Multi-Space Alignments Towards Universal LiDAR Segmentation

Youquan Liu[*,1] Lingdong Kong[*,2,3] Xiaoyang Wu[4] Runnan Chen[4] Xin Li[5]
Liang Pan[2] Ziwei Liu[6] Yuexin Ma[1]
[1]ShanghaiTech University, [2]Shanghai AI Laboratory, [3]National University of Singapore
[4]University of Hong Kong, [5]East China Normal University, [6]S-Lab, Nanyang Technological University
https://github.com/youquanl/M3Net

Abstract

*A unified and versatile LiDAR segmentation model with strong robustness and generalizability is desirable for safe autonomous driving perception. This work presents **M3Net**, a one-of-a-kind framework for fulfilling **m**ulti-task, **m**ulti-dataset, **m**ulti-modality LiDAR segmentation in a **universal** manner using just a **single** set of parameters. To better exploit data volume and diversity, we first combine large-scale driving datasets acquired by different types of sensors from diverse scenes and then conduct alignments in three spaces, namely data, feature, and label spaces, during the training. As a result, M3Net is capable of taming heterogeneous data for training state-of-the-art LiDAR segmentation models. Extensive experiments on twelve LiDAR segmentation datasets verify our effectiveness. Notably, using a **shared** set of parameters, M3Net achieves 75.1%, 83.1%, and 72.4% mIoU scores, respectively, on the official benchmarks of SemanticKITTI, nuScenes, and Waymo Open.*

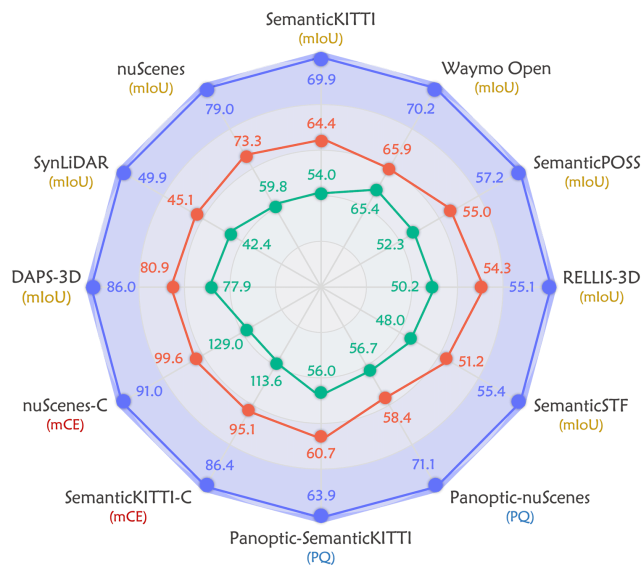

Figure 1. Performance comparisons among **M3Net** [•], *Single-Dataset Training* [•], and *Naïve Joint Training* [•] across **twelve** LiDAR segmentation datasets. For better comparisons, the radius is normalized based on M3Net's scores. The larger the area coverage, the higher the overall performance. Best viewed in colors.

1. Introduction

Dense and structural 3D surrounding scene understanding provides crucial information for autonomous vehicles to make proper decisions [70]. With the recent advancements in sensing technologies, especially the Light Detection and Ranging (LiDAR) sensor, a holistic scene perception can be achieved by segmenting the acquired sensor data [28, 83].

Most existing LiDAR segmentation models [1, 36, 111, 121, 128] are trained and tested in a *single-task, single-dataset, single-modality* manner. Despite achieving commendable results in the single domain, there is a significant performance drop when transitioning to new domains [40, 48]. The limited generalization capability hinders their facilitation of real-world applications [47, 50, 87]. In reality, LiDAR datasets are marred by significant variances, encompassing variations in data patterns due to different sensor types and weather conditions, diverse class distributions arising from varying capture scenarios, and distinct label spaces shaped by specific annotation protocols. These factors collectively pose a formidable challenge in harmonizing disparate LiDAR point clouds and jointly optimizing model parameters to effectively address multiple tasks across a range of sensor modalities [89, 118]. Empirical evidence in Fig. 3 further reveals that naïvely combining heterogeneous data to train a LiDAR segmentation model – without strategic alignments – often leads to sub-opt results.

Recent works [6, 39, 48, 79, 87, 94, 106] resort to unsupervised domain adaptation (UDA) for utilizing training data from both source and target domains to optimize one parameter set. Nevertheless, they either focus on only the sharing mapping between two domains (by ignoring dis-

[*] The first two authors contributed equally to this work.

joint classes) or directly merge source domain labels to align with the target domain [40, 113]. The overlook of the performance degradation on the source dataset and the destruction of original label mappings inevitably constrains such a learning paradigm. Furthermore, there have been efforts [85, 96, 104, 118] to employ multi-dataset learning strategies to bolster the generalization prowess of 3D perception models. However, they either necessitate dataset-specific fine-tuning, deviating from a truly universal learning approach, or converge label spaces to a coarser set, resulting in the dilution of fine-grained segmentation capabilities across diverse semantic categories.

In this work, we define a novel paradigm towards leveraging LiDAR point clouds from different datasets to tame *a single set of parameters* for multi-task LiDAR segmentation. Sibling to image segmentation communities [43, 53, 124], we call this paradigm **universal LiDAR segmentation**. The ultimate goal of such a synergistic way of learning is to build a powerful segmentation model that can absorb rich cross-domain knowledge and, in return, achieve strong resilience and generalizability for practical usage. Given the substantial differences among datasets in terms of data characteristics, feature distributions, and labeling conventions, we introduce a comprehensive *multi-space alignment* approach that encompasses data-, feature-, and label-level alignments, to effectively pave the path for efficient and universally applicable LiDAR segmentation. In particular, the multi-modal data, including images and texts, is fully exploited to assist the alignment process with the guidance of more general knowledge. Through aforementioned processes, we propose **M3Net** to learn common knowledge across datasets, modalities, and tasks, thereby significantly enhancing its applicability in practical scenarios.

To substantiate the efficacy of M3Net and the utility of each module developed, we have carried out a series of thorough comparative and ablation studies across an extensive array of driving datasets, as shown in Fig. 1. Notably, our best model achieves state-of-the-art LiDAR segmentation performance with 75.1%, 83.1%, 72.4% mIoU scores on *SemanticKITTI* [3], *nuScenes* [27], *Waymo Open* [90], respectively, using a *shared* set of parameters. Moreover, our approach also performs well for direct knowledge transfer and out-of-distribution adaptations, further underscoring its robust capability for effective knowledge transfer.

2. Related Work

LiDAR Segmentation. A holistic perception of 3D scenes is crucial for safe autonomous driving [4, 7, 33, 51, 61]. Various LiDAR segmentation models have been proposed, with distinct focuses on aspects include LiDAR representations [20, 74, 91–93, 103, 121, 128], model architectures [1, 17, 24, 35, 46, 52, 80, 112], sensor fusion [18, 62, 64, 113, 129], post-processing [111, 123], data augmentations [75, 84, 105], *etc*. Most recently, researchers started to explore data efficiency [49, 56], annotation efficiency [57, 63, 65, 86, 97], annotation-free learning [10, 11, 122], zero-shot learning [12, 69], domain adaptation [6, 39, 48, 54, 73, 79, 106], and robustness [47] in LiDAR segmentation, shedding lights for practitioners. Existing pursues, however, learn *separate* parameter sets for each dataset, impeding the scalability. This motivates us to explore LiDAR segmentation in a multi-task, multi-dataset, multi-modality manner with just a *single* set of parameters.

