Building a Strong Pre-Training Baseline for Universal 3D Large-Scale Perception

Haoming Chen[1], Zhizhong Zhang[1†], Yanyun Qu[3], Ruixin Zhang[4], Xin Tan[1,2], Yuan Xie[1,2]
[1]East China Normal University, Shanghai, China
[2]Chongqing Institute of East China Normal University, Chongqing, China
[3]Xiamen University, Fujian, China
[4]Tencent Youtu Lab, Shanghai, China
chenhaomingbob@gmail.com, {zzzhang, xtan, yxie}@cs.ecnu.edu.cn,
yyqu@xmu.edu.cn, ruixinzhang@tencent.com
† Corresponding Author

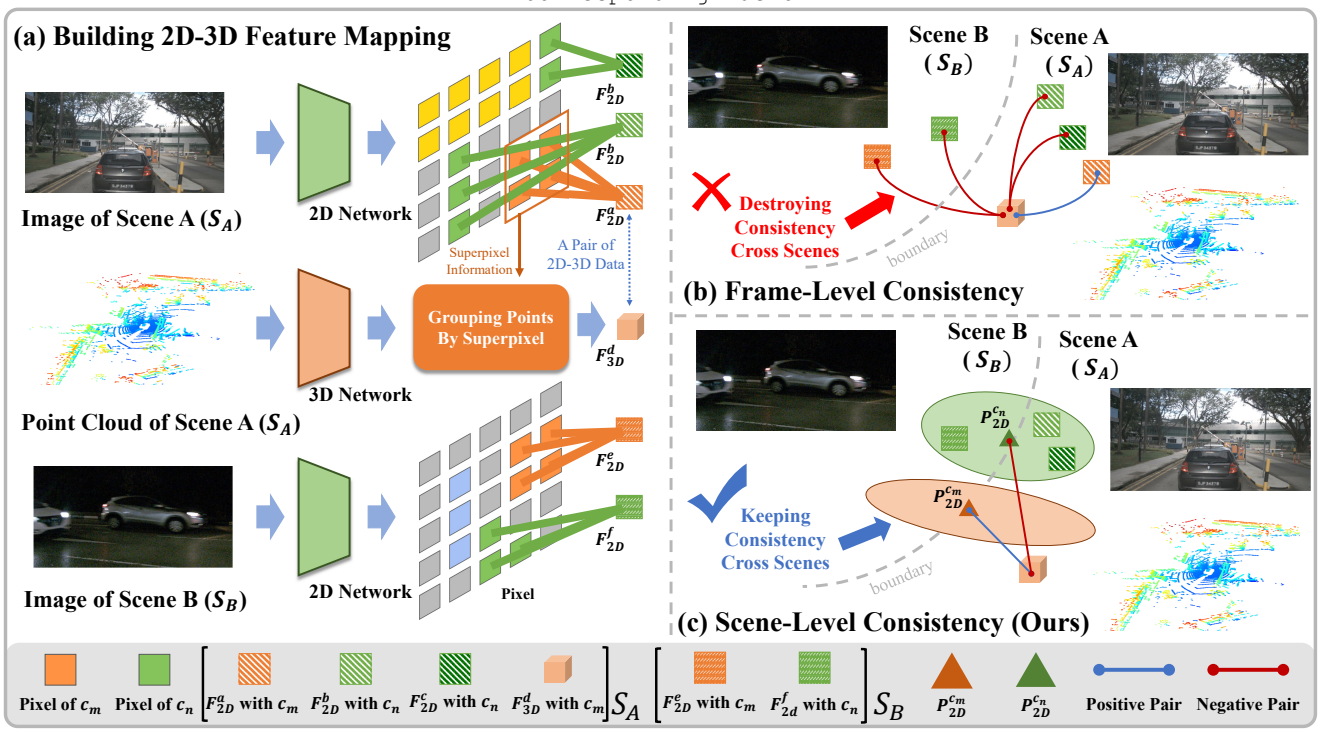

Figure 1. Brief illustration of the current multi-modality 3D pre-training paradigm and our proposed scene-level consistency. (a) we show the process of superpixel-superpoint association and clearly observe that superpixels with the same semantics can exist in the same scene or in different scenes, *e.g.*, the green squares. (b) We summarize existing pre-training methods and find that they all use frame-level consistency to learn 3D representations. Moreover, we believe that this constraint breaks the semantic consistency across views/frames and visualize the drawback. (c) We show our proposed scene-level consistency, which keeping the semantic consistency across various scenes. As a result, we achieve SOTA on three perceptual tasks with limited 3D annotation (Tab. 1). Our CSC framework builds a strong pre-training baseline for universal 3D Large-scale perception.

