. B.1). Training on partial label coverage presents a challenge for our model.

In Tab. 1, we only evaluate class-agnostic segmentation by applying SO and SM (see Sec. 4.1). We compare SAL models trained using *pseudo* and *GT* labels. The *Frustum Filter* removes all unlabeled points not visible in the camera frustum during training and/or evaluation. As shown in Tab. 1, removing unlabeled points from pseudo-labeled point clouds during training shows a *significant* benefit on the performance (3<sup>rd</sup> row, 59.3 PQ) as compared to the variant, where we merely ignore unlabeled region during the training (2<sup>nd</sup> row, 22.2 PQ). This is likely due to data imbalance in a single-camera setup, leading to a class imbalance between labeled/unlabeled points, incentivizing SAL to suppress predictions.

While utilizing standard augmentations (rotations, flipping, scaling, and translations) is crucial, we observe that performing our *FrankenFrustum* augmentation, which concatenates the visible portion of the point cloud around the z-axis (see Sec. 3.4), significantly improves results (4<sup>th</sup> row, 62.5 PQ). Mixing point clouds from different scans (5<sup>th</sup> row, 62.8 PQ) further boosts results. Overall, we obtain 69 PQ from GT supervision (1<sup>st</sup> row) and 62.8 PQ when using our pseudo-labels (5<sup>th</sup> row). Remarkably, our labels yield 91% of the GT performance while covering only 14% of the input point cloud (see Tab. B.1). To further contextualize these results, we note that this evaluation is performed *only* on classes for which GT instance labels are available. As shown in Fig. 1, our model learns to segment a much larger variety of classes. Evaluating only the subset of points visible in the camera (14% of all points, rows 6&7), the gap between the GT-supervised model (71.8 PQ) and pseudo-supervised model (70.7 PQ) shrinks further. *We conclude that by training SAL model using pseudo-labels, we distill the notion of objectness from SAM [31] into our model and obtain segmentation capabilities, similar to the model trained on GT data.*

**Table 1: Class-agnostic segmentation.** By cropping unlabeled points and performing data augmentations (in combination with our FrankenFrustum), SAL successfully learns to segment full point clouds even when only 14% of points are (pseudo)-labeled.

Labels	Frust.-Filter Train Eval	Franken Frust.	PQ	SQ	PQ <sup>Th</sup>	PQ <sup>St</sup>
GT			69.0	83.5	81.6	59.7
Pseu.			22.2	67.5	45.1	5.5
Pseu.	✓		59.3	78.2	65.8	54.5
Pseu.	✓	✓	62.5	79.1	67.6	58.8
Pseu.	✓	✓(mix)	62.8	79.0	69.0	58.3
GT			71.8	84.8	84.6	62.5
Pseu.	✓	✓	70.7	81.9	75.4	67.3

Train Labels	Set	Num. Quer.	PQ	SQ	PQ <sup>Th</sup>	PQ <sup>St</sup>
GT	train	100	67.8	82.4	79.7	59.2
GT	train	200	67.9	83.6	79.8	59.3
GT	train	300	69.0	83.5	81.6	59.7
Pseu.	train	100	53.9	75.3	62.2	47.8
Pseu.	train	200	60.8	77.0	66.7	56.5
Pseu.	train	300	62.8	79.0	69.0	58.3
Pseu.	bigtrain	300	65.3	79.5	71.9	60.5

**Scaling queries and data.** Analogous to the grid inference of SAM, which operates with hundreds of spatial queries (prompts), we analyze the effect of different numbers of decoder queries. We subsample instances during training to train models with fewer queries than the maximum number of segments per scan. As visualized in Tab. 2, increasing the number of decoder queries improves the recognition of **things** for both training with pseudo and GT labels. However, we observe a significant performance boost with *pseudo* labels for **stuff** classes (58.3 PQ<sup>St</sup>, +10.8), while the improvement for GT labels is marginal (59.7 PQ<sup>St</sup>, +0.5). Our model benefits from a larger number of queries for **stuff** classes as it learns a fine-grained segmentation model that is not limited to a prefixed set of **things** classes. In the last row, we concatenate pseudo-labeled *train* and *test* sets (neither used for the validation) into a *bigtrain* set and observe an improvement of +2.5 PQ, suggesting that scaling the amount of training data has the potential to improve the performance of our models further.

**Table 3: Semantic distillation.** We linearly probe the SAL model, trained with and without semantic distillation loss  $\mathcal{L}_{token}$ . We train a linear classifier with GT labels while keeping backbone and decoder features frozen. As can be seen,  $\mathcal{L}_{token}$  successfully distills a notion of semantics into our model.

Linear prob. $\mathcal{L}_{token}$	PQ	RQ	SQ	mIoU
×	20.0	26.3	55.2	23.4
×	33.1	41.9	68.3	40.0
×	24.8	32.3	66.8	29.7

decoder remain frozen) and train it using GT labels. As can be seen in Tab. 3, distilling CLIP features into our model *significantly* improves linear probing

**Semantic distillation.** At inference, SAL predicts a binary segmentation mask and objectness score for each query. To perform zero-shot classification, our model additionally predicts CLIP image tokens. This allows us to not rely on image features during inference but prompt our predicted tokens with arbitrary texts, *i.e.*, object classes. *Does the token head loss  $\mathcal{L}_{token}$  successfully distill a notion of object semantics from CLIP into our model?* To verify this, we perform *linear probing* experiments: we replace the token distillation head of a trained model with a linear classifier (backbone and instance de-



**Table 4: Zero-shot panoptic segmentation.** We utilize prior efforts in the image domain [31] and Lidar [58] domain to craft multiple baselines that only unproject segmentation masks and lift image features to Lidar. By contrast, SAL distills outputs of such baselines (pseudo-labels) into a stronger Lidar segmentation model. With *Image Feat.* we denote methods that require image features at inference time, and *Frust. Eval.* denotes the evaluation of a subset of points visible in the camera.

Method	Frust. Image		Default classes					Super classes				
	Eval	Feat.	PQ	SQ	PQ <sup>Th</sup>	PQ <sup>St</sup>	mIoU	PQ	SQ	PQ <sup>Th</sup>	PQ <sup>St</sup>	mIoU
Class-agnostic Segmentation (Semantic Oracle)												
SAM	✓	✓	46.0	72.1	49.7	43.4	-	-	-	-	-	-
SAM+Erosion	✓	✓	42.2	69.4	45.6	39.6	-	-	-	-	-	-
SAM+DBS (filter)	✓	✓	46.7	70.3	76.8	24.8	-	-	-	-	-	-
SAM+DBS (replace)	✓	✓	48.7	73.7	53.1	45.4	-	-	-	-	-	-
SAL	✓	✗	70.7	81.9	75.4	67.3	-	-	-	-	-	-
Zero-Shot Lidar Panoptic Segmentation												
SAM+DBS+CLIP	✓	✓	27.5	71.5	31.7	24.5	30.6	51.1	77.5	71.2	41.0	54.3
SAL	✓	✗	33.1	71.4	22.8	40.5	33.5	63.9	84.2	88.3	51.7	66.4
SAM+DBS+CLIP	✗	✗	8.2	56.4	18.6	0.6	7.5	11.5	47.6	0.0	17.3	11.2
SAL	✗	✗	24.8	66.8	17.4	30.2	28.7	48.5	78.8	80.4	32.6	52.8

performance (33.1 PQ, 40 mIoU) compared to the baseline, not trained with the semantic distillation loss  $\mathcal{L}_{token}$  (20.0 PQ, 23.4 mIoU). This confirms that by distilling CLIP features, we inject a notion of semantics into our model, even though our model is not explicitly supervised with any labels containing semantic information – the notion of semantics is learned implicitly via CLIP feature distillation. Without this loss, our zero-shot model (24.8 PQ, 29.7 mIoU) outperforms the linearly probed model tuned using GT semantic labels.

### 4.3 Class-Agnostic and Zero-Shot Lidar Panoptic Segmentation

As the first study tackling *Zero-Shot Lidar Panoptic Segmentation* (ZS-LPS), we devise several strong baselines inspired by prior works in RGB-D semantic segmentation. During inference, these methods utilize image-based models and lift image features to 3D [58]. To ensure these are evaluated fairly, we report results in Tab. 4 on the subset of points visible in at least one camera (*Frust. Eval.*). Our SAL model does not have this limitation. Therefore, we additionally report results evaluated on *full* point clouds in the bottom of Tab. 4.

**Class-agnostic segmentation.** Our first baseline (SAM [31]) generates masks in images and unprojects them to Lidar (see Sec. 3.3), and leads to 46 PQ. We observe bleeding edges (see Fig. 4b) after unprojecting from the cameras to Lidar caused by imperfect calibrations. To mitigate these artifacts, we experiment with slightly eroding SAM predictions in the image domain. However, this leads to a performance drop (42.2 PQ), likely because smaller segments after erosion fail to pass  $> 0.5$  overlap with labeled instances.

*Density-based filtering:* Alternatively, to mitigate unprojection errors, we experiment with density-based clustering methods. We generate a large pool of DBSCAN (DBS) clusters and test two strategies to improve our SAM-based segmentation labels: (i) *filtering* segments without a sufficient overlap with any DBSCAN cluster or (ii) *replacing* segments with large overlaps.

*SAM+DBS (filter):* Intuitively, the filter approach removed Lidar segments only segmentable in the image domain, *e.g.*, a SAM mask that segments a shadow on a flat wall. While this strategy improves precision for **things** classes (76.8 PQ<sup>Th</sup>, +27.1), it significantly degrades performance on **stuff** classes (24.8 PQ<sup>St</sup> −18.6) and overall degrades PQ to 46.7.

*SAM+DBS (replace):* The replace strategy (that replaces SAM masks with DBSCAN clusters with sufficient mutual overlap), on the other hand, improves both **things** and **stuff** classes to an overall PQ of 48.7 (+3.9) and and SQ 73.7 (+1.1), respectively. This improvement can be visually verified in Fig. 4c.

*SAL:* We train our SAL model using top-performing *SAM+DBS (replace)* and observe a significant improvement (70.7 PQ) in terms of both **things** and **stuff** classes. Even though SAL obtains significantly higher PQ than hand-crafted baselines, it also improves in terms of SQ (81.9 *vs.* 73.7 closest competitor), suggesting that the distilled model is robust to projection artifacts.

**Do we need a vision segmentation foundation model?** To generate a set with a sufficiently high recall, we need to cluster points with varying DBSCAN radius parameters (details in Appendix A.1). This leads to a large set of predicted masks per scan (5,413 avg.). While most of the *correct* segments are in this set, separating the signal from the noise is difficult. Prior works, such as [52], are limited to **moving thing** classes as they rely on motion cues to filter static point cloud regions and DBSCAN to group remaining points. Our core insight is that we can utilize vision foundation models that already learned a notion of objectness: SAM generates *only* 171 masks per image, reduced to 45 via flattening, and to 39 masks after the unprojection, achieving a high recall with *only a few instances*, or in other words, a high signal-to-noise ratio.

**Zero-shot segmentation.** So far, we discussed class-agnostic segmentation. To obtain semantic class predictions per mask in a zero-shot fashion, we devise a simple baseline inspired by [58]. Once masks are unprojected from the image to Lidar space, we utilize associated CLIP features (directly extracted from images using [16]) to perform zero-shot classification via dot product between encoded class prompts and (lifted) CLIP features [67]. *This baseline (SAM+DBS+CLIP) can be understood as an approach that directly lifts image features from state-of-the-art image-based models [16] to Lidar.* We detail prompts in Appendix A.3.

There is a significant performance drop between *SAM+DBS (replace)* baseline, evaluated with Semantic Oracle (5<sup>th</sup> row, 48.7 PQ), and zero-shot baseline (*SAM+DBS (replace)+CLIP*), 7<sup>th</sup> row, 27.5 PQ. This is not surprising, as zero-shot segmentation is a challenging problem in both image [16] and Lidar domains. To this end, even when transferring features of state-of-the-art image-based models to Lidar we observe a significant performance drop. Remarkably,

**Table 5: Lidar Panoptic Segmentation (LPS) on SemanticKITTI and nuScenes validation sets.** We prompt our zero-shot SAL model with the respective class vocabularies and compare its performance to fully-supervised baselines. On SemanticKITTI and nuScenes, we reach 42% and 54% of the fully-supervised model, respectively. This gap reduces significantly when we evaluate super classes.