Multi-Task Learning. A proper pipeline design could enable the model to generate suitable predictions to fulfill multiple tasks simultaneously [16, 31]. The current research endeavors mainly focus on building image or video segmentation models to handle semantic, instance, and panoptic segmentation tasks [38, 59, 98, 100, 119, 120, 130]. Recently, several attempts have been made to enable multi-task segmentation on LiDAR point clouds. MaskRange [30] and MaskPLS [71] extend the mask classification paradigm [15] for joint semantic and panoptic LiDAR segmentation. LidarMultiNet [117] uses global context pooling and task-specific heads to handle LiDAR-based detection and segmentation. P3Former [108] proposed a specialized positional embedding to handle the geometry ambiguity in panoptic LiDAR segmentation. Our framework also supports multi-task learning. Different from existing approaches, the proposed M3Net stands out by combining knowledge from different sensor data across multiple data sources, which achieves superior performance on each task.

Multi-Dataset Learning. Leveraging data samples from different sources for training has been proven effective in enhancing robustness and generalizability [72]. Various approaches have been proposed to merge image datasets for object detection [14, 58, 60, 99, 125, 126], image segmentation [29, 42, 43, 53, 124], depth estimation [13, 82], *etc*. Due to large domain gaps, the image-based methods are often hard to be transferred to 3D. To combine multiple LiDAR datasets for 3D object detection, MDT3D [89] defines a coarse label set to handle the label space conflicts in different point cloud datasets. MS3D++ [95, 96] ensembles pre-trained detectors from different source datasets for multi-domain adaptation. Uni3D [118] resorts to dataset-specific detection heads and feature re-coupling for training a unified 3D object detector. Recently, PPT [104] proposed to pre-train a point cloud segmentation network using data from multiple datasets. However, the pre-trained weights are then fine-tuned on each specific dataset, which breaks the universal learning manner. The closest work to us is COLA [85], which trains a single model across multiple sources by converting dataset-specific labels to a common coarse set. Such a conversion, however, leads to the loss of fine-grained segmentation across the various semantic categories. Differently, our M3Net is tailored to tame a *single*

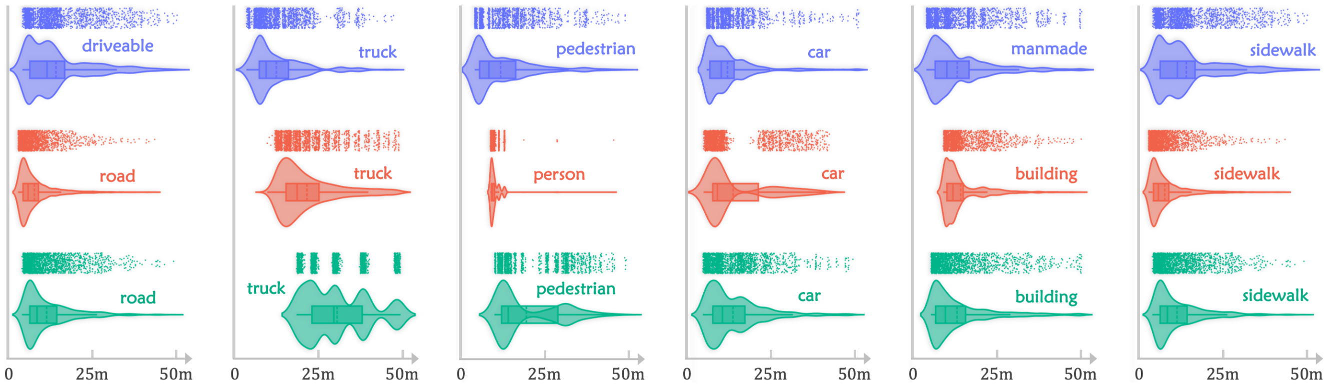

Figure 2. **Statistical analysis** of six sharing semantic classes in the *nuScenes* [•], *SemanticKITTI* [•], and *Waymo Open* [•] datasets. Each violin plot shows the class distribution across LiDAR scenes spanning 50 meters, centered around the ego-vehicle. Best viewed in colors.

parameter set to fulfill multi-task prediction across multiple datasets while still maintaining the original label mappings.

Multi-Modality Learning. Recent trend favors synergistic learning from data of different modalities, such as vision, language, and speech [2, 8, 19, 26, 76, 81, 101]. For LiDAR segmentation, several works [9, 39, 40, 67, 115] explored the distillation of image features to point clouds. Recently, OpenScene [78] and CLIP2Scene [11] proposed to leverage point clouds along with multi-view images and language for open-vocabulary learning. PPKT [66], SLidR [86], and Seal [65] form cross-sensor contrastive learning objectives to pre-train the LiDAR segmentation models. The advantages of sensor fusion have been consistently proven. In this work, to pursue universal LiDAR segmentation, we propose to align multi-space point clouds via images and texts.

3. Approach

Our study serves as an early attempt at combining *multi-task, multi-dataset, multi-modality* knowledge into a *single* set of parameters to fulfill **universal LiDAR segmentation**. We start with a pilot study to unveil the difficulties in merging heterogeneous LiDAR point clouds (*cf*. Sec. 3.1). We then present M3Net, a versatile LiDAR segmentation network tailored to pursue *i)* statistical consistency in the data space (*cf*. Sec. 3.2), *ii)* cross-modality-assisted alignment in the feature space (*cf*. Sec. 3.3), and *iii)* language-guided unification in the label space (*cf*. Sec. 3.4).

3.1. Pilot Study

The current de facto of training a LiDAR segmentation network adopts a *task-by-task* and *dataset-by-dataset* pipeline. Despite the superior performance achieved under such standalone settings, the trained parameter sets cannot be shared to satisfy out-of-domain requirements and, therefore, limits their use cases for practical applications.

Naïve Joint Training. A natural alternative to breaking the above constraint is to jointly train a network across multiple datasets for better generalizability. However, as depicted in Fig. 2, it is often non-trivial to naïvely combine heterogeneous data with large data distribution gaps to train a universal LiDAR segmentation model without proper alignments. To testify this, we conducted a pilot study using the prior art MinkUNet [20] for both standalone and joint training on three large-scale datasets [3, 27, 90]. As shown in Fig. 3 (a) and (d), a brutal combination undermines the segmentation performance. Due to large discrepancies in aspects like sensor configurations, data acquisitions, label mappings, and domain shifts, the jointly trained representations tend to be disruptive instead of being more general.

LiDAR Sensor Discrepancy. To understand the root cause of performance degradation, we conducted another study that controls point cloud density discrepancies when merging datasets. As shown in Fig. 3 (b) and (c), joint training on data collected by sensors with different beam numbers tends to suffer more severely than merging less density variant data. We hypothesize that this is mainly caused by the data statistical variations. In light of these observations, we propose a bag of suitable operations in the following sections to alleviate the large domain gaps among different LiDAR segmentation datasets [3, 4, 7, 27, 90].