Abstract

An effective pre-training framework with universal 3D representations is extremely desired in perceiving large-scale dynamic scenes. However, establishing such an ideal framework that is both task-generic and label-efficient poses a challenge in unifying the representation of the same primitive across diverse scenes. The current contrastive 3D pre-training methods typically follow a frame-level consistency, which focuses on the 2D-3D relationships in each detached image. Such inconsiderate consistency greatly hampers the promising path of reaching an universal pre-training framework: (1) The cross-scene semantic self-conflict, i.e., the intense collision between primitive segments of the same semantics from different scenes; (2) Lacking a globally unified bond that pushes the cross-scene semantic consistency into 3D representation learning. To

address above challenges, we propose a CSC framework that puts a scene-level semantic consistency in the heart, bridging the connection of the similar semantic segments across various scenes. To achieve this goal, we combine the coherent semantic cues provided by the vision foundation model and the knowledge-rich cross-scene prototypes derived from the complementary multi-modality information. These allow us to train a universal 3D pre-training model that facilitates various downstream tasks with less fine-tuning efforts. Empirically, we achieve consistent improvements over SOTA pre-training approaches in semantic segmentation (+1.4% mIoU), object detection (+1.0% mAP), and panoptic segmentation (+3.0% PQ) using their task-specific 3D network on nuScenes. Code is released at https://github.com/chenhaomingbob/CSC, hoping to inspire future research.

1. Introduction

As one of the most promising applications of artificial intelligence, autonomous driving has undergone rapid development in recent years [15, 20, 21, 29, 51]. In it, 3D scene perception plays a fundamental role to perceive and understand surroundings, and thus has become increasingly attended in recent researches [14, 20, 21, 31]. Within the 3D perception, there are three essential tasks across different granularities, *i.e.,* semantic segmentation [12, 13, 23, 25, 27, 41, 45], object detection [9, 17, 19, 47], and panoptic segmentation [24, 54, 57]. All of them build upon powerful 3D representations and recent success of vision foundation model shows great potential for such fabulous goal [5, 32, 39, 40, 42].

Thanks to the paired image-lidar data captured by the multi-sensors [2, 4, 11, 44], it lays the foundation for improving 3D representation through cross-modality learning from 2D image priors. However, how VFM knowledge could benefit 3D scene perception without using any or limited point cloud annotations remains under-explored.

Inspired by the groundbreaking work SLidR [42], a succession of multi-modality 3D self-supervised approaches [32, 36, 39] has been proposed subsequently. To obtain desired 3D representation, all of them adopt a **frame-level consistency**, which conduct superpixel-superpoint contrastive distillation from the association between a point cloud frame and a image. We show the 2D-3D association in Fig. 1 (a). We observe that achieving the pre-trained 3D backbone with strong generalization ability still faces the following **two challenges**: **(1)** the superpixels sharing the identical semantic from the same or different images are erroneously treated as the negative pair. Despite the introduction of VFMs [32] or the adoption of semantically tolerant loss [36] as coping strategies, the challenge remains unresolved, especially in the presence of numerous identical semantic superpixels observed across various views or scenes. As presented in Fig. 1 (b), the cross-scene superpixels ($\boldsymbol{F}^{a}_{2D}$ & $\boldsymbol{F}^{e}_{2D}$) with the same semantic $\mathcal{C}^{m}_{sem}$ are not consistent in these methods. **(2)** The challenge of keeping global semantic consistency in large-scale scenarios. As an example, for objects with identical semantic labels but from different frames, their features should be as close as possible, whether the frames are from the same or disparate scenes.

In this paper, we introduce a strong pre-trained baseline, termed CSC (**C**oherent **S**emantic **C**ues Framework), for universal 3D large-scale perception by learning the **scene-level consistency** (as shown in Fig. 1 (c)). The core idea is that we push the cross-scenes semantic consistency into the pristine 3D backbone, leveraging the coherent semantic cues provided by powerful VFMs and information-rich semantic prototypes from multi-modality. Specifically, CSC consists of two key components: **(i)** A VFM-Assisted Semantic Prototype Generation, where we first utilize the VFM to provide reliable and coherent semantic cues for all superpixels across diverse scenes and then produce the multi-modality semantic prototypes to cover the representative features of involved semantics. **(ii)** A Coherent Semantic Consistency for 3D universal representation learning. In this component, we propose a multi-modality prototype blending module designed to fuse these unaligned prototypes that, while semantically aligned, reside in distinct feature spaces. We individually process each prototypes through a modality-specific prototype projection module, yielding implicitly aligned prototypes. Subsequently, these updated prototypes are combined and fed into a multi-modality prototype fusion module, resulting in the mixed prototypes that incorporate information from both image and lidar modalities. Ultimately, we perform a cross-scene semantic contrastive loss between superpoints and mixed prototypes, thereby obtaining the universal 3D representation with scene-level semantic consistency.

Compared to the current methods, the CSC brings a new SOTA performance to all three mainstream 3D perception tasks, semantic segmentation, object detection, and panoptic segmentation. This is particularly notable when only limited 3D annotations are available for a specific task. The main contributions of this work are summarized as follows:

- To the best of our knowledge, CSC is the first work to explore cross-scene semantic consistency in multi-modality 3D pre-training, which achieves the semantic consistency of all frames from all scenes.
- We utilize unexplored semantic cues from the VFM to maintain the coherent semantic prototypes including multi-modality information, resulting in the general pre-trained 3D backbone.
- Our CSC establishes a strong pre-training baseline for universal 3D large-scale perception, surpassing prior self-supervised approaches, remarkably evidenced by three

annotation-efficient downstream tasks, semantic segmentation improved by 1.4% mIoU, object detection by 1.0% mAP, and panoptic segmentation by 3% PQ.