Method	Supervision	Default classes						Super classes				
		PQ	RQ	SQ	PQ <sup>Th</sup>	PQ <sup>St</sup>	mIoU	PQ	RQ	SQ	mIoU	
SemanticKITTI	DS-Net [26]	Full	57.7	68.0	77.6	61.8	54.8	63.5	-	-	-	-
	PolarSeg [98]	Full	59.1	70.2	78.3	65.7	54.3	64.5	-	-	-	-
	EfficientLPS [74]	Full	59.2	69.8	75.0	58.0	60.9	64.9	-	-	-	-
	GP-S3Net [70]	Full	63.3	75.9	81.4	70.2	58.3	73.0	-	-	-	-
	MaskPLS [45]	Full	59.8	69.0	76.3	-	-	-	78.4	87.1	88.2	84.5
SAL	Full	59.5	69.2	75.7	62.3	57.4	63.8	81.7	90.0	89.2	85.9	
	Zero-shot	24.8	32.3	66.8	17.4	30.2	28.7	48.5	59.4	78.8	52.8	
nuScenes	PHNet [36]	Full	74.7	84.2	88.2	74.0	75.9	79.7	-	-	-	-
	DS-Net [26]	Full	51.2	59.0	86.1	38.4	72.3	73.5	-	-	-	-
	GP-S3Net [70]	Full	61.0	72.0	84.1	56.0	66.0	75.8	-	-	-	-
	EfficientLPS [74]	Full	62.0	73.9	83.4	56.8	70.6	65.6	-	-	-	-
	PolarSeg [98]	Full	63.4	75.3	83.9	59.2	70.4	66.9	-	-	-	-
	MaskPLS [45]	Full	57.7	66.0	71.8	64.4	52.2	62.5	71.5	81.0	86.2	80.6
SAL	Full	70.5	80.8	85.9	79.4	61.7	72.8	74.2	82.7	87.1	84.0	
	Zero-shot	38.4	47.8	77.2	47.5	29.2	33.9	52.6	63.5	77.3	52.6	

SAL distills such noisy image-based features to a stronger model (8<sup>th</sup> row, 33.1 PQ, +5.6 PQ). Finally, to emphasize the inherent limitation of baselines, relying on image features, we evaluate on full 360° point clouds. SAL performs similarly in both settings. The baseline, on the other hand, relies on image features and is, therefore, unable to predict labels out of the camera frustum. This problem can be mitigated by utilizing a setup with denser camera coverage.

*Evaluation on super-classes.* We observe that the CLIP classifier often confuses related classes, e.g., **car** and **other vehicle**. We additionally evaluate our semantics by prompting with super-classes (as defined in [6]). This alone increases PQ to 63.9 (48.5 when evaluating on full point clouds) and suggests significant potential for improving prompting and text-to-image alignment.

#### 4.4 Lidar Panoptic Segmentation Evaluation

Finally, we compare SAL with state-of-the-art LPS methods on standard benchmarks [6, 18]. We focus this discussion on comparing SAL, trained using labeled data, to our SAL zero-shot model. For the zero-shot evaluation, we specify classes, defined in respective dataset vocabularies [6, 18], as text prompts (we detail these prompts in Appendix C.1). Importantly, zero-shot results are obtained without training on *any* labeled data. *In contrast, all baselines are trained using human-labeled data and are constrained to their respective pre-defined class vocabularies.*

**SemanticKITTI.** When evaluating with the *default class* vocabulary (Tab. 5), our zero-shot model reaches 24.8 PQ or 42% of the fully-supervised model (59.5 PQ). As discussed in Sec. 4.2, our method is mainly limited by the quality of the generated CLIP features. In practice, our zero-shot semantic classifier mainly

confuses objects within super classes, *e.g.*, **car** *vs.* **other-vehicle**. Therefore, **without any retraining**, we prompt our zero-shot model with *super classes* and obtain 48.5 PQ, 59% of the supervised model (81.7 PQ).

**nuScenes.** We report our result in Tab. 5 and reach similar conclusions. SAL reaches 38.4 PQ on *default classes*, 54% of the supervised model (70.5 PQ). As with SemanticKITTI, results consistently improve when evaluated with *super-classes*. In this setting, zero-shot model yields 52.6 PQ. Note that SAL model, supervised with labeled data, significantly outperforms MaskPLS.

**Towards closing the gap.** Compared to SemanticKITTI (24.8 PQ, 42%), we observe a significantly smaller gap between the model trained on pseudo- and ground truth labels for nuScenes (38.4 PQ, 54%). Despite its sparse Lidar signal and more diverse input scenarios, SAL performs better on this challenging Lidar dataset. We attribute this to a larger pseudo-label coverage of nuScenes (48%, compared to 14% coverage on SemanticKITTI), and, therefore, effectively more labeled data, which is also consistent with our observations on SemanticKITTI w.r.t. increasing the amount of labeled data (Tab. 2). This trend suggests a clear path forward: *applying SAL to pseudo-label more data with zero annotation cost*.

**Beyond labeled datasets.** Datasets, such as SemanticKITTI and nuScenes, provide dense, point-level panoptic segmentation labels, which we use for analysis and the evaluation of our method on a small set of classes that were labeled in the respective dataset (14/11 **things/stuff** in SemanticKITTI, respectively, and 23/6 in nuScenes). To demonstrate the versatility of SAL, we additionally pseudo-label and train a model on Waymo Open [75] dataset and show qualitative results in Fig. 1 and Appendix E. This effort takes less than a week of pseudo-labeling and model training efforts (we provide details in Appendix B.2).

**The good, the (breaking) bad, and the ugly.** While we are thrilled about class-agnostic segmentation performance, our experimental analysis reveals a substantial gap between full supervision and our zero-shot model. Finally, at the moment, SAL is limited to training on specific datasets. As in fully-supervised Lidar perception, cross-sensor generalization remains a challenge for SAL. This could be addressed by utilizing temporal context, data scaling efforts and investigating sensor-agnostic backbone networks.

## 5 Conclusions

We proposed SAL, a method for *Zero-Shot Lidar Panoptic Segmentation*. The key SAL components are a pseudo-label engine that utilizes vision foundation models and a zero-shot model trained via self-supervision. While SAL is the first step in the direction of learning general Lidar segmentation models by distilling image-based foundation to Lidar, we believe we just scratched the surface and opened the door for training Lidar segmentation models without manual supervision.

**Acknowledgments.** This project was funded, in parts, by ERC Starting Grant DynAI (ERC-101043189). We are grateful to Žan Gojčič, Guillem Braso, Cristiano Saltori,

Sérgio Agostinho, and Jonas Schult for their feedback on the paper and their insightful comments. Special thanks to Maxim Maximov for his help on figures.

## References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2274–2282 (2012)
2. Agarwalla, A., Huang, X., Ziglar, J., Ferroni, F., Laura, L.T., Hays, J., Osep, A., Ramanan, D.: Lidar panoptic segmentation and tracking without bells and whistles. In: *Int. Conf. Intel. Rob. Sys.* (2023)
3. Aksoy, E.E., Baci, S., Cavdar, S.: Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving. In: *Intel. Veh. Symp.* (2020)
4. Aygün, M., Ošep, A., Weber, M., Maximov, M., Stachniss, C., Behley, J., Leal-Taixé, L.: 4d panoptic lidar segmentation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2021)
5. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: *Eur. Conf. Comput. Vis.* (2018)
6. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: *Int. Conf. Comput. Vis.* (2019)
7. Behley, J., Milioto, A., Stachniss, C.: A Benchmark for LiDAR-based Panoptic Segmentation based on KITTI. In: *Int. Conf. Rob. Automat.* (2021)
8. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. *Adv. Neural Inform. Process. Syst.* (2019)
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *Eur. Conf. Comput. Vis.* (2020)
10. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inform. Process. Syst.* (2020)
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2022)
12. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2019)
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2016)
14. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2017)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2009)
16. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with maskclip. In: *Int. Conf. Mach. Learn.* (2023)
17. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Rob. Sci. Sys.* (1996)
18. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Rob. Automat. Letters* (2021)
19. Gasperini, S., Mahani, M.A.N., Marcos-Ramiro, A., Navab, N., Tombari, F.: Panoster: End-to-end panoptic segmentation of lidar point clouds. *IEEE Rob. Automat. Letters* (2021)



20. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: *Eur. Conf. Comput. Vis.* (2022)
21. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921* (2021)
22. Harley, A.W., Zuo, Y., Wen, J., Mangal, A., Potdar, S., Chaudhry, R., Fragkiadaki, K.: Track, check, repeat: An em approach to unsupervised tracking. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2021)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2020)
24. Held, D., Guillory, D., Rebsamen, B., Thrun, S., Savarese, S.: A probabilistic framework for real-time 3d segmentation using spatial, temporal, and semantic cues. In: *Rob. Sci. Sys.* (2016)
25. Held, D., Levinson, J., Thrun, S., Savarese, S.: Combining 3d shape, color, and motion for robust anytime tracking. In: *Rob. Sci. Sys.* (2014)
26. Hong, F., Zhou, H., Zhu, X., Li, H., Liu, Z.: Lidar-based panoptic segmentation via dynamic shifting network. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2021)
27. Hu, P., Held, D., Ramanan, D.: Learning to optimally segment point clouds. *IEEE Robotics and Automation Letters* **5**(2), 875–882 (2020)
28. Hurtado, J.V., Mohan, R., Valada, A.: Mopt: Multi-object panoptic tracking. *arXiv preprint arXiv:2004.08189* (2020)
29. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2019)
30. Kirillov, A., He, K., Girshick, R.B., Rother, C., Dollár, P.: Panoptic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.* (2018)
31. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Int. Conf. Comput. Vis.* (2023)
32. Kreuzberg, L., Zulfikar, I.E., Mahadevan, S., Engelmann, F., Leibe, B.: 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In: *ECCV AVVision Workshop* (2022)
33. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2019)
34. Lee, S., Lim, H., Myung, H.: Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In: *Int. Conf. Intel. Rob. Sys.* (2022)
35. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: *Int. Conf. Learn. Represent.* (2022)
36. Li, J., He, X., Wen, Y., Gao, Y., Cheng, Y., Zhang, D.: Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2022)
37. Li, S., Chen, X., Liu, Y., Dai, D., Stachniss, C., Gall, J.: Multi-scale interaction for real-time lidar data segmentation on an embedded platform. *IEEE Rob. Automat. Letters* **7**(2), 738–745 (2021)
38. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2023)
39. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: *Eur. Conf. Comput. Vis.* (2014)