3.2. Data-Space Alignment

Given a total of S datasets $D^s = \{(x^s, y^s)|1 \leq s \leq S\}$, where (x^s, y^s) denotes the data-label pairs constituting a dataset. For the LiDAR segmentation task, x^s often encompasses the LiDAR point cloud $P^s = \{p_x, p_y, p_z\}^s \in \mathbb{R}^{N\times 3}$ and synchronized multi-view camera images $V^s = \{\mathcal{I}_1, ..., \mathcal{I}_l\}|l = 1, ..., L\}$, where $\mathcal{I}_t \in \mathbb{R}^{H\times W\times 3}$, N is the number of points, L denotes the number of camera sensors, H and W are the height and width of the image, respectively. $y^s \in \mathbb{R}^N$ denotes point cloud labels in the label space $\mathbb{Y}^s$, we unify the label space as $\mathbb{Y}^u = \mathbb{Y}^1 \cup \mathbb{Y}^2 ... \cup \mathbb{Y}^S$.

Cross-Modality Data Alignment. As a multi-sensing system, the information encoded in P_i^s and V_i^s are intuitively

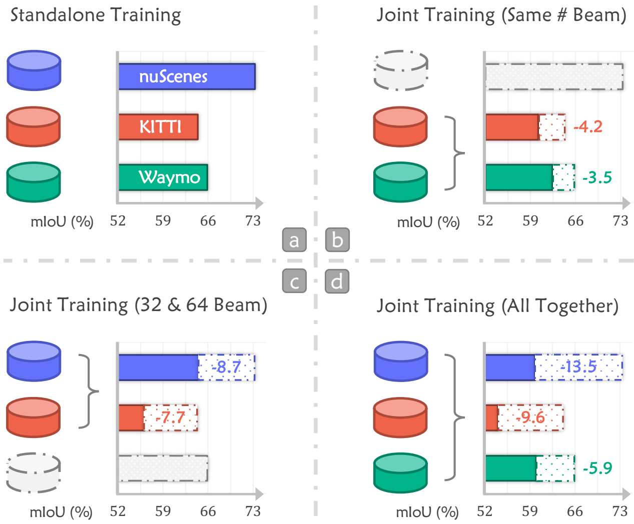

Figure 3. **A pilot study** of naïvely merging different datasets for training the MinkUNet [20] model. Compared to the standalone training in **(a)**, either jointly training with **(b)** the same, **(c)** different, or **(d)** all sensor-acquired data will cause severe degradation.

complementary to each other [3, 7, 90]. To leverage such an advantage, we resort to the correspondences embedded in camera calibration matrices to bridge the LiDAR points and camera image pixels. Specifically, for each point $\mathbf{p} = (p^x, p^y, p^z)$ in P^s, the corresponding pixel (u, v) can be found by the following transformations:

$$[u, v, 1]^{\mathrm{T}} = \frac{1}{p^z} \cdot T_s \cdot T \cdot [p^x, p^y, 1]^{\mathrm{T}}, \tag{1}$$

where $T \in \mathbb{R}^{4\times4}$ is the camera extrinsic matrix that consists of a rotation matrix and a translation matrix, and $T_s \in \mathbb{R}^{3\times4}$ is the camera intrinsic matrix. As we will show in the following sections, such a cross-sensor data alignment serves as the foundation for alignments in other spaces.

Cross-Sensor Statistical Alignment. To mitigate the discrepancies in sensor installations across different datasets, we incorporate a point coordinate alignment operation. Specifically, drawing upon insights from prior domain adaptation approaches [102, 116], we adjust the coordinate origins of point clouds from different datasets by introducing an offset $\sigma \in \mathbb{R}^{1\times3}$ to the ground plane. We find empirically that such an alignment can largely reduce the degradation caused by the variations in different sensor setups.

Dataset-Specific Rasterization. It is conventional to rasterize LiDAR point clouds P^s using unified rasterization parameters, *e.g.*, voxel size [91, 128] or horizontal range view resolution [74, 111]. However, the point clouds acquired in different LiDAR datasets naturally differ in density, range, intensity, *etc.*, which tends to favor different rasterization parameters. To meet such a requirement, we select dataset-specific parameters for rasterization on each dataset through empirical experiments and analyses.

Decoupled BN. Another challenge in training across multiple datasets is the presence of domain gaps, which can result in significant statistical shifts of feature learning among datasets. Such shifts can hinder the convergence and affect the model's ability to generalize well across diverse datasets. We adopt a decoupled batch norm (BN) for point cloud features in each dataset. Instead of using the traditional BN, which calculates mean and variance across all samples in a mini-batch, the decoupled BN tends to adapt each dataset's specific characteristics independently.

3.3. Feature-Space Alignment

We aim to acquire a generalized feature representation for downstream tasks. Compared to point clouds, images contribute stronger visual, textural, and semantic information. Thus, the collaboration between pixels and points could enrich the overall representation. Previous research [11, 60, 64] has consistently demonstrated that such a combination results often leads to improved performance.

Cross-Modality Assisted Alignment. In the context of multi-dataset joint training, our objective is to establish a unified feature space by leveraging image features to assist point cloud features. Acknowledging that images used in training lack ground truth labels [3, 27], we utilize image features from a pre-trained model as an alternative, facilitating a more universally applicable representation. We feed camera images V^s into a pre-trained DeepLab [98] and a vision-language model (VLM) and visualize the output image features by t-SNE [25]. As shown in Fig. 4, we observe that image features from DeepLab appear disorderly and lack semantics. In contrast, features from VLM share a more unified feature space. Motivated by this, we propose a cross-modality assisted alignment that uses VLM to help align the feature space. Specifically, the camera images V^s are fed to the frozen image encoder from VLM to obtain image features $F_v = \{\mathcal{F}_v^1, \mathcal{F}_v^2, ..., \mathcal{F}_v^s\}$, where $\mathcal{F}_v^s \in \mathbb{R}^{c\times h\times w}$. The LiDAR point clouds P^s, on the other hand, are fed to the point encoder followed by a projection layer to generate the point features $F_p = \{\mathcal{F}_p^1, \mathcal{F}_p^2, ..., \mathcal{F}_p^s\}$, where $\mathcal{F}_p^s \in \mathbb{R}^{m\times c}$; m denotes the number of non-empty grids. We then leverage the paired image features $\hat{F}_v \in \mathbb{R}^{m_p\times c}$ and point feature $\hat{F}_p \in \mathbb{R}^{m_p\times c}$ for alignment, where m_p is the number of point-pixel pairs. After obtaining $\hat{F}_p$ and $\hat{F}_v$, the cross-modality alignment is expressed as follows:

$$\mathcal{L}_{\text{cma}}(\hat{F}_v, \hat{F}_p) = 1 - \frac{\hat{F}_v \cdot \hat{F}_p}{\|\hat{F}_v\| \cdot \|\hat{F}_p\|} . \tag{2}$$

Domain-Aware Cross-Modality Alignment. With cross-modality alignment, we transfer the knowledge of VLM to the point encoder, enabling the point features to gain a more comprehensive representation. However, during the execution of the above alignment, we have narrowed