40. Lin, Z., Pathak, D., Wang, Y.X., Ramanan, D., Kong, S.: Continual learning with evolving class ontologies. *Adv. Neural Inform. Process. Syst.* (2022)
41. Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z.: Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347* (2023)
42. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: *Int. Conf. Comput. Vis.* (2021)
43. Lu, Y., Jiang, Q., Chen, R., Hou, Y., Zhu, X., Ma, Y.: See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In: *Int. Conf. Comput. Vis.* (2023)
44. Ma, Y., Peri, N., Wei, S., Hua, W., Ramanan, D., Li, Y., Kong, S.: Long-tailed 3d detection via 2d late fusion. *arXiv preprint arXiv:2312.10986* (2023)
45. Marcuzzi, R., Nunes, L., Wiesmann, L., Behley, J., Stachniss, C.: Mask-based panoptic lidar segmentation for autonomous driving. *IEEE Rob. Automat. Letters* **8**(2), 1141–1148 (2023)
46. Marcuzzi, R., Nunes, L., Wiesmann, L., Marks, E., Behley, J., Stachniss, C.: Mask4d: End-to-end mask-based 4d panoptic segmentation for lidar sequences. *IEEE Rob. Automat. Letters* (2023)
47. Marcuzzi, R., Nunes, L., Wiesmann, L., Vizzo, I., Behley, J., Stachniss, C.: Contrastive instance association for 4d panoptic segmentation using sequences of 3d lidar scans. *IEEE Rob. Automat. Letters* (2022)
48. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013)
49. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: RangeNet++: Fast and Accurate LiDAR Semantic Segmentation. In: *Int. Conf. Intel. Rob. Sys.* (2019)
50. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: *Int. Conf. Rob. Automat.* (2018)
51. Moosmann, F., Stiller, C.: Joint self-localization and tracking of generic objects in 3d range data. In: *Int. Conf. Rob. Automat.* (2013)
52. Najibi, M., Ji, J., Zhou, Y., Qi, C.R., Yan, X., Ettinger, S., Anguelov, D.: Motion inspired unsupervised perception and prediction in autonomous driving. In: *Eur. Conf. Comput. Vis.* (2022)
53. Najibi, M., Ji, J., Zhou, Y., Qi, C.R., Yan, X., Ettinger, S., Anguelov, D.: Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In: *Int. Conf. Comput. Vis.* (2023)
54. Nunes, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. *IEEE Rob. Automat. Letters* **7**(2), 2116–2123 (2022)
55. Ošep, A., Voigtlaender, P., Luiten, J., Breuers, S., Leibe, B.: Towards large-scale video video object mining. In: *ECCV Workshop on Interactive and Adaptive Learning in an Open World* (2018)
56. Ošep, A., Mehner, W., Voigtlaender, P., Leibe, B.: Track, then decide: Category-agnostic vision-based multi-object tracking. In: *Int. Conf. Rob. Automat.* (2018)
57. Ošep, A., Voigtlaender, P., Luiten, J., Breuers, S., Leibe, B.: Large-scale object mining for object discovery from unlabeled video. In: *Int. Conf. Rob. Automat.* (2019)
58. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: Openscene: 3d scene understanding with open vocabularies. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2023)
59. Peri, N., Dave, A., Ramanan, D., Kong, S.: Towards long-tailed 3d detection. In: *Conf. Rob. Learn.* (2023)

60. Peri, N., Li, M., Wilson, B., Wang, Y.X., Hays, J., Ramanan, D.: An empirical analysis of range for 3d object detection. In: ICCV Workshops (2023)
61. Peri, N., Luiten, J., Li, M., Ošep, A., Leal-Taixé, L., Ramanan, D.: Forecasting from lidar via future object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
62. Petrovskaya, A., Thrun, S.: Model based vehicle detection and tracking for autonomous urban driving. *Aut. Rob.* **26**, 123–139 (2009)
63. Pot, E., Toshev, A., Kosecka, J.: Self-supervisory signals for object discovery and detection. arXiv preprint arXiv:1806.03370 (2018)
64. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: IEEE Conf. Comput. Vis. Pattern Recog. (2012)
65. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
66. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Adv. Neural Inform. Process. Syst. (2017)
67. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (2021)
68. Rahman, S., Khan, S.H., Porikli, F.: Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *Asian Conf. Comput. Vis.* (2018)
69. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
70. Razani, R., Cheng, R., Li, E., Taghavi, E., Ren, Y., Bingbing, L.: Gp-s3net: Graph-based panoptic sparse semantic segmentation network. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
71. Razani, R., Cheng, R., Taghavi, E., Bingbing, L.: Lite-hdseg: Lidar semantic segmentation using lite harmonic dense convolutions. In: Int. Conf. Rob. Automat. (2021)
72. Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-lidar self-supervised distillation for autonomous driving data. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
73. Seidenschwarz, J., Ošep, A., Ferroni, F., Lucey, S., Leal-Taixé, L.: Semoli: What moves together belongs together. *IEEE Conf. Comput. Vis. Pattern Recog.* (2024)
74. Sirohi, K., Mohan, R., Büscher, D., Burgard, W., Valada, A.: Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics* (2021)
75. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
76. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023)
77. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: Eur. Conf. Comput. Vis. (2020)
78. Teichman, A., Levinson, J., Thrun, S.: Towards 3D object recognition via classification of arbitrary object tracks. In: Int. Conf. Rob. Automat. (2011)

79. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: *Int. Conf. Comput. Vis.* (2019)
80. Thorpe, C., Herbert, M., Kanade, T., Shafer, S.: Toward autonomous driving: the cmu navlab. i. perception. *IEEE expert* **6**(4), 31–42 (1991)
81. Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G.: Stanley: The robot that won the darpa grand challenge. *Journal of field Robotics* (2006)
82. Wang, Y., Chen, X., You, Y., Li, L., Hariharan, B., Campbell, M., Weinberger, K., Chao, W.: Train in germany, test in the usa: Making 3d object detectors generalize. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2020)
83. Wong, K., Wang, S., Ren, M., Liang, M., Urtasun, R.: Identifying unknown instances for autonomous driving. In: *Conference on Robot Learning*. pp. 384–393. PMLR (2020)
84. Wu, B., Wan, A., Yue, X., Keutzer, K.: Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In: *Int. Conf. Rob. Automat.* (2018)
85. Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K.: Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: *Int. Conf. Rob. Automat.* (2019)
86. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018)
87. Xiong, X., Munoz, D., Bagnell, J.A., Hebert, M.: 3-D Scene Analysis via Sequenced Predictions over Points and Regions. In: *Int. Conf. Rob. Automat.* pp. 2609–2616 (2011)
88. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2023)
89. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2023)
90. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. *Sensors* **18**(10), 3337 (2018)
91. Yilmaz, K., Schult, J., Nekrasov, A., Leibe, B.: Mask4d: Mask transformer for 4d panoptic segmentation. *arXiv preprint arXiv:2309.16133* (2023)
92. Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2021)
93. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2021)
94. Zhang, L., Yang, A.J., Xiong, Y., Casas, S., Yang, B., Ren, M., Urtasun, R.: Towards unsupervised object detection from lidar point clouds. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2023)
95. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2022)
96. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: *Eur. Conf. Comput. Vis.* (2022)
97. Zhou, Y., Tuzel, O.: Voxynet: End-to-end learning for point cloud based 3d object detection. In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2018)

98. Zhou, Z., Zhang, Y., Foroosh, H.: Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
99. Zhu, M., Han, S., Cai, H., Borse, S., Ghaffari, M., Porikli, F.: 4d panoptic segmentation as invariant and equivariant field prediction. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
100. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)

## Appendix

**Abstract.** The appendix provides additional implementation and design details of the SAL pseudo-label engine (A.1) and zero-shot model (A.2), along with a discussion on different training strategies with partial pseudo-labels (A.4). In A.3, we discuss the zero-shot text prompt engineering. Further ablations on our pseudo-labels and model are provided in B and C, respectively, including fine-grained, per-class results (D) and detailed statistics on the generated pseudo-labels (B.1) and time and compute analysis for pseudo-labeling (B.2). We conclude with a discussion on detailed qualitative results in E.

### A Implementation Details

#### A.1 Pseudo-label Engine

This section details the overview of our pseudo-labeling procedure (Sec. 3.3 of the main paper). We detail our pseudo-label generation steps in Algorithm 1. As input, we assume a sequence of multi-modal data that consists of  $k$  images  $\mathcal{I}_{t,k} \in \mathbb{R}^{W \times H \times 3}$  (one per camera view, per frame  $t, t \in 1, \dots, T$ ), Lidar point clouds  $P_t \in \mathbb{R}^{N \times 4}$ , and image and camera-to-Lidar calibration. The output is pairs of Lidar segmentation masks  $\tilde{m}_t \in \{0, 1\}^N$  and corresponding CLIP feature tokens  $f_t \in \mathbb{R}^{768}$ ,  $t \in 1, \dots, T$ .

**Mask generation.** We start by generating an overlapping set of segmentation masks for each image (camera view) (Algorithm 1, L6) using *Segment Anything* (SAM) foundation model [31]. To generate masks, we utilize parameters, as described in Tab. A.1. This step yields 171 masks per image on average. We then flatten SAM’s output mask hierarchy by non-maxima suppression (NMS) with a small overlap threshold (Algorithm 1, L7, govern by *NMS IoU threshold* in Tab. A.1) to obtain a small, non-overlapping set of masks per-scan (45 on average after flattening). This way, we obtain a set of non-overlapping binary masks  $m_t^k \in \{0, 1\}^{W \times H}$  for each camera view  $k$  with an image plane of size  $W \times H$ .

In practice, during mask suppression, we sort masks based on the mask area (rather than score, usually done in NMS) – this criterion favors objects over their parts and subparts. As can be seen in Tab. B.3, this approach (*NMS area*) performs significantly better as compared to objectness score-based suppression (*NMS score*), which may favor object parts over objects. We visualize the flattened segmentation masks in images in Fig. E.3 (SemanticKITTI) and Fig. E.4 (nuScenes).

**CLIP image token generation.** We proceed by generating localized CLIP [67] image feature tokens  $f_t^k \in \mathbb{R}^{768}$  for each binary mask  $m_t^k$  (Algorithm 1, L8). To obtain image tokens for a masked region of the input image, we utilize MaskCLIP [16] and their relative mask attention in the CLIP image encoder feature space. The original MaskCLIP pipeline forwards the entire image and all masks at once. For our use case, we observed better per-mask classification results by generating image tokens for each mask separately. To this end, we forward individual image crops. In Fig. E.3 and Fig. E.4, we visualize for each generated mask in the image the most-likely class, assigned using the generated CLIP token (according to the class vocabularies, as detailed in Tab. E.1). Importantly, we report class names only for visualization purposes



in Fig. E.3, Fig. E.4, and Fig. E.5. We **never** classify segments in the image domain but merely distill CLIP image feature tokens to Lidar to facilitate zero-shot classification during inference where we do not use any image features.

**Image-to-Lidar unprojection.** From the Lidar perspective, we unproject each image mask  $m_{t,i}^k \in \{0, 1\}^{W \times H}$  to a binary Lidar segmentation mask  $\tilde{m}_{t,i}^k \in \{0, 1\}^N$  by transforming the respective camera coordinate frame to the Lidar space (Algorithm 1, L9–L10). The unprojection yields pairs  $\{\tilde{m}_t^k, f_t^k\}$  of Lidar masks and their corresponding CLIP features w.r.t. camera  $k$ . On average, from 45 masks generated in images, 39 are transferred to the Lidar domain – the rest are either too small or not supported by Lidar measurements (due to signal sparsity or lack of Lidar coverage). We visualize masks, unprojected from a single camera to Lidar in the middle row of Fig. E.6 (SemanticKITTI).

We then insert these masks to the output sets  $\tilde{m}_t, f_t$  as follows. To all masks in  $\tilde{m}_t^k$  that do not overlap significantly overlap with masks in  $\tilde{m}_t$  (Tab. A.1, *multi-view IoU threshold*), we assign a new ID, and insert masks and their corresponding CLIP features to the output sets  $\tilde{m}_t, f_t$ . For masks whose overlap threshold exceeds this limit, we update existing masks with a union of the two, and the feature with the average. We visualize masks, unprojected to Lidar from several views, followed by fusion, in the middle row of Fig. E.7 (nuScenes).

**Refinement via clustering.** Finally, we improve our pseudo-label quality by creating an ensemble of DBSCAN [17] clusters  $\tilde{m}_t^{DBSCAN} \in \{0, 1\}^N$ , obtained by varying the density thresholds (Tab. A.1, *DBSCAN density thresholds*) to compensate for varying density in Lidar point clouds. We replace each  $\tilde{m}_i \in \tilde{m}_t$  with its best-matching  $\tilde{m}_l \in \tilde{m}_t^{DBSCAN}$  in case their IoU exceeds a minimal overlap threshold (Tab. A.1, *DBSCAN IoU overlap threshold*) and retain the original mask otherwise to obtain a refined set of pseudo-labels that retain their original cardinality and associated CLIP features.