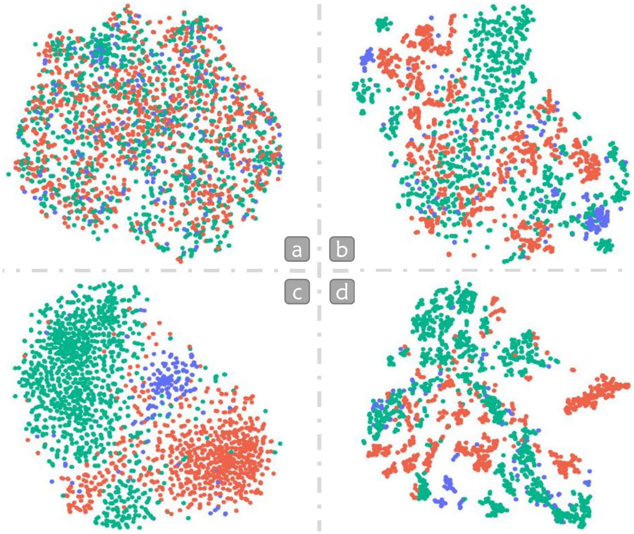

Figure 4. **The t-SNE plots of learned features** before and after the feature-space alignment in merging the *nuScenes* [•], *SemanticKITTI* [•], and *Waymo Open* [•] datasets. We show image features from **(a)** standalone networks; **(b)** SAM [44], and point cloud features **(c)** before and **(d)** after the feature-space alignment.

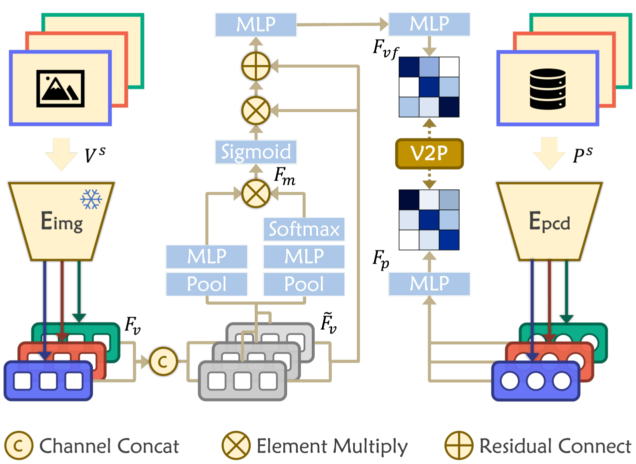

Figure 5. **Feature-space alignment** in M3Net. We leverage both image features F_v and LiDAR point cloud features F_p extracted from image encoder E_{img} and point encoder E_{pcd} to employ the regularization via V2P loss and achieve feature-space alignment.

it exclusively to image and point features from the same dataset. In this mode, point features solely learn from matching image features, restricting their knowledge acquisition. Ideally, we aim to ensure that image features encompass not only scenes identical to those represented in point clouds but also scenes from other datasets. To address this, we propose a domain-aware cross-modality guided alignment, as illustrated in Fig. 5. Specifically, we first extract, for each dataset, F_v and F_p from the same image encoder E_{img} and point encoder E_{pcd} during the cross-modality assisted alignment. The sets of features from all datasets are concatenated along the channel dimension to form $\widetilde{F}_v \in \mathbb{R}^{c_v \times h \times w}$. Subsequently, we sequentially feed $\widetilde{F}_v$ through a branch that consists of a global average pooling and an MLP. Simultaneously, $\widetilde{F}_v$ is fed to an auxiliary branch that undergoes the same processing flow and generates an output after the softmax function $\mathcal{G}(\cdot)$. The outputs from both branches are multiplied to obtain $F_m \in \mathbb{R}^{c_v \times 1 \times 1}$. The overall process can be described as follows:

$$F_m = MLP(Pool(\widetilde{F}_v)) \cdot \mathcal{G}(MLP(Pool(\widetilde{F}_v)))\ . \tag{3}$$

Next, we forward F_m to a sigmoid activation function $\mathcal{H}(\cdot)$ and multiply it with input image features $\widetilde{F}_v$. The resulting output is added to $\widetilde{F}_v$ and passed through the MLP layers to obtain the final image features $F_{vf} \in \mathbb{R}^{c \times h \times w}$. The forward process of this operation is depicted as follows:

$$F_{vf} = MLP((\mathcal{H}(F_m) \cdot \widetilde{F}_v) + \widetilde{F}_v)\ . \tag{4}$$

Finally, we leverage the cross-modality data alignment to acquire paired image features $\hat{F}_{vf} \in \mathbb{R}^{m_p \times c}$ and paired point feature $\hat{F}_p$. The overall objective function is:

$$\mathcal{L}_{v2p}(\hat{F}_{vf}, \hat{F}_p) = 1 - \frac{\hat{F}_{vf} \cdot \hat{F}_p}{\|\hat{F}_{vf}\| \cdot \|\hat{F}_p\|}\ . \tag{5}$$

3.4. Label-Space Alignment

Label Conflict. In multi-dataset joint training settings, label conflicts emerge as a significant challenge. This often refers to the inconsistencies in class labels across different datasets involved in the training process. The discrepancy can arise due to variations in annotation conventions, labeling errors, or even differences in the underlying semantics of classes between datasets. In our baseline, we unionize the different label spaces across datasets into $\mathbb{Y}^u$, where all datasets share a single LiDAR segmentation head. However, this may introduce several potential drawbacks:

- *Loss of granularity:* Unified label spaces could lose semantic granularity, particularly when dealing with subtle category differences in between different datasets.
- *Information loss:* During label space consolidation, details unique to each dataset may be obscured or lost, especially for those related to domain-specific categories.
- *Increased complexity:* Handling a unified label space may necessitate more complex model architectures or training strategies, thereby increasing overall complexity.

To address these issues, we introduce a language-guided label-space alignment to facilitate a more holistic semantic correlation across datasets. Given the natural correspondence between images and texts and the strong correlation between images and point clouds, we aim to strategically utilize the image modality as a bridge to establish language-guided alignments. Such a process consists of a text-driven

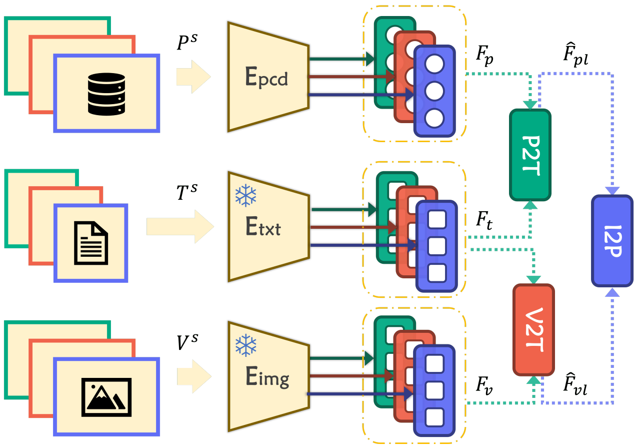

Figure 6. **Label-space alignment** in M3Net. We leverage image features F_v, point cloud features F_p, and text embedding F_t extracted from E_{img}, E_{pcd}, and E_{txt}, respectively, for regularization via the I2P, P2T, and V2T losses in the label-space alignment.

point alignment, a text-driven image alignment, and a cross-modality-assisted label alignment.

Text-Driven Alignments. As depicted in Fig. 6, images V^s are fed into the frozen image encoder E_{img} to extract the image features F_v. Concurrently, the LiDAR point clouds P^s are processed by the point encoder E_{pcd} to generate the point features F_p. Additionally, given the text input T^s, text embedding features $F_t \in \mathbb{R}^{Q\times c}$ are obtained from a frozen text encoder E_{txt}, where Q represents the number of categories across datasets. The text is composed of class names from unified label space $\mathbb{Y}_u$ placed into pre-defined templates, and the text embedding captures semantic information of the corresponding classes. Subsequently, pixel-text pairs $\{v_k, t_k\}_{k=1}^M$ and point-text pairs $\{p_k, t_k\}_{k=1}^M$ are generated, where M represents the number of pairs. Leveraging the semantic information contained in the text, we selectively choose positive and negative samples for both images and points for contrastive learning. It is noteworthy that negative samples are confined to the specific dataset category space. The overall objective of the text-driven point alignment function is shown as follows:

$$\mathcal{L}_{p2t} = -\sum_{q=1}^{Q} \log\left(\frac{\sum_{t_k\in q, p_k} \exp(< t_k, p_k > /\tau)}{\sum_{t_k\in q, t_k\notin q, p_j} \exp(< t_k, p_k > /\tau)}\right) , \tag{6}$$

where $t_k \in q$ indicates that t_k is generated by the q-th classes name, and Q is the number of classes. Symbol $<, >$ denotes the scalar product operation and τ is a temperature term ($\tau > 0$). Similarly, the objective of the text-driven image alignment function is illustrated as follows:

$$\mathcal{L}_{v2t} = -\sum_{q=1}^{Q} \log\left(\frac{\sum_{t_k\in q, v_k} \exp(< t_k, v_k > /\tau)}{\sum_{t_k\in q, t_k\notin q, v_j} \exp(< t_k, v_k > /\tau)}\right) . \tag{7}$$

Cross-Modality-Assisted Label Alignment. After text-driven alignments, the subsequent crucial step entails aligning the point and image modalities within the label space. We first obtain image logits $F_{vl} = \{\mathcal{F}_{vl}^1, \mathcal{F}_{vl}^2, ..., \mathcal{F}_{vl}^s\}$ and point logits $F_{pl} = \{\mathcal{F}_{pl}^1, \mathcal{F}_{pl}^2, ..., \mathcal{F}_{pl}^s\}$ from text-driven alignments, where $\mathcal{F}_{vl}^s \in \mathbb{R}^{Q\times H\times W}$, $\mathcal{F}_{pl}^s \in \mathbb{R}^{N\times Q}$. Subsequently, we conduct cross-modality alignment to obtain paired image logits $\hat{F}_{vl} \in \mathbb{R}^{m_p\times Q}$ and paired point logits $\hat{F}_{pl} \in \mathbb{R}^{m_p\times Q}$. Formally, the cross-modality-assisted alignment in the label space is formulated as follows:

$$\mathcal{L}_{i2p}(\hat{F}_{vl}, \hat{F}_{pl}) = 1 - \frac{\hat{F}_{vl}\cdot\hat{F}_{pl}}{\|\hat{F}_{vl}\|\cdot\|\hat{F}_{pl}\|} . \tag{8}$$

Finally, the complete objective function for the language-guided label-space alignment is expressed as follows:

$$\mathcal{L}_{label} = \mathcal{L}_{p2t} + \mathcal{L}_{i2p} + \mathcal{L}_{v2t} . \tag{9}$$

3.5. Universal LiDAR Segmentation

We enhance the versatility of M3Net via multi-tasking learning. This integration involves an instance extractor to enable joint semantic and panoptic LiDAR segmentation.

Panoptic LiDAR Segmentation. Motivated by DSNet [32, 33], our instance extractor comprises an instance head and a clustering step. The instance head encompasses several MLPs designed to predict the offsets between instance centers. The clustering step uses semantic predictions to filter out *stuff* points, thereby retaining only those associated with *thing* points. The remaining points undergo a mean-shift clustering [21], utilizing features from the instance head to discern distinct instances. Lastly, we employ the L1 loss $\mathcal{L}_{l1}$ to optimize the *thing* point regression process.

Overall Objectives. Putting everything together, the overall objective of M3Net is to minimize the following losses:

$$\mathcal{L} = \mathcal{L}_{v2p} + \mathcal{L}_{label} + \mathcal{L}_{ce} + \mathcal{L}_{lovasz} + \mathcal{L}_{l1} , \tag{10}$$

where $\mathcal{L}_{ce}$ and $\mathcal{L}_{lovasz}$ denote the cross-entropy loss and the Lovasz-softmax [5] loss, respectively.

4. Experiments

4.1. Experimental Setups

Datasets. Our M3Net framework and baselines are trained on a combination of ***nuScenes*** [27], ***SemanticKITTI*** [3], and ***Waymo Open*** [90]. Meanwhile, we resort to another five LiDAR-based perception datasets [41, 45, 77, 106, 107] and two 3D robustness evaluation datasets [47] to verify the strong generalizability of M3Net. Due to space limits, additional details regarding the datasets are in the Appendix.

Implementation Details. M3Net is implemented based on Pointcept [23] and MMDetection3D [22]. We use two backbones in our experiments, *i.e.*, MinkUNet [20] and

Table 1. **Ablation study** on the M3Net alignments happen in the **Data**, **Feature**, and **Label** spaces, respectively, when combining the *SemanticKITTI* [3], *nuScenes* [27], and *Waymo Open* [90] datasets. The mAcc and mIoU scores are in percentage. Best scores are in **bold**.

-	Configurations			MinkUNet [20]						PTv2+ [103]					
				SemKITTI		nuScenes		Waymo		SemKITTI		nuScenes		Waymo	
				mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
Baseline	Naïve Joint Training			62.43	54.03	65.05	59.84	73.76	65.39	67.96	61.59	76.53	69.65	75.68	67.00
M3Net (Ours)	Data	Feature	Label	-		-		-		-		-		-	
	✓			73.82	69.01	83.66	76.89	77.88	69.37	78.55	69.95	86.22	79.13	80.96	72.15
	✓	✓		74.36	69.64	85.17	78.88	78.31	69.70	79.43	70.87	87.10	80.26	80.74	72.33
	✓		✓	73.85	69.34	85.20	78.90	78.04	69.55	80.30	71.13	87.44	80.45	80.69	72.30
	✓	✓	✓	**74.40**	**69.85**	**85.30**	**79.00**	**78.66**	**70.15**	**80.00**	**72.00**	**87.91**	**80.90**	**81.11**	**72.40**

Table 2. **Panoptic LiDAR segmentation** results on the *val* sets of the *Panoptic-SemanticKITTI* [3] and *Panoptic-nuScenes* [27] datasets. All scores are given in percentage. The best and second-best scores are highlighted in **bold** and underline, respectively.