**DBSCAN clustering implementation.** To create the ensemble of DBSCAN clusters  $\tilde{m}_t^{DBSCAN}$ , we first perform geometric plane fitting and remove ground points (to estimate the ground points, we use [34] and its publicly-available implementation). We then perform DBSCAN on the ground-filtered point clouds using six density thresholds (reported in Tab. A.1, *DBSCAN IoU overlap threshold*). This leads to a large set of (overlapping) segmentation masks, induced by the estimated point clusters (in SemanticKITTI, 5,413 on average per scan)

In Tab. B.3, we report two alternatives. Firstly, rather than using DBSCAN, we perform erosion in the image domain to minimize the “edge bleeding” artifacts. As can be seen, this variant decreases the PQ score (46.0 to 42.2). Second, rather than replacing segments with DBSCAN (*DBSCAN replace*), we filter the pool of SAM-generated segments, only retaining those that have a sufficient overlap (in terms of intersection-over-union), (*DBSCAN filter*). While this variant improves PQ for **things** classes, it significantly reduces PQ for **stuff** classes, leading to overall lower performance compared to the replace variant (46.7 PQ for *filter vs.* 48.7 PQ for *replace*).

**Parameter tuning.** We tune all hyperparameters on a subset of 40 Lidar scans of SemanticKITTI validation set (we sample every 100<sup>th</sup> scan). We perform no parameter tuning for nuScenes.

---

**Algorithm 1** Pseudo-label Engine

---

**Require:** Lidar point clouds  $P_t$ ,  $k$  camera views  $\mathcal{I}_{t,k}$ ,  $k$  camera calibrations  $c_k$ ,  $t \in 1, \dots, T$

**Ensure:**  $\{\tilde{m}_t, f_t\}, t \in 1, \dots, T$  // Lidar segmentation masks and corresponding CLIP image feature tokens.

- 1: **for** each timestamp  $t$  **do**
- 2:      $P_t \leftarrow \text{load\_lidar}(t)$
- 3:      $\tilde{m}_t = \emptyset, f_t = \emptyset$
- 4:      $\tilde{m}_t^{DBSCAN} \leftarrow \text{DBSCAN\_ensemble}(P_t)$  // Generate (overlapping) Lidar mask ensemble
- 5:     **for** each camera  $k$  **do**
- 6:          $\mathcal{I}_{t,k} \leftarrow \text{load\_image}(t, k)$
- 7:          $m_t^k \leftarrow \text{SAM}(\mathcal{I}_{t,k})$  // Generate masks in image  $k$
- 8:          $m_t^k \leftarrow \text{NMS}(m_t^k)$  // Apply Non-Maximum Suppression (NMS) to masks
- 9:          $f_t^k \leftarrow \text{MaskCLIP}(\mathcal{I}_{t,k}, m_t^k)$  // Obtain localized CLIP features for each mask
- 10:         $\tilde{m}_t^k \leftarrow \text{label\_point\_cloud}(P_{t,k}, m_t^k, c_k)$  // Generate segmentation pseudo-labels w.r.t. camera  $k$
- 11:         $\tilde{m}_t \leftarrow \text{replace}(\tilde{m}_t, \tilde{m}_t^{DBSCAN})$  // Replace with sufficiently overlapping DBSCAN masks
- 12:         $\{f_t, f_t^k\} \leftarrow \text{insert\_or\_merge}(\tilde{m}_t, f_t, \tilde{m}_t^k, f_t^k)$  // Cross-camera fusion: Insert new or merge with existing
- 13:     **end for**
- 14: **end for**

---

## A.2 Zero-shot Model

We use our SAL model, trained using labels generated by the pseudo-label engine (Appendix A.1), to segment and classify *any* object. The model implements a Transformer decoder architecture and adds a CLIP token prediction head (see Fig. A.1). In contrast to MaskPLS [45], our decoder operates not in point but voxel space and only on a single backbone feature scale. These adaptations obtained empirically better results from an overall more lightweight model. As it is common for Transformer decoders, the computation of its loss  $\mathcal{L}_{\text{SAL}}$  relies on a bipartite cost matching between predictions and (pseudo-)labels. Empirically, we obtain better results if this matching only relies on the class-agnostic segmentation and does not incorporate the image token prediction as an additional cost. Unlike [45], we train all models from scratch and do not initialize our backbone with SegContrast [54]. In Tab. A.1, we highlight hyperparameter choices that deviate from the default setup in [45].

The *overlap threshold* filters segments during inference. Since the segmentation head predicts binary masks per query, the resulting segments are potentially overlapping. To flatten the output and obtain panoptic segmentation results, overlapping points are assigned based on probability. If a percentage of points are assigned to another segment, the overlap threshold removes the segment entirely from the output. Training on partial and noisy pseudo-labels increases the amount of overlap during inference, in particular, for areas that are out of the camera frustum(s) and hence not labeled in the training data. Deactivating the overlap filter entirely yields the best results for our SAL model.

**Table A.1: SAL hyperparameters.** We show parameters for both components of our framework: (i) SAM model [31], which we use to generate segmentation masks in images; (ii) the pseudo-label generation engine and (iii) our zero-shot model. For the latter, we only highlight parameters that deviate from [45].

Parameter	Value
SAM [31]	
Model	sam_vit_h_4b8939
Inference POINTS_PER_SIDE	32
Inference PRED_IOU_THRESH	0.84
Inference STABILITY_SCORE_THRESH	0.86
Inference MIN_MASK_REGION_AREA	100
Pseudo-label engine	
NMS IoU threshold	0.01
Multi-view IoU threshold	0.01
DBSCAN IoU overlap threshold	0.5
DBSCAN density thresholds	(1.2488, 0.8136, 0.6952, 0.594, 0.4353, 0.3221)
Zero-shot model	
GPUs	8 × 32GB (V100)
Batch size	24 (3 per GPU)
Learning rate (LR)	0.0003
Number of epochs	30
LR drop	15
Number of queries	300
Overlap threshold	0.0
Loss weights	2.0, 5.0, 5.0, 5.0, 2.0

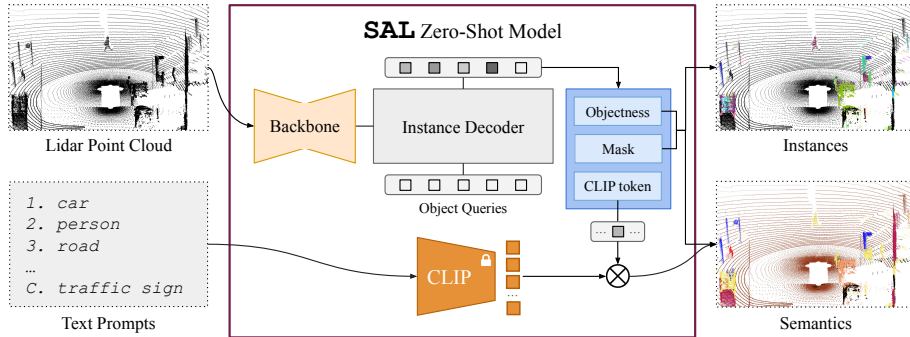
The *loss weights* balance each component of our full model loss:

$$\begin{aligned} \mathcal{L}_{\text{SAL}} = & \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{dice}} \\ & + \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{token\_aux}}. \end{aligned} \quad (2)$$

The  $\mathcal{L}_{\text{token\_aux}}$  is an auxiliary segmentation loss as applied in [45]. This loss computes the semantic segmentation quality not for the decoder queries but based on the backbone features alone. To this end, a per-point semantic segmentation head is added after the backbone. In contrast to [45], our head regresses a token per-point and not class per point. This auxiliary head is only added for training and discarded during inference. All of our trainings apply common spatial augmentations, including random rotations, flipping, scaling, and translations.

### A.3 Text Prompt Engineering

To perform zero-shot classification within a pre-defined class vocabulary, we complete and enrich the otherwise ambiguous and uninformative class names. As shown in Tab. E.1, each class is predictable not only by its own name but a set of additional prompts. In particular, all **other-X** classes are ambiguous prompts. The otherness only works in a fully supervised setting where a model can learn to predict, for example, all vehicles except the types already covered by other classes. In our case, we must directly prompt for other vehicle types, such as **trailer**, **bus**, **tram**, or **train**. This problem could also be solved by adding negations/exclusions to prompts. However, this approach did not yield the desired outcome in our experiments. We apply the same set of super classes for SemanticKITTI [6] and nuScenes [18] and merge their



**Fig. A.1: SAL zero-shot model.** Our model takes Lidar point clouds and text prompts as inputs. Its architecture relies on a 3D sparse-convolutional Minkowski backbone [12] followed by a Transformer decoder for object instance segmentation. The decoder computes cross-attention between object queries and backbone features. Three task heads predict objectness scores, segmentation masks, and CLIP tokens for each query. Once trained, we forward the dataset class vocabulary through the CLIP text encoder and perform zero-shot classification via matching with predicted CLIP tokens. The model requires no retraining for different vocabularies.

respective default class prompts. Furthermore, we follow [67] and wrap every prompt into a list of full-sentence templates. This results in text prompts like a **photo of a car**, which better align with the image caption training data of the CLIP text encoder. In Appendix C.1, we show additional ablations on the aforementioned text prompt engineering.

For panoptic segmentation outputs, every point is classified to one dataset vocabulary class. To avoid classifying all segments to the same class when given a single text prompt, we append a second background prompt to the prompt set. More specifically, all predicted CLIP tokens are prompted with the target text and the word **other**. Empirically, we observed that this broad term reliably matches to all objects unless the actual text prompt, *e.g.*, **fire hydrant**, is being segmented.

#### A.4 How to Train on Partial Labels?

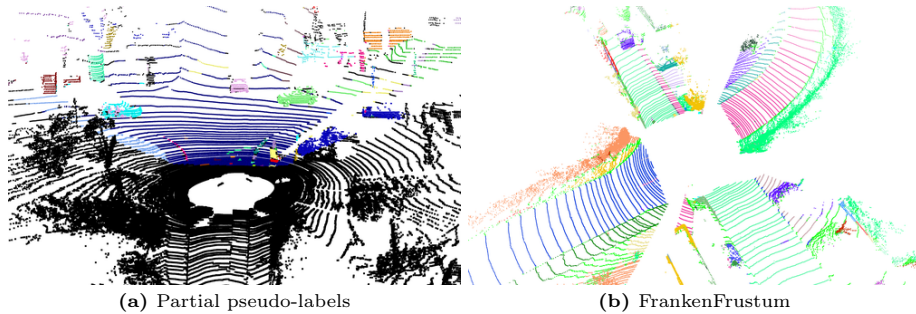
Pseudo-labels provide only partial supervision within the camera frustum (see Fig. A.2a), leaving the majority of Lidar points unlabeled (see Tab. B.1). How can we train SAL with such partial supervision?

**Ignore unlabeled region + standard data augmentations.** During training, we remove all unlabeled points from the point cloud. Otherwise, the  $\mathcal{L}_{obj}$  loss would penalize any segmentation in these regions as a false positive, thereby teaching the model to ignore them entirely. As visualized in Fig. A.2a and quantified in Tab. B.1, the 360° Lidar label coverage is particularly low for single camera setups, as in SemanticKITTI [6].

**FrankenFrustum.** To generalize to full clouds during inference, we propose a simple but very effective *FrankenFrustum* augmentation (Fig. A.2b). It mimics full point clouds during training by randomly removing unlabeled points and replicating labeled frustum regions around the vertical axis. This augmentation does not increase the over-

**Table A.2: CLIP token distillation and text prompt engineering.** To evaluate our token prediction, we prompt the SemanticKITTI class vocabulary to generate labeled training data and train a non-zero-shot model (row 1). Furthermore, we demonstrate the insufficiency of vanilla class names (`car`) as text prompts and the boost from engineering a rich set of terms (`car`, `jeep`, `SUV`, `van`) as explained in Appendix A.3.