Method	Configurations			Panoptic-SemanticKITTI					Panoptic-nuScenes				
				PQ	$PQ^\dagger$	RQ	SQ	mIoU	PQ	$PQ^\dagger$	RQ	SQ	mIoU
Panoptic-TrackNet [37]	Single-Dataset Training			40.0	-	48.3	73.0	53.8	51.4	56.2	63.3	80.2	58.0
Panoptic-PolarNet [127]				59.1	64.1	70.2	78.3	64.5	63.4	67.2	75.3	83.9	66.9
EfficientLPS [88]				59.2	65.1	69.8	75.0	64.9	59.2	62.8	82.9	70.7	69.4
DSNet [32]				61.4	65.2	72.7	79.0	69.6	64.7	67.6	76.1	83.5	76.3
Panoptic-PHNet [55]				61.7	-	-	-	65.7	**74.7**	**77.7**	**84.2**	**88.2**	79.7
Baseline	Naïve Joint Training			56.03	59.64	65.78	73.72	61.59	56.67	60.61	66.75	83.49	69.65
M3Net (Ours)	Data	Feature	Label			-					-		
	✓			62.34	65.17	72.60	74.67	69.95	68.49	71.11	79.13	85.49	79.13
	✓	✓		62.91	65.73	73.32	75.47	70.87	71.47	73.86	81.53	86.71	80.26
	✓		✓	63.23	67.89	**73.61**	81.66	71.13	71.53	73.91	81.80	86.92	80.45
	✓	✓	✓	**63.87**	**68.66**	73.10	**82.35**	**72.00**	71.70	74.01	82.20	86.47	**80.90**

PTv2+ [103]. We trained M3Net on four A100 GPUs for 50 epochs with a batch size of 6 for each GPU. The initial learning rate is set to 0.002. We adopt the AdamW optimizer [68] with a weight decay of 0.005 and cosine decay learning rate scheduler. For the dataset-specific rasterization, we set voxel sizes to 0.05m, 0.1m, and 0.05m for *SemanticKITTI* [3], *nuScenes* [27], and *Waymo Open* [90], respectively. For the data augmentation, we employ random flipping, jittering, scaling, rotation, and Mix3D [75]. Due to space limits, kindly refer to Appendix for additional details.

Evaluation Metrics. We adopt conventional reportings of *mAcc* and *mIoU* for LiDAR semantic segmentation, *PQ*, $PQ^\dagger$, *SQ*, and *RQ* for panoptic segmentation, and *mCE* and *mRR* for 3D robustness evaluation. Due to the space limit, kindly refer to our Appendix for more detailed definitions.

4.2. Ablation Study

Multi-Space Alignments. The effectiveness of three proposed alignments over the joint training baselines is shown in Tab. 1. We observe that the data-space alignment plays the most crucial role in improving the universal LiDAR segmentation performance. Without proper data alignments, joint training with either MinkUNet [20] or the stronger PTv2+ [103] will suffer severe degradation, especially on sparser point clouds [27]. On top of the data-space alignment, the combinations of multi-view images at the feature space and the language-guided knowledge at the label space further enhance the learned feature representations. The results show that they work synergistically in merging knowledge from heterogeneous domains during joint training.

Panoptic LiDAR Segmentation. In Tab. 2, we present another ablation study focusing on panoptic LiDAR segmentation. All three alignments incorporated in M3Net demonstrate significant improvements over the baselines. This highlights the pronounced efficacy of our multi-space alignments. Moreover, our approach outperforms the single-dataset state-of-the-art method Panaptic-PHNet [55] by a notable 2.17% PQ on *Panoptic-SemanticKITTI* [3] and achieves compelling results on *Panoptic-nuScenes* [27].

Visual Feature Alignments. We conduct a qualitative analysis of the learned visual feature distributions in the form of t-SNE [25]. Fig. 4 (a) and (b) represent the distributions of learned visual features among three datasets from DeepLab and VLM backbones, respectively. The features obtained by the latter exhibit more distinct semantics in feature space. The concentrated distribution space is advantageous for achieving feature alignments across multiple datasets. Additionally, Fig. 4 (c) and (d) illustrate the distribution of point cloud features before and after feature-space alignment. As can be seen, the feature distribution distances

Table 3. **Knowledge transfer and generalization** analyses across five LiDAR segmentation datasets and two 3D robustness evaluation datasets. All scores are given in percentage. The best and second-best scores are highlighted in **bold** and underline, respectively.

Method	RELLIS-3D		SemanticPOSS		SemanticSTF		SynLiDAR		DAPS-3D		SemKITTI-C		nuScenes-C	
	1%	10%	Half	Full	Half	Full	1%	10%	Half	Full	mCE	mRR	mCE	mRR
PPKT [66]	49.71	54.33	50.18	56.00	50.92	54.69	37.57	46.48	78.90	84.00	-	-	105.64	76.06
SLidR [86]	49.75	54.57	51.56	55.36	52.01	54.35	42.05	47.84	81.00	85.40	-	-	106.08	75.99
Seal [65]	51.09	55.03	53.26	56.89	53.46	55.36	43.58	49.26	81.88	85.90	-	-	92.63	**83.08**
Naïve Joint	37.77	50.23	42.19	52.31	46.70	48.00	18.56	42.37	73.91	77.89	113.65	84.73	128.97	81.45
Single-Dataset	40.17	54.25	47.69	55.00	50.33	51.19	23.17	45.08	75.10	80.87	95.11	84.95	99.63	79.06
M3Net (Ours)	**51.27**	**55.05**	**53.60**	**57.17**	**53.78**	**55.42**	**44.10**	**49.93**	**82.08**	**86.00**	**86.43**	**85.77**	**91.03**	79.15

Table 4. **LiDAR semantic segmentation** results on the *val* and *test* sets of *SemanticKITTI* [3] and *nuScenes* [27], and the *val* set of *Waymo Open* [90]. All scores are in percentage. The best and second-best scores are highlighted in **bold** and underline.

Method	SemKITTI		nuScenes		Waymo	
	Val	Test	Val	Test	mIoU	mAcc
RangeNet++ [74]	-	52.2	-	65.5	-	-
PolarNet [121]	57.2	54.3	71.0	69.8	-	-
SalsaNext [24]	-	59.5	-	72.2	-	-
RangeViT [1]	60.7	64.0	75.2	-	-	-
MinkUNet [20]	63.8	63.7	73.3	-	65.9	76.6
SPVNAS [91]	64.7	66.4	-	77.4	67.4	-
AMVNet [62]	65.2	65.3	77.2	77.3	-	-
RPVNet [112]	65.5	70.3	77.6	-	-	-
(AF)^2-S3Net [18]	-	69.7	-	78.3	-	-
Cylinder3D [128]	65.9	67.8	76.1	77.9	66.0	-
PVKD [34]	66.4	71.2	76.0	-	-	-
WaffleIron [80]	66.8	70.8	79.1	-	-	-
RangeFormer [46]	67.6	73.3	78.1	80.1	-	-
SphereFormer [52]	67.8	74.8	78.4	81.9	69.9	-
FRNet [114]	68.7	73.3	79.0	82.5	-	-
PTv2+ [103]	70.3	70.6	80.2	82.6	70.6	80.2
LidarMultiNet [117]	-	-	-	81.4	**73.8**	-
M3Net (Ours)	**72.0**	**75.1**	**80.9**	**83.1**	72.4	**81.1**

between the three datasets have been largely reduced, providing evidence of the alignment effectiveness.

4.3. Comparative Study

Comparisons to State of the Arts. In Tab. 4, we compare M3Net with current best-performing models on the benchmarks of *SemanticKITTI* [3], *nuScenes* [27], and *Waymo Open* [90]. Remarkably, M3Net consistently outperforms existing approaches across all three datasets. Specifically, on *SemanticKITTI* [3], M3Net achieves a 72.0% mIoU on the validation set, surpassing the closest method by a notable margin of 1.7% mIoU. Similarly, on *nuScenes* [27], M3Net achieves 80.9% mIoU and 83.1% mIoU on the validation and test sets, demonstrating its robustness and generalization capabilities. Additionally, the performance of M3Net on *Waymo Open* [90] is competitive with prior arts. We achieve a mIoU of 72.4% and a mAcc of 81.1%. These results highlight again the superiority of M3Net in handling complex diverse LiDAR segmentation tasks.