Text prompt engineering	$\mathcal{L}_{token}$	PQ	RQ	SQ	mIoU
Default classes					
×	×	25.1	33.5	68.4	25.9
		20.6	27.1	65.2	20.9
×	×	24.8	32.3	66.8	29.7
Super classes					
	×	27.4	33.8	71.8	24.9
×	×	48.5	59.4	78.8	52.8



**Fig. A.2: Training on partial labels.** Unprojecting image-based pseudo labels results in a partially (pseudo) labeled point cloud (Fig. A.2a). We construct supervisory signal by concatenating multiple partially labeled point clouds (Fig. A.2b).

all label coverage. However, our ablations in the main paper (Sec. 4.2, Tab. 1) show its effectiveness in reducing the domain gap between training and inference input clouds.

## B Pseudo-label Analysis

### B.1 Pseudo-label Statistics

**Point coverage.** As shown in Tab. B.1, the single-camera setup of SemanticKITTI [6] allows us to label a significantly smaller portion of Lidar point clouds (14%) compared to GT labels (98%). Interestingly, even for nuScenes dataset [18], which provides a setup with five cameras and 360° view coverage, pseudo-labels cover only 48% of all points. Furthermore, we quantify pseudo-label coverage on SemanticKITTI when filtering all points outside of the camera frustum. This leads to coverage of 89%. The remaining missing labels can be explained by errors committed by our image-based segmentation foundation model, SAM [31]. *This analysis confirms that even when utilizing strong foundation models, image-to-Lidar distillation remains a challenging problem.*

**Things vs. stuff.** Furthermore, the *segment anything* philosophy transferred from SAM [31] to our pseudo-labels yields a significantly larger number of instances per

**Table B.1: Pseudo-label statistics.** We outline the label coverage of point clouds, the total, max, and mean number of instances per scan, and the ratio of things/stuff instances on the full point cloud and point cloud areas that overlap with the camera view frustum (Filter Frustum). As can be seen, due to the single-camera setup, pseudo-label coverage in SemanticKITTI [6] is very low (14% of points). Even though nuScenes [18] dataset provides 360° view coverage, only 48% are labeled due to blind spots. Even when only retaining points, that overlap with the camera view frustum (SemanticKITTI, Filter Frustum), we observe coverage of 89%. This can be explained by mistakes (*e.g.*, false negatives) committed by our segmentation foundation model (SAM [31]).

Label	Filter Frustum	Label coverage	Instances					
			Total	Max	Mean	Things	Stuff	$\frac{\#things}{\#stuff}$
<i>SemanticKITTI [6]</i>								
GT		98%	372478	65	19	54%	46%	0.84
Pseudo		14%	767450	122	40	13%	87%	0.15
GT	×	98%	224104	39	11	32%	68%	0.47
Pseudo	×	89%	756203	120	39	13%	87%	0.15
<i>nuScenes [18]</i>								
GT		70%	818971	119	29	63%	37%	1.70
Pseudo		48%	5594800	647	198	15%	84%	0.18

scan (19 GT *vs.* 40 pseudo on average). This transfer is particularly notable in the shift of ratios between **things** and **stuff** instances (0.84 *vs.* 0.14 for GT and pseudo-labels, respectively), as existing datasets (*e.g.*, SemanticKITTI and nuScenes) merge individual instances of classes such as **pole** or **trees** into single instances. *By contrast, our pseudo-labels provide a finer-grained segmentation of object instances, as needed for learning to segment a variety of objects.*

**Fine-grained analysis.** In Tab. B.2, we additionally report per-class label statistics for train and validation splits for SemanticKITTI dataset [6]. As can be seen, the label coverage is consistent in train and val splits (97% GT *vs.* 14% pseudo-labels in val, and 98% GT *vs.* 15% pseudo-labels in train). Overall, the max. number of instances is larger in the train set (65 GT & 122 pseudo) as compared to the validation set (53 GT & 82 pseudo), whereas the average number of instances remains consistent for pseudo-labels, while for GT labels are lower in the train (19 GT & 40 pseudo) compared to val (24 GT & 41 pseudo). The larger number of instances reflect the most frequent classes: on the train set (Filter Frustum), the highest percentage of instances are due to **vegetation** (27.1%), **building** (16.9%), and **road** (13.4%) classes. As can be confirmed in Fig. E.3 and Fig. E.4, our pseudo-labels indeed often localize individual trees, bushes, and buildings, leading to a large number of overall instances for these classes.

## B.2 Pseudo-label Timing & Effort Analysis

We report information and statistics on three datasets (SemanticKITTI [6], Panoptic nuScenes [18] and Waymo Open [75]) that we pseudo-label and use to train SAL, in Tab. B.5. As can be seen, the processing time needed to process a single scan depends on multiple dataset characteristics, such as the number of cameras. In the case of the Waymo Open dataset, we pseudo-label every 10<sup>th</sup> scan, and we do not perform



**Table B.2: Pseudo-label statistics per class on SemanticKITTI.** We compare ground truth label and pseudo-label statistics on SemanticKITTI [6] train and validation sets. Label coverage and instance class distributions are reported in percentage. Due to camera visibility coverage, pseudo-labels cover a significantly smaller portion of the dataset than GT labels. This is especially prominent in the SemanticKITTI dataset (14% coverage). Pseudo-labels cover a significantly larger number of instances per scan than GT labels (3× more when only compared in the *camera view frustum*), as needed to learn to segment a large variety of objects. While GT labels treat stuff classes as a “single instance,” our pseudo-labels hypothesize a variety of plausible segmentations of **stuff** classes, leading to a higher percentage of **stuff** class labels. We use a semantic oracle to report per-class statistics on pseudo-labels

Label	Coverage (%)	Max / Avg. Inst.	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	
<i>Full point cloud</i>																						
GT	97	53/24	46.9	4.2	1.3	0.3	3.2	4.5	1.5	0.2	4.2	1.4	4.2	0.9	4.2	3.9	4.2	4.0	4.2	4.2	2.9	
Pseudo	14	82/41	15.6	0.2	0.2	0.3	1.0	0.6	0.5	0.0	12.5	0.8	9.0	0.2	15.1	2.2	27.6	2.6	9.1	2.0	0.6	
Val	<i>Camera view frustum</i>																					
	GT	97	27/13	31.4	2.2	0.8	0.2	1.9	2.8	1.3	0.1	7.3	1.9	7.3	0.5	6.7	5.2	7.3	6.3	6.9	6.8	3.1
Pseudo	88	81/41	15.6	0.2	0.2	0.3	1.0	0.6	0.5	0.0	12.5	0.8	9.0	0.2	15.0	2.2	27.6	2.6	9.1	2.0	0.6	
Train	<i>Full point cloud</i>																					
	GT	98	65/19	44.8	2.0	1.0	0.7	2.1	2.5	0.5	0.1	5.1	2.1	4.8	1.4	4.6	5.0	5.1	4.6	5.0	5.0	3.5
	Pseudo	15	122/40	11.4	0.1	0.1	0.5	0.6	0.2	0.1	0.1	13.4	1.1	8.8	0.5	16.8	9.3	27.1	2.1	5.3	1.9	0.6
	<i>Camera view frustum</i>																					
GT	98	39/11	27.4	1.0	0.6	0.5	1.1	1.3	0.3	0.2	8.5	2.6	8.0	1.2	6.8	7.7	8.5	5.8	7.4	7.7	3.2	
Pseudo	89	120/39	11.4	0.1	0.1	0.5	0.6	0.2	0.1	0.1	13.4	1.1	8.8	0.5	16.9	9.3	27.1	2.1	5.3	1.9	0.6	

DBSCAN refinement. The reason is two-fold: (i) we observe image-lidar calibration on Waymo is more accurate compared to KITTI and nuScenes, and (ii) we save processing effort/time to pseudo-label Waymo. Importantly, our pseudo-labeling setup is general: it supports various sensory setups (single camera, multi-camera) and multiple Lidar types and can cope well with datasets with different degrees of accuracy of the image-lidar calibration/synchronization.

In Tab. B.6, we provide fine-grained per-scan timing analysis for SAL pseudo-labeling engine. The most costly component is extracting image-level segmentation masks using the segmentation foundation model (SAM [31]), which needs to be performed, in general, for each camera.

### B.3 Qualitative Analysis

Both SemanticKITTI [6] and nuScenes [18] do not provide ground truth annotations in 2D. To illustrate the performance of the foundation models we utilize for pseudo-labeling in the image domain, we visualize predicted masks and dataset class vocabularies from SAM [31] and CLIP [67]. In Fig. E.3, we show the single (front) camera view corresponding to the first scan of four different SemanticKITTI sequences. For

**Table B.3: Segmentation pseudo-labels analysis.** This table extends Tab. 4 from the main paper with additional metrics and baselines. We report results on the SemanticKITTI validation set using the semantic oracle, as detailed in the main paper (Sec. 4.1). We evaluate the region of the point cloud visible in the camera (Filter Frustum).

Model	PQ	RQ	SQ	IoU	PQ <sup>Th</sup>	RQ <sup>Th</sup>	SQ <sup>Th</sup>	PQ <sup>St</sup>	RQ <sup>St</sup>	SQ <sup>St</sup>
SAM (NMS area)	46.0	62.3	72.1	58.9	49.7	66.2	73.1	43.4	59.5	71.4
SAM (NMS score)	28.5	39.3	71.9	45.8	19.4	26.4	73.8	35.0	48.6	70.5
SAM+Erosion	42.2	58.6	69.4	55.9	45.6	62.3	70.0	39.6	55.9	69.0
SAM+DBS (filter)	46.7	56.3	70.3	45.8	76.8	85.5	89.4	24.8	35.1	56.4
SAM+DBS (replace)	48.7	64.8	73.7	59.8	53.1	69.2	75.1	45.4	61.6	72.7
SAL	70.7	85.6	81.9	79.7	75.4	87.1	86.4	67.3	84.4	78.7

**Table B.4: Segmentation pseudo-labels ablation - per class.** This table extends Tab. 4 from the main paper with per-class PQ scores. We report results on the SemanticKITTI validation set using the semantic oracle, as detailed in the main paper (Sec. 4.1). We evaluate the region of the point cloud visible in the camera (Filter Frustum).

Model	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
SAM (NMS area)	74.2	13.4	44.0	77.2	66.3	61.9	39.1	21.5	83.1	24.7	38.4	27.9	58.3	34.5	71.8	26.2	40.9	21.1	49.9
SAM (NMS score)	26.2	9.5	17.4	15.8	29.1	24.4	13.2	20.0	56.2	32.5	34.8	14.9	13.7	28.0	57.6	33.9	35.6	16.3	62.0
SAM+Erosion	72.2	5.6	45.0	78.4	65.4	55.1	22.9	20.3	82.4	23.1	35.9	24.3	54.4	31.1	68.8	22.0	37.3	13.4	43.3
SAM+DBS (filter)	84.9	56.7	78.5	90.4	78.8	75.0	85.9	64.1	26.9	0.0	0.0	0.8	51.3	23.3	43.6	37.3	1.1	32.2	56.5
SAM+DBS (replace)	75.8	17.5	50.0	78.2	67.3	66.9	42.9	26.4	83.1	24.7	38.4	28.0	62.2	36.2	71.5	33.5	40.9	27.0	54.4
SAL	90.7	65.8	79.9	59.5	78.1	89.8	65.4	74.1	92.8	24.8	74.2	47.4	82.4	59.5	85.0	73.3	60.7	75.3	65.5

nuScenes Fig. E.4 and Waymo Open, we focus on the first scan of a single sequence and show all camera views. As can be seen, in all datasets, the output contains many correct segmentations and classifications, *e.g.*, for the **road**, **car**, or **vegetation** classes. Moreover, the class-agnostic SAM masks introduce many smaller instances by segmenting individual road markings. However, without any corresponding points in the cloud, such segments are filtered by the unprojection/transfer to Lidar.

While the SAM masks are generally correct, noisy CLIP predictions confirm there is room for improvement. For example, the sky is always misclassified since it is not part of the class vocabulary. As shown in the main paper (Tab. 3), the distillation of SAM and CLIP to 3D yields an analogous behavior of our SAL model. It should be noted that the clip predictions shown in Fig. E.3 and Fig. E.4 are only for visualization purposes. We do not directly transfer class labels to 3D, but only the CLIP embeddings.

## B.4 Pseudo-label Quality

We additionally report per-class pseudo-label results in Tab. B.4. As can be seen, *DBSCAN replace* consistently improves over tuned SAM variant on **things** classes (*e.g.*, for **car** class, +1.6 PQ, **person** class +4.8 PQ). While *DBSCAN filter* leads to remarkable improvements on some classes (*e.g.*, for **car** +10.7, **bicycle** +43.2),

**Table B.5: Dataset pseudo-labeling analysis.** We report dataset information and statistics, along with pseudo-labeling effort and time analysis. We note that in SemanticKITTI, we have four cameras in total. However, only one is used for pseudo-labeling, as all cameras have (roughly) the same field of view. We report the total number of scans, as well as per-scan processing time (NVIDIA A100D-80C GPU). Processing time depends on the number of cameras, as well as individual dataset characteristics (such as camera resolution, point cloud size, density, *etc.*). Finally, we report the total time needed to pseudo-label a dataset if processed sequentially (in practice, we can pseudo-label datasets in 1-3 days using a compute cluster). We note that in the case of Waymo, due to sheer size, we (i) pseudo-label only 10% of point clouds and (ii) skip the postprocessing with DBSCAN.

Dataset	# cam.	Cam. cov. (°)	Lidar	# scans		Time (s)	Time (days)
				Total / Pseudo-lab.	Per scan	Total	
SemanticKITTI	1	90	Velodyne HDL-64E	43592/43592	32	16	
Panoptic nuScenes	6	360	Velodyne HDL-32E	40157/40157	129	59	
Waymo Open	5	270	Waymo Proprietary	227101/22710	92	24	

**Table B.6: Per-scan pseudo-labeling timing analysis.** We report fine-grained per-dataset timing analysis for our pseudo-labeling system. In particular, we report individual timings for core components of **SAL** pseudo-label engine: (i) running segmentation foundation model (SAM [31]) in the image domain, (ii) extracting corresponding CLIP features [16, 67], (iii) image-to-Lidar unprojection, and finally, (iv) DBSCAN refinement. As can be seen, the most costly step is due to SAM; this step could be reduced in the future using newer and more time-efficient variants of SAM. DBSCAN is the second bottleneck, however, it is only executed once per point cloud, and does not depend on the number of cameras. The analysis was conducted using NVIDIA A100D-80C GPU.

Dataset	# cam.	Time (s)				
		Scan	SAM	CLIP	Unprojection	DBSCAN
SemanticKITTI	1	31.9	6.7	1.8	0.9	22.2
Panoptic nuScenes	6	129.2	56.4	14.8	14.34	41.1
Waymo Open	5	92	56.1	16.0	17.5	–

it severely degrades stuff classes (*e.g.*,  $-39.8$  for **terrain** class). On the other hand, *DBSCAN replace* either does not affect **stuff** classes or leads to improvements (*e.g.*, **building**  $+3.9$ ).

## C Model Ablations

### C.1 Zero-shot Classification

The **SAL** model performs zero-shot classification by matching text prompts with predicted CLIP tokens. To predict tokens only with Lidar input, our label engine generates pairs of Lidar segments and corresponding CLIP image tokens. Training on these pairs distills the CLIP image encoder into our model. The first row of Tab. A.2 demonstrates a different approach where we train on pairs of Lidar segments and class labels obtained by directly prompting the CLIP image encoder. Examples of these labels are visualized in Fig. E.3. The difference ( $-1.5$ ) between the first and third row in Tab. A.2 indicates the performance drop of the distillation into our model. This further demonstrates the

**Table C.1: SAL per class PQ results on the SemanticKITTI and nuScenes validation sets.** We report metrics for zero-shot (ZS) and linear probing (LP) supervision.

Supervision		SemanticKITTI [6]																		
		all	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole
ZS	24.8	82.3	22.3	10.9	9.5	9.2	5.4	0.0	0.0	67.0	0.2	27.6	0.1	33.2	3.6	80.7	16.9	32.4	45.8	24.9
LP	33.1	78.0	1.5	25.0	25.8	20.3	41.5	62.6	0.0	79.7	17.6	33.2	0.0	74.2	13.3	74.0	21.5	34.4	32.9	9.8

Supervision		nuScenes [18]															
		all	barrier	bicycle	bus	car	construction_vehicle	motorcycle	pedestrian	traffic_cone	trailer	truck	driveable_surface	other_flat	sidewalk	terrain	manmade
ZS	38.4	0.7	53.4	46.6	82.2	17.9	52.5	50.6	35.4	30.0	47.2	57.6	0.1	13.5	29.2	29.7	67.5
LP	40.1	3.2	17.6	53.8	77.3	15.8	40.8	83.2	27.2	28.2	47.7	67.1	23.1	10.2	15.8	64.2	65.8

effectiveness of our framework. In particular, our insufficiencies with respect to zero-shot classification are not caused by our design choices in 3D but merely a reflection of the same limitations in the image domain.

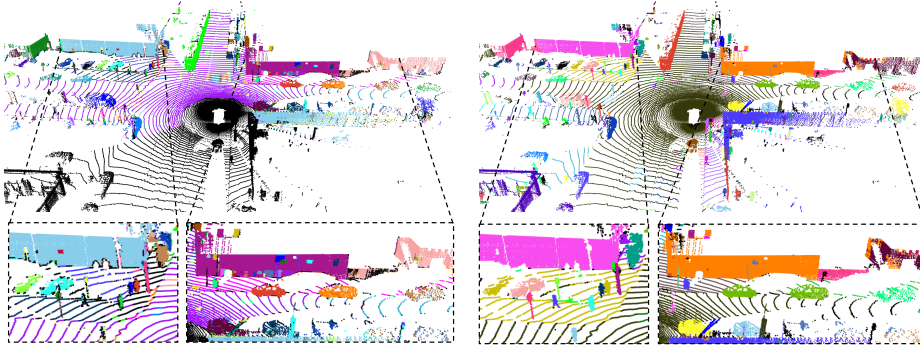
**Text prompt engineering.** The CLIP tokens predicted by our model can be prompted with any arbitrary text. This allows our model to classify potentially *any* object. The text prompts must adhere to the respective class vocabulary to perform zero-shot classification on an annotated dataset. In Tab. A.2, we demonstrate the performance boost from engineering the set of text prompts, *e.g.*, by adding similar class terms to each class set (see Tab. E.1). In particular, for super-classes, where we add all default class names to the respective super set, our engineering yields a huge gain of +23.1. This is due to the ambiguous nature of the super class names, *e.g.*, **object** or **structure**.

## D Per-class Results

In Tab. C.1, we report per-class PQ results of the SAL model on SemanticKITTI [6] and nuScenes [18]. These results correspond to the default class evaluations shown in the main paper (Tab. 5). For **other-X** classes which are defined in delimitation to other classes, *e.g.*, **other-vehicle** than **car**, our zero-shot model suffers from the ambiguous class prompt (see Appendix A.3). Linear probing improves across most classes except, for example, **bicycle**. The performance drop can be explained by the different object-notion between our pseudo and the ground truth labels.

## E Qualitative Results

**Class-agnostic segmentation.** In Fig. E.2, we provide qualitative results for class-agnostic segmentation on Waymo Open dataset. Colors encode identities (IDs) of individual instances. For reference, we provide images of corresponding camera views,

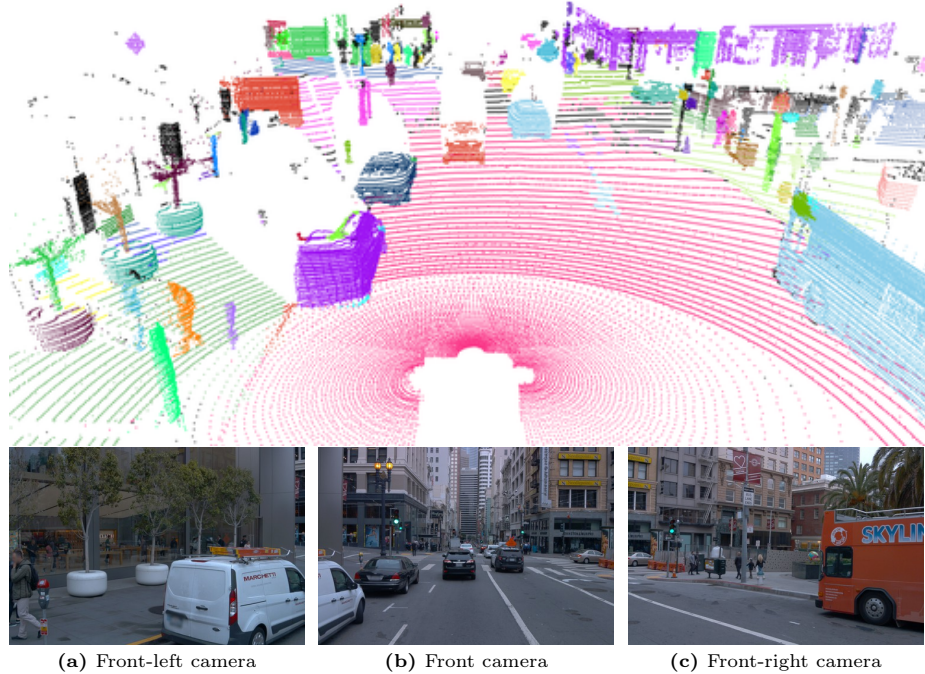


**Fig. E.1: Class-Agnostic Segmentation on Waymo Open [75].** We visually compare class-agnostic segmentation results. Colors encode object instance IDs. *Left*: baseline (SAM [31], unprojected to Lidar), and, *right*, SAL. As can be seen, the baseline that directly lifts SAM masks to Lidar is limited to  $270^\circ$  field of view, which overlaps with the camera ring. By contrast, SAL segments the full Lidar point cloud and is not limited by the camera coverage. Zoomed-in regions show that the baseline is sensitive to edge bleeding (*e.g.*, see **pedestrian** and **traffic sign** masks, partially projected to the blue wall). SAL, by contrast, distills noisy SAM masks into crisp segmentation masks.

even though these are not used during the inference. As can be seen, SAL learns to segment full point clouds, **things** and **stuff** classes, even though supervision is only partial. Interestingly, SAL segments well large structures (often classified as **stuff** classes), such as **building**, **road**, and **sidewalk**. In addition to canonical **things** classes, such as **car**, **van**, **bus**, and **pedestrian**, SAL also learns a variety of classes, that are not covered in class vocabularies of existing datasets of Lidar Panoptic Segmentation (SemanticKITTI [6] and nuScenes [18]). Examples of such classes, segmented in Fig. E.2, are parking meters, potted trees, rooftop ladder, water hydrant, post box, traffic cone and traffic barrier.

In Fig. E.1, we contrast a baseline that simply lifts SAM [29] masks to Lidar (this baseline is evaluated in Tab. B.3, as well as Tab. 4 in the main paper). As can be seen in Fig. E.1, *left*, Waymo Open provides  $270^\circ$  coverage (four camera views), leaving a “blind spot” behind the vehicle. This is an inherent limitation of the baseline that requires camera views for the inference. By contrast, SAL (Fig. E.1, *right*) learns to distill such image-generated pseudo-labels to a full *Lidar Panoptic Segmentation* model. Therefore, it segments full  $360^\circ$  point clouds and does not require camera views during the inference (only during the model training).

For further insights, we zoom in on certain regions of point clouds. As can be seen in Fig. E.1, *left*, masks transferred from images to Lidar often lead to bleeding edges (*e.g.*, red and blue pedestrians “bleed” to the pale blue wall), and thinner structures are often not segmented well due to non-perfect calibration and rolling shutter nature of the Lidar sensor. While such issues can be (partially) mitigated by postprocessing via density-based clustering (we refer to Tab. B.3 and Tab. 4 in the main paper for quantitative analysis), we show in Fig. E.1, *right*, results obtained with SAL model, trained directly on SAM-transferred masks (without DBSCAN as postprocessing). Remarkably, the distilled model does not suffer from these artifacts.