Direct Knowledge Transfer. To further validate the strong knowledge transfer capability of M3Net, we conduct extensive experiments on five different LiDAR-based perception datasets [41, 45, 77, 106, 107]. These datasets have unique data collection protocols and data distributions. As shown in Fig. 1 and the first ten columns in Tab. 3, our framework constantly outperforms the prior arts, the naïve joint training, and the single-dataset baselines across all five datasets. This concretely supports the strong knowledge transfer efficacy brought by multi-space alignments in M3Net.

Out-of-Distribution Generalization. Evaluating the generalization ability of models on out-of-training-distribution data is crucial, particularly in safety-critical fields like autonomous driving [50, 109, 110]. In this context, we resort to the two corruption datasets from the Robo3D [47] benchmark, *i.e.*, *SemanticKITTI-C* and *nuScenes-C*, to conduct our assessment. From the last four columns of Tab. 3, we observe that M3Net achieves better results than the naïve joint training and other single-dataset approaches, proving the strong generalizability of the learned representations.

5. Conclusion

In this work, we presented M3Net, a universal framework capable of fulfilling multi-task, multi-dataset, multi-modality LiDAR segmentation using a single set of parameters. Through extensive analyses, we validated the effectiveness of applying data-, feature-, and label-space alignments to handle such a challenging task. In addition, our comprehensive analysis and discourse have delved into the fundamental challenges of acquiring the general knowledge for scalable 3D perception, which holds substantial potential to propel further research in this domain. Our future strides focus on combining more data resources to further enhance the alignments and adaptations in our framework.

Acknowledgements. This work was partially supported by NSFC (No.62206173) and MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University). This work was also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5240–5250, 2023. 1, 2, 8

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312, 2022. 3

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 2, 3, 4, 6, 7, 8

[4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Jürgen Gall, and Cyrill Stachniss. Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The semantickitti dataset. *International Journal of Robotics Research*, 40:959–96, 2021. 2, 3

[5] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018. 6

[6] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. Also: Automotive lidar self-supervision by occupancy estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13455–13465, 2023. 1, 2

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 2, 3, 4

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3

[9] Jun Cen, Shiwei Zhang, Yixuan Pei, Kun Li, Hang Zheng, Maochun Luo, Yingya Zhang, and Qifeng Chen. Cmdfusion: Bidirectional fusion network with cross-modality knowledge distillation for lidar semantic segmentation. *arXiv preprint arXiv:2307.04091*, 2023. 3

[10] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, 2023. 2

[11] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 2, 3, 4

[12] Runnan Chen, Xinge Zhu, Nenglun Chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Bridging language and geometric primitives for zero-shot point cloud segmentation. In *ACM International Conference on Multimedia*, pages 5380–5388, 2023. 2

[13] Tian Chen, Shijie An, Yuan Zhang, Chongyang Ma, Huayan Wang, Xiaoyan Guo, and Wen Zheng. Improving monocular depth estimation by leveraging structural awareness and complementary datasets. In *European Conference on Computer Vision*, pages 90–108, 2020. 2

[14] Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, and Davide Modolo. Scaledet: A scalable multi-dataset object detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7288–7297, 2023. 2

[15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 17864–17875, 2021. 2

[16] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2

[17] Huixian Cheng, Xianfeng Han, and Guoqiang Xiao. Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022. 2

[18] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. Af2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021. 2, 8

[19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[20] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2, 3, 4, 6, 7, 8

[21] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2005. 6

[22] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 6

[23] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. https://github.com/Pointcept/Pointcept, 2023. 6

[24] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, pages 207–222, 2020. 2, 8

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 4, 7

[26] Christoph Feichtenhofer, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems*, 2022. 3

[27] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7:3795–3802, 2022. 2, 3, 4, 6, 7, 8

[28] Biao Gao, Yancheng Pan, Chengkun Li, Sibo Geng, and Huijing Zhao. Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1

[29] Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, and David A. Ross. Dataseg: Taming a universal multi-dataset multi-task segmentation model. *arXiv preprint arXiv:2306.01736*, 2023. 2

[30] Yi Gu, Yuming Huang, Chengzhong Xu, and Hui Kong. Maskrange: A mask-classification model for range-view based lidar segmentation. *arXiv preprint arXiv:2206.12073*, 2022. 2

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[32] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13090–13099, 2021. 6, 7

[33] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2024. 2, 6

[34] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022. 8

[35] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 2

[36] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4977–4987, 2021. 1

[37] Juana Valeria Hurtado, Rohit Mohan, Wolfram Burgard, and Abhinav Valada. Mopt: Multi-object panoptic tracking. *arXiv preprint arXiv:2004.08189*, 2020. 7

[38] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[39] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12605–12614, 2020. 1, 2, 3

[40] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2023. 1, 2, 3

[41] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripallig. Rellis-3d dataset: Data, benchmarks and analysis. In *IEEE International Conference on Robotics and Automation*, pages 1110–1116, 2021. 6, 8

[42] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C. V. Jawahar. Universal semi-supervised semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 5259–5270, 2019. 2

[43] Dongwan Kim, Yi-Hsuan Tsai, Yumin Suh, Masoud Faraki, Sparsh Garg, Manmohan Chandraker, and Bohyung Han. Learning semantic segmentation from multiple datasets with label shifts. In *European Conference on Computer Vision*, pages 20–36, 2022. 2

[44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[45] Alexey Klokov, Di Un Pak, Aleksandr Khorin, Dmitry Yudin, Leon Kochiev, Vladimir Luchinskiy, and Vitaly

Bezuglyj. Daps3d: Domain adaptive projective segmentation of 3d lidar point clouds. *IEEE Access*, 11:79341–79356, 2023. 6, 8

[46] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023. 2, 8

[47] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023. 1, 2, 6, 8

[48] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023. 1, 2

[49] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023. 2

[50] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, 2023. 1, 8

[51] Lingdong Kong, Xiang Xu, Jun Cen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Calib3d: Calibrating model preferences for reliable 3d scene understanding. *arXiv preprint arXiv:2403.17010*, 2024. 2

[52] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17545–17555, 2023. 2, 8

[53] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2879–2888, 2020. 2

[54] Guangrui Li, Guoliang Kang, Xiaohan Wang, Yunchao Wei, and Yi Yang. Adversarially masking synthetic to mimic real: Adaptive noise injection for point cloud segmentation adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20464–20474, 2023. 2

[55] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11809–11818, 2022. 7

[56] Li Li, Hubert PH Shum, and Toby P. Breckon. Less is more: Reducing task and model complexity for 3d point cloud semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9361–9371, 2023. 2

[57] Rong Li, Raoul de Charette, and C. A. O. Anh-Quan. Coarse3d: Class-prototypes for contrastive learning in weakly-supervised 3d point cloud segmentation. In *British Machine Vision Conference*, 2022. 2