**Fig. E.2: Class-agnostic segmentation on Waymo Open [75] from first-person perspective.** We visually outline the Lidar point cloud, where points are colored according to estimated instance IDs, estimated by SAL. We show corresponding camera views (not used for inference) for reference. As can be seen, SAL accurately segments a large variety of objects, including parking meters, potted trees (pots as well as trees), rooftop ladder, water hydrant, post box, traffic cone, traffic barrier, and more. Canonical objects, such as car, van, bus, and pedestrian are segmented as well. This class-agnostic segmentation is a basis for zero-shot classification.

**Zero-shot prompting.** In Fig. E.9, we highlight the capability of SAL for prompting specific semantic classes. On the *left* side, we visualize class-agnostic segmentations of point clouds, and on the *right*, we highlight prompts and highlight segmented regions. We show two **things** classes (**tram** and **trash bin**) and two **stuff** classes (**store front** and **curb**). A basis for such zero-shot prompting is our class-agnostic segmentation model, which segments input point clouds into a set of segmented objects. SAL model predicts for each segmented object CLIP feature token that we use for zero-shot prompting (for details, we refer to Sec. 3 of the main paper).

**Zero-shot semantic segmentation.** To provide further insight, we illustrate semantic ground truth labels, SAL pseudo-labels, that we use to train the SAL model, and SAL model outputs for SemanticKITTI [6] (Fig. E.6), nuScenes [18] (Fig. E.7) and Waymo Open [75] datasets.

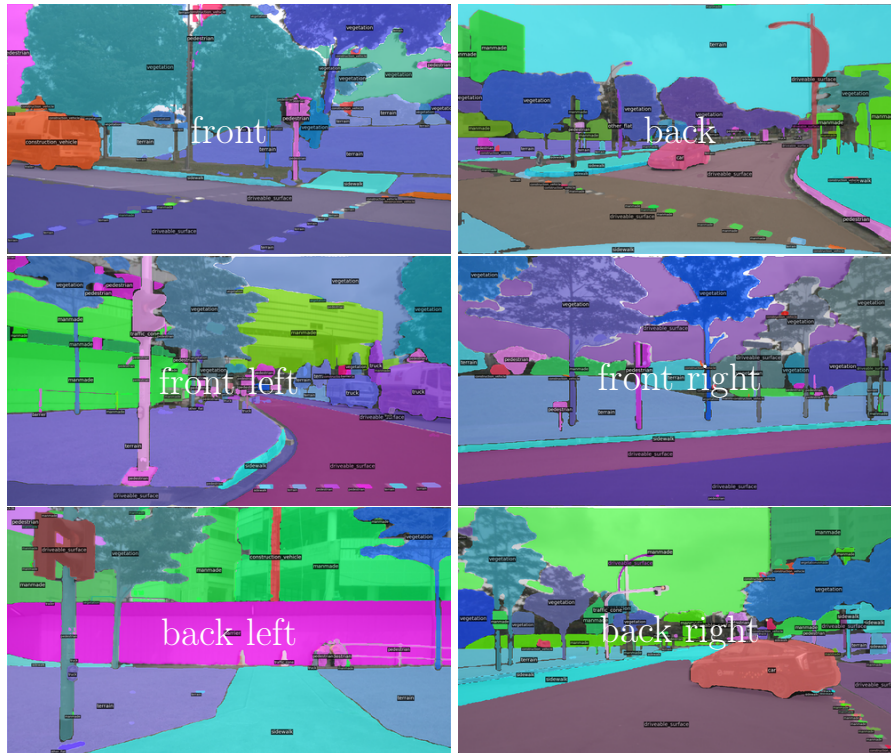
For ground truth and model outputs, we visualize semantic classes. Since pseudo-labels are class-agnostic, colors encode object instance IDs in the middle column. As can

be seen in [6], the single-camera setup in SemanticKITTI provides limited supervision only in the frustum. Across all shown output examples, the segmentation of **things** and **stuff** objects is close to the expected ground truth.

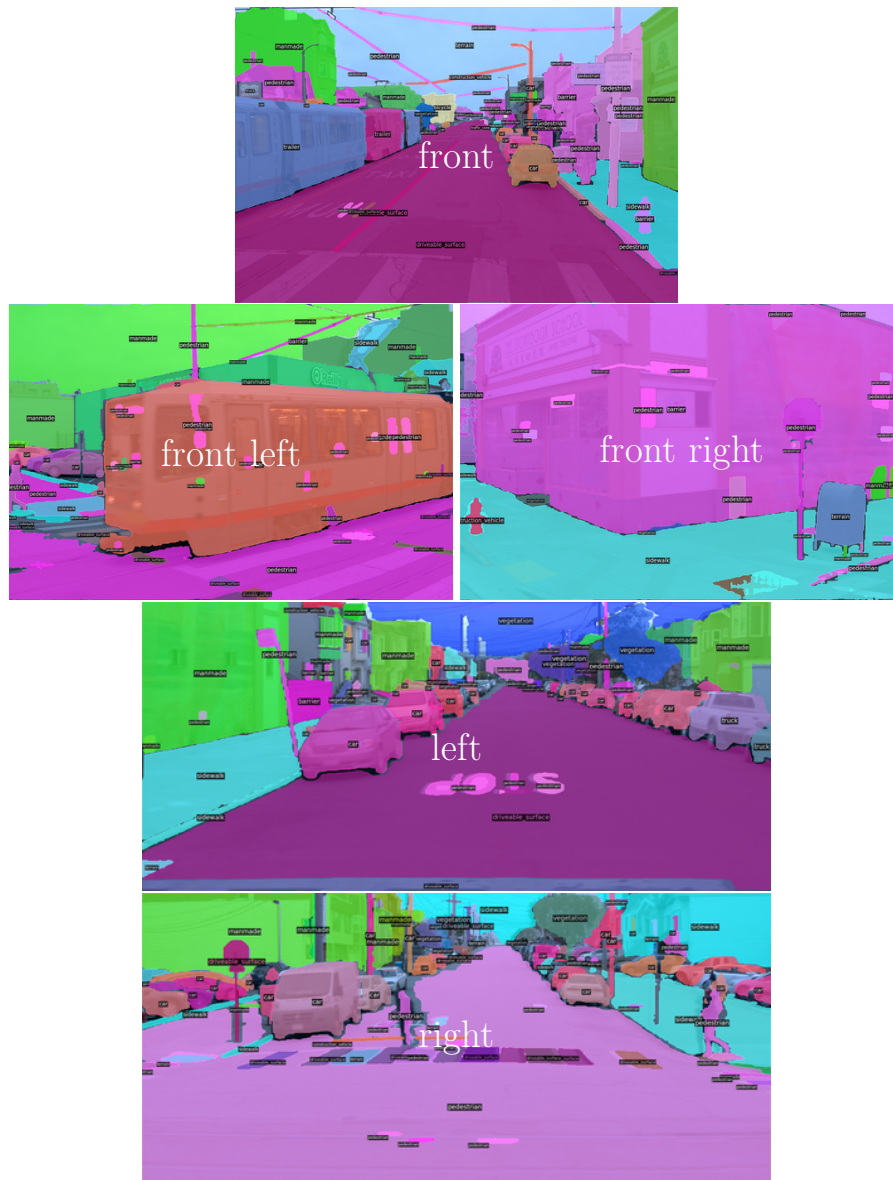








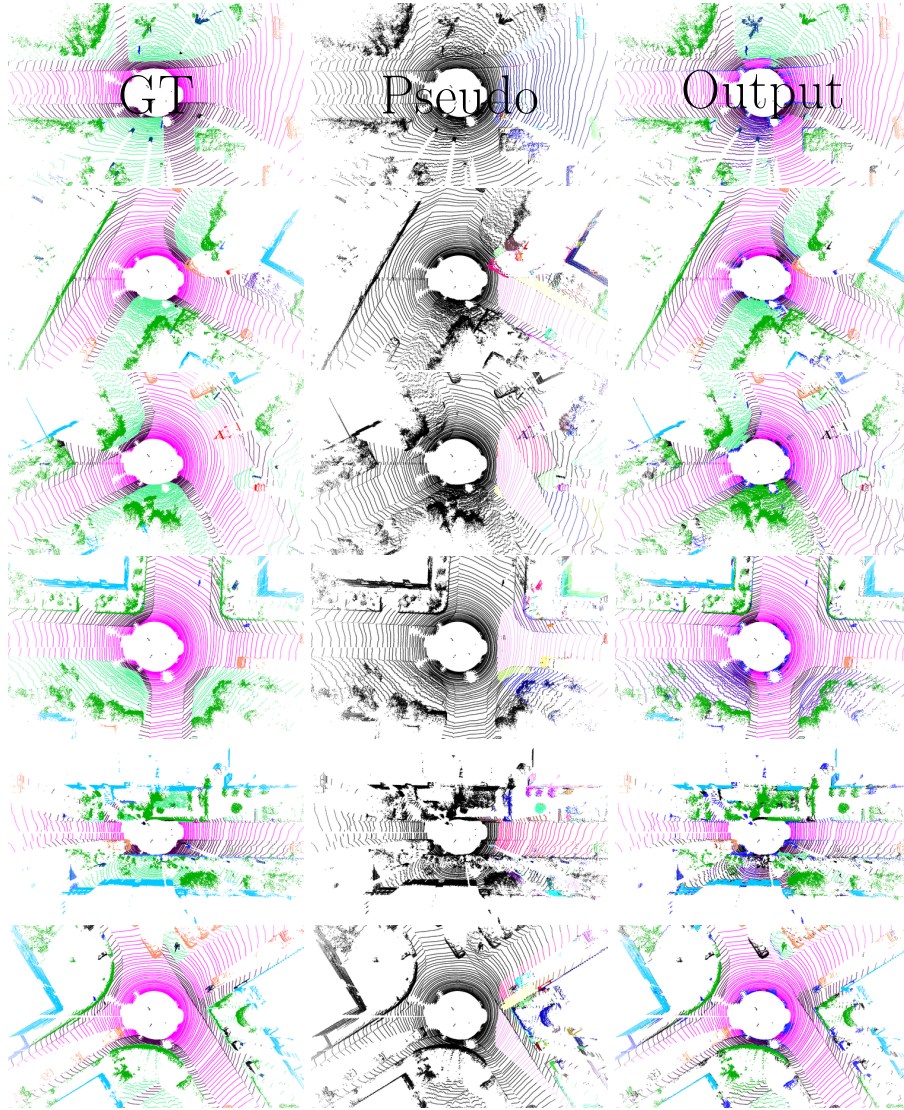
**Fig. E.4: SAM mask and CLIP vocabulary predictions - nuScenes.** We show all six camera views of the first scan of the *0001* sequence. We obtain masks with SAM [31] and compute per-mask CLIP image tokens with MaskCLIP [16]. To visualize classes, we prompt generated tokens with nuScenes [18] class vocabulary. These classes are not transferred to Lidar. Our pseudo-labels contain masks and image tokens, no explicit class labels. Instances of the same class are indicated in different tones of the same color.



**Fig.E.5: SAM mask and CLIP vocabulary predictions - Waymo.** We show all five camera views of the first frame of sequence emph008 of the test set. We obtain masks with SAM [31] and compute per-mask CLIP image tokens with MaskCLIP [16]. Since WAYMO has no panoptic ground truth, we visualize classes by prompting generated tokens with the nuScenes [18] class vocabulary. These classes are not transferred to Lidar. Our pseudo-labels contain masks and image tokens, no explicit class labels. Instances of the same class are indicated in different tones of the same color.