[58] Xin Li, Botian Shi, Yuenan Hou, Xingjiao Wu, Tianlong Ma, Yikang Li, and Liang He. Homogeneous multi-modal feature fusion and interaction for 3d object detection. In *European Conference on Computer Vision*, pages 691–707. Springer, 2022. 2

[59] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18857, 2022. 2

[60] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023. 2, 4

[61] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Optimizing lidar placements for robust driving perception in adverse conditions. *arXiv preprint arXiv:2403.17009*, 2024. 2

[62] Venice Erin Liong, Thi Ngoc Tho Nguyen, Sergi Widjaja, Dhananjai Sharma, and Zhuang Jie Chong. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*, 2020. 2, 8

[63] Minghua Liu, Yin Zhou, Charles R. Qi, Boqing Gong, Hao Su, and Dragomir Anguelov. Less: Label-efficient semantic segmentation for lidar point clouds. In *European Conference on Computer Vision*, pages 70–89, 2022. 2

[64] Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi Bai, Xinge Zhu, Yuexin Ma, Yikang Li, Yu Qiao, and Yuenan Hou. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023. 2, 4

[65] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, 2023. 2, 3, 8

[66] Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.0468*, 2021. 3, 8

[67] Zhijian Liu, Alexander Amini Haotian Tang, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, pages 2774–2781, 2023. 3

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7

[69] Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, and Yuexin Ma. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In *IEEE/CVF International Conference on Computer Vision*, pages 21674–21684, 2023. 2

[70] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 2023. 1

[71] Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Mask-based panoptic lidar segmentation for autonomous driving. *IEEE Robotics and Automation Letters*, 8(2):1141–1148, 2023. 2

[72] Panagiotis Meletis and Gijs Dubbelman. Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation. In *IEEE Intelligent Vehicles Symposium*, pages 1045–1050, 2018. 2

[73] Björn Michele, Alexandre Boulch, Gilles Puy, Tuan-Hung Vu, Renaud Marlet, and Nicolas Courty. Saluda: Surface-based automotive lidar unsupervised domain adaptation. *arXiv preprint arXiv:2304.03251*, 2023. 2

[74] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4213–4220, 2019. 2, 4, 8

[75] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. In *International Conference on 3D Vision*, pages 116–125, 2021. 2, 7

[76] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[77] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *IEEE Intelligent Vehicles Symposium*, pages 687–693, 2020. 6, 8

[78] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 3

[79] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Tai Wang, Xinge Zhu, and Yuexin Ma. Sam-guided unsupervised domain adaptation for 3d segmentation. *arXiv preprint arXiv:2310.08820*, 2023. 1, 2

[80] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Using a waffle iron for automotive point cloud semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 3379–3389, 2023. 2, 8

[81] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 3

[82] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020. 2

[83] Giulia Rizzoli, Francesco Barbato, and Pietro Zanuttigh. Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. *Technologies*, 10(4), 2022. 1

[84] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *European Conference on Computer Vision*, pages 586–602, 2022. 2

[85] Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Cola: Coarse-label multi-source lidar semantic segmentation for autonomous driving. *arXiv preprint arXiv:2311.03017*, 2023. 2

[86] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 2, 3, 8

[87] Alvari Seppänen, Risto Ojala, and Kari Tammi. 4denoisenet: Adverse weather denoising from adjacent point clouds. *IEEE Robotics and Automation Letters*, 8:456–463, 2022. 1

[88] Kshitij Sirohi, Rohit Mohan, Daniel Büscher, Wolfram Burgard, and Abhinav Valada. Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 2021. 7

[89] Louis Soum-Fontez, Jean-Emmanuel Deschaud, and François Goulette. Mdt3d: Multi-dataset training for lidar 3d object detection generalization. *arXiv preprint arXiv:2308.01000*, 2023. 1, 2

[90] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 2, 3, 4, 6, 7, 8

[91] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702, 2020. 2, 4, 8

[92] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019.

[93] Larissa T Triess, David Peter, Christoph B Rist, and J Marius Zöllner. Scan-based semantic segmentation of lidar point clouds: An experimental study. In *IEEE Intelligent Vehicles Symposium*, pages 1116–1121, 2020. 2

[94] Larissa T. Triess, Mariella Dreissig, Christoph B. Rist, and J. Marius Zöllner. A survey on deep domain adaptation for lidar perception. In *IEEE Intelligent Vehicles Symposium Workshop*, pages 350–357, 2021. 1

[95] Darren Tsai, Julie Stephany Berrio, Mao Shan, Eduardo Nebot, and Stewart Worrall. Ms3d: Leveraging multiple detectors for unsupervised domain adaptation in 3d object detection. *arXiv preprint arXiv:2304.02431*, 2023. 2

[96] Darren Tsai, Julie Stephany Berrio, Mao Shan, Eduardo Nebot, and Stewart Worrall. Ms3d++: Ensemble of experts for multi-source unsupervised domain adaption in 3d object detection. *arXiv preprint arXiv:2308.05988*, 2023. 2

[97] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2697–2707, 2022. 2

[98] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 2, 4

[99] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019. 2

[100] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *Advances in Neural Information Processing Systems*, 2023. 2

[101] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 3

[102] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In *European Conference on Computer Vision*, pages 179–195. Springer, 2022. 4

[103] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Advances in Neural Information Processing Systems*, 2022. 2, 7, 8

[104] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. *arXiv preprint arXiv:2308.09718*, 2023. 2

[105] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. In *Advances in Neural Information Processing Systems*, pages 11035–11048, 2022. 2

[106] Aoran Xiao, Jiaxing Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *AAAI Conference on Artificial Intelligence*, pages 2795–2803, 2022. 1, 2, 6, 8

[107] Aoran Xiao, Jiaxing Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El Saddik, Shijian Lu, and Eric Xing. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9382–9392, 2023. 6, 8

[108] Zeqi Xiao, Wenwei Zhang, Tai Wang, Chen Change Loy, Dahua Lin, and Jiangmiao Pang. Position-guided point cloud panoptic segmentation transformer. *arXiv preprint arXiv:2303.13509*, 2023. 2

[109] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023. 8

[110] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and analyzing bird's eye view perception robustness to corruptions. *Preprint*, 2023. 8

[111] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19, 2020. 1, 2, 4

[112] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021. 2, 8

[113] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Rui Zhang, Qingyuan Zhou, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *arXiv preprint arXiv:2403.10001*, 2024. 2

[114] Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *arXiv preprint arXiv:2312.04484*, 2023. 8

[115] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, pages 677–695, 2022. 3

[116] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2021. 4

[117] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. In *AAAI Conference on Artificial Intelligence*, pages 3231–3240, 2023. 2, 8

[118] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9253–9262, 2023. 1, 2

[119] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 2

[120] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *Advances in Neural Information Processing Systems*, pages 10326–10338, 2021. 2

[121] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020. 1, 2, 8

[122] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17619–17629, 2023. 2

[123] Yiming Zhao, Lin Bai, and Xinming Huang. Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4453–4458, 2021. 2

[124] Qiang Zhou, Yuang Liu, Chaohui Yu, Jingliang Li, Zhibin Wang, and Fan Wang. Lmseg: Language-guided multi-dataset segmentation. In *International Conference on Learning Representations*, 2022. 2

[125] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368, 2022. 2

[126] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7571–7580, 2022. 2

[127] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13194–13203, 2021. 7

[128] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021. 1, 2, 4, 8

[129] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021. 2

[130] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, 2023. 2