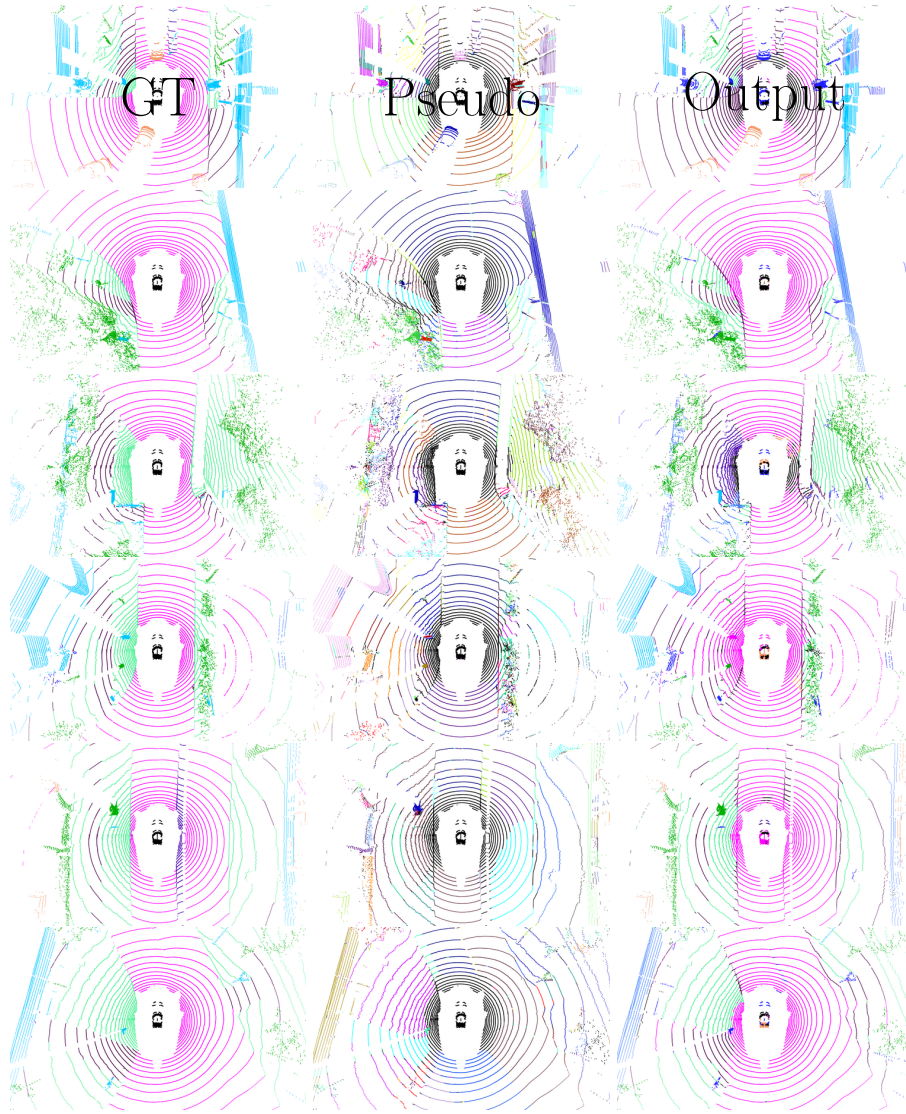
**Table E.1: Dataset vocabulary text prompts.** To circumvent ambiguous or uninformative class names, we prompt each class with a set of possible text prompts. Furthermore, we follow [67] and wrap each class prompt in full-sentence templates.

Class	Text prompts
SemanticKITTI [6]	
car	car, jeep, SUV, van
bicycle	bicycle, bike
motorcycle	motorcycle, moped
truck	truck, pickup truck
other-vehicle	other-vehicle, caravan, trailer, bus, tram, train
person	person, pedestrian
bicyclist	bicyclist, bicycle rider
motorcyclist	motorcyclist, motorcycle rider
road	road, lane
parking	parking, parking lot
sidewalk	sidewalk, curb, driveway
other-ground	other-ground, traffic island
building	building, garage, wall, window, stair
fence	fence, separator, small wall, crash barrier
vegetation	vegetation, bush, shrub, foliage, treetop
trunk	trunk, tree trunk
terrain	terrain, gras, soil
pole	pole, lamp post, traffic-sign pole
traffic-sign	traffic-sign, traffic-sign mounting
nuScenes [18]	
bicycle	bicycle, bike
bus	bus
car	car, jeep, SUV, van
construction vehicle	construction vehicle, crane, excavator
motorcycle	motorcycle, moped
pedestrian	pedestrian, person
trailer	trailer
truck	truck, pickup truck
barrier	barrier, fence, separator, small wall, crash barrier
traffic cone	traffic cone
driveable surface	driveable surface, road, service lanes, bike lanes
flat surface	flat surface, ground
sidewalk	sidewalk, curbs, driveways
terrain	terrain, gras, soil
manmade	manmade, building, garage, walls, windows, stairs, bench
vegetation	vegetation, bush, shrub, foliage, treetop
Super classes	
vehicle	vehicle, car, truck, bicycle, motorcycle, other-vehicle, jeep, SUV, van, bike, moped, pickup truck, caravan, trailer, bus, tram, train, construction vehicle, crane, excavator
human	human, person, bicyclist, motorcyclist, pedestrian, bicycle rider, motorcycle rider
ground	ground, road, sidewalk, parking, other-ground, driveable area, service lane, bike lane, parking lot, curb, driveway, traffic island
structure	structure, building, garage, wall, window, stair
nature	nature, vegetation, trunk, terrain, bush, shrub, foliage, treetop, tree trunk, gras, soil
object	object, fence, pole, traffic-sign, lamp post, traffic-sign pole, traffic-sign mounting, separator, small wall, crash barrier, traffic cone, bench

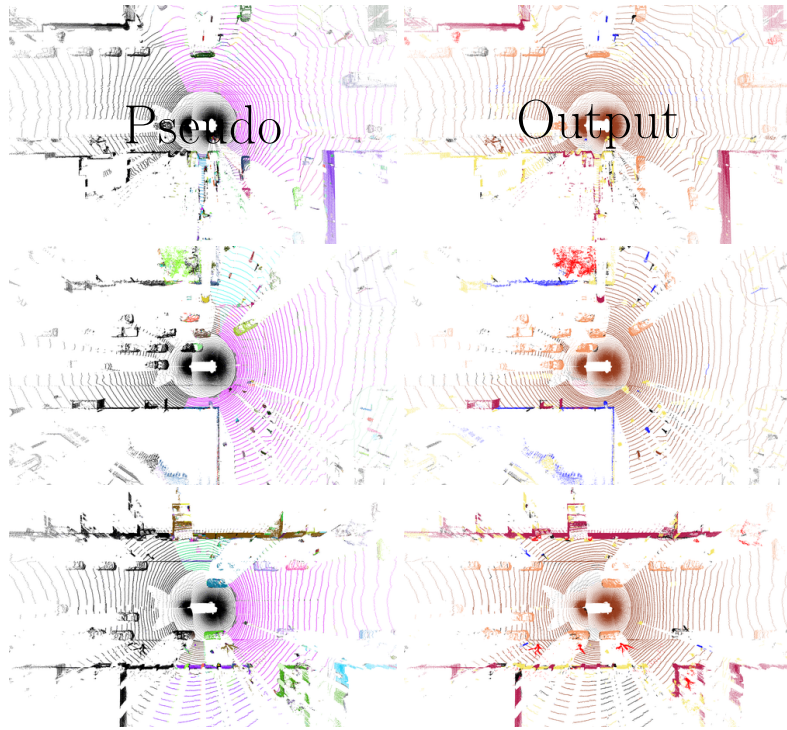


**Fig.E.6: Qualitative results for SemanticKITTI.** We visualize ground truth (GT), pseudo labels, and our model output for several scans of validation sequence *08* of SemanticKITTI [6]. While GT and our output display semantics, the class-agnostic pseudo labels show instances.

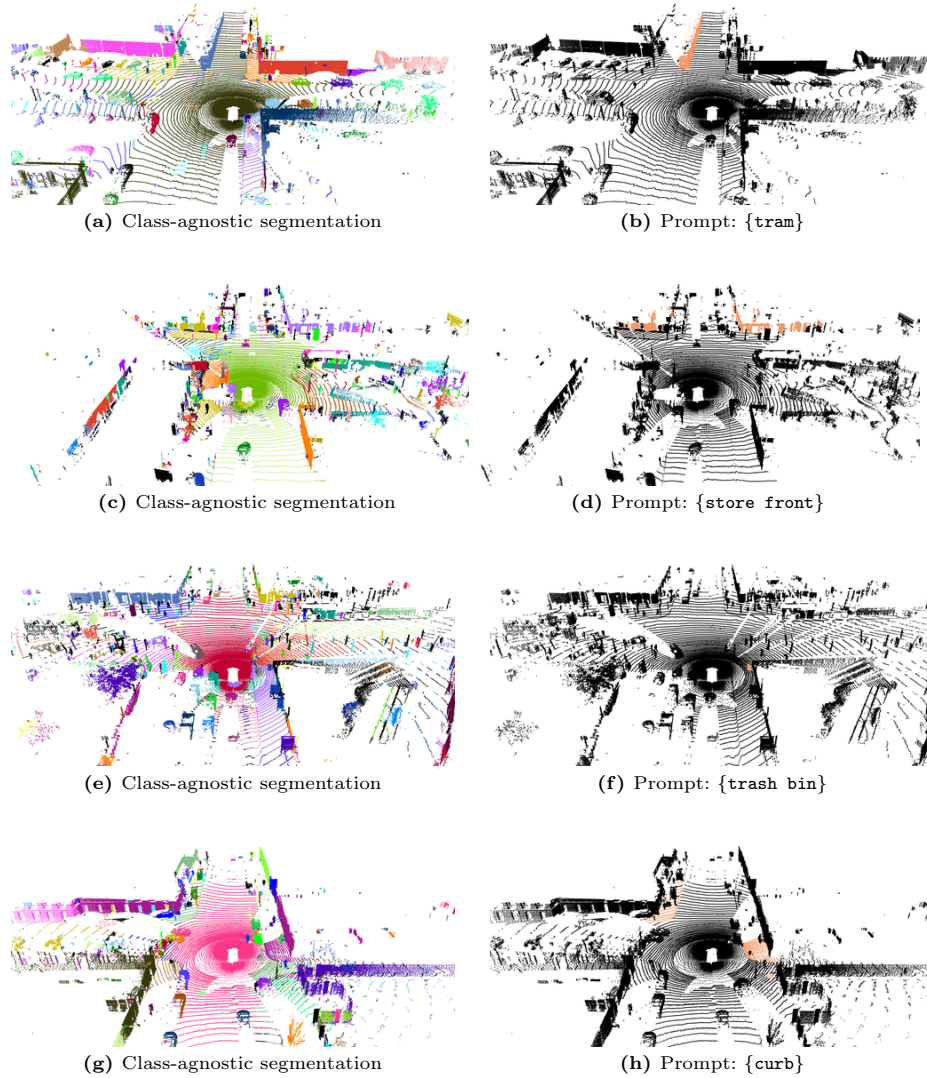




**Fig.E.7: Qualitative results for nuScenes.** We visualize ground truth (GT), pseudo labels, and our model output for the first scan of validation sequences *0003*, *0013*, *0015*, *0017*, *0035*, and *0038* of nuScenes [18]. While GT and our output display semantics, the class-agnostic pseudo labels show instances.



**Fig. E.8: Qualitative results for Waymo.** We visualize pseudo labels and our model output for the first scan of test sequences *008*, *024*, and *032* of Waymo [75]. Waymo does not provide panoptic ground truth labels. While our output displays semantics, the class-agnostic pseudo labels show instances.



**Fig. E.9: Zero-shot per-class prompting on Waymo Open [75].** SAL predicts a set of object instances (*left*), along with their objectness scores and *distilled* CLIP [67] features. We can use text prompts and query these instances for specific classes specified as prompts. On the *right*, we highlight several such examples that are outside of class-vocabularies of SemanticKITTI, nuScenes, and Waymo Open datasets. As can be seen on the left, a basis for such zero-shot prompting is accurate and, importantly, *diverse* class-agnostic segmentation.