2. Related Works

2.1. Vision Foundation Models

In light of the advancement of massive and diverse image sources and inspired by large-scale vision-language pre-training techniques, the computer vision community [10, 16, 18, 22, 28, 34, 38] is now witnessing hot attention in building powerful vision systems. Among these vision foundation models, DINOV2 and SAM have gathered the most widespread attention. The DINOv2 [38] is an advanced self-supervised learning framework that leverages vision transformers (with 1B parameters) and enough curated data sources (about 142M images) to producing high-performance visual features. The segment anything model (SAM) [18] demonstrates a new paradigm with robust zero-shot transferability, excelling in new image distributions and tasks. Apart from the mentioned works, approaches including SEEM [59] and OneFormer [16] also expand the visual model landscape, offering alternatives for the vision community. In this study, we explore the potential of VFMs for universal 3D scene perception. We leverage the VFM to acquire reliable and stable semantic cues across images from diverse scenes, harnessing these cues to promote global semantic consistency learning of 3D representations.

2.2. Self-Supervised 3D Representation Learning

Here, we focus on the line of contrastive-based self-supervised methods [3, 5, 32, 36, 37, 39, 40, 42, 43, 49]. According to the types of input modalities, these methods can be further categorized into uni-modality and multi-modality self-supervised framework. For uni-modality, they commonly perform the multi-view consistency constraint, which seek to the consistency of points/regions from various view transformations. TARL [37] exploits vehicle motion to extract different views of the same object in consecutive point cloud frames to learn spatio-temporal view-consistent. When we pay attention to the point cloud acquisition, we will find that almost most of the point cloud data will have its corresponding image data. For multi-modality, current studies not only consider 2D images but also involve text. In this paper, we primarily discuss the assistance of 2D images to advance 3D perception. SLidR [42] is the pioneering work that take superpixels obtained from SLIC [1] as units, and achieves superpixel-driven contrastive distillation to initialize the 3D network. Subsequently, Seal [32], the current SOTA method, introduces the popular VFM into this field and gains breakthrough performance improvements. Compared to these methods, we propose semantic prototype to manage the coherent semantic cues of all visual inputs, and enable scene-level semantic consistency to promote 3D representations.

2.3. Prototype-based Self-Supervised Learning

In 2D self-supervised realm, a wide range of applications adopts the idea of clustering or prototype. In the USL-VI-ReID [46], the existing SOTA methods [6] are developed on the ClusterContrast [8], which first generates prototypes using a clustering algorithm and then optimizes their networks via a cluster comparison mechanism. Meanwhile, in the field of unsupervised semantic segmentation, many excellent works [26, 35, 48] use the concept of prototypes and show inspiring performance. Inspired by these promising study, our CSC use prototypes for 3D pre-training. Through the prototypes, we could bridge the connection of various superpixels from different views or scenes.

3. Methodology

3.1. Preliminaries

To learn powerful 3D representations, the current 3D pre-training paradigm (stemming from SLidR [42]) utilizes the calibrated relationships between 2D images and 3D point clouds with the assistance of powerful 2D models. Technically, for a point cloud frame $\boldsymbol{\mathcal{P}} = \{\mathbf{P}_k \mid k = 1, \dots, K\}$ comprising K points, each point $\mathbf{P}_k \in \mathbb{R}^4$ represents the k-th point's three-dimensional location (x, y, z) coupled with its intensity feature. Meanwhile, the point cloud frame is equipped with L surrounding images $\boldsymbol{\mathcal{I}} = \{\mathbf{I}_l \mid l = 1, \dots, L\}$, where $\mathbf{I}_l \in \mathbb{R}^{H \times W \times 3}$ denotes l-th RGB image with the shape of $H \times W$.

In the pre-training phase, the images $\boldsymbol{\mathcal{I}}$ and points $\boldsymbol{\mathcal{P}}$ are respectively fed into a 2D embedding network, $\boldsymbol{\Theta_I}$: $\mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H \times W \times D}$, and a 3D embedding network, $\boldsymbol{\Theta_P}$: $\mathbb{R}^{K \times 4} \rightarrow \mathbb{R}^{K \times D}$, to produce pixel-wise features and point-wise features. Next, grouping these features via 2D masks $\boldsymbol{\mathcal{S}}_{\text{2D}}$ (*a.k.a*, superpixels introduced in [42]) obtained from a 2D segmentation algorithm, $\boldsymbol{\Theta_F}$: $\mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H \times W \times 1}$, we can harvest superpixel embeddings $\boldsymbol{\mathcal{F}}_{\text{2D}}$ and superpoint embeddings $\boldsymbol{\mathcal{F}}_{\text{3D}}$. Finally, the existing state-of-the-art methods [32, 36] will conduct frame-level superpixel-superpoint contrastive loss by utilizing the 2D-3D mapping, which pulls the matched superpixel-superpoint features while pushing away unmatched pairs, to optimize the 3D backbone $\boldsymbol{\Theta_P}$.

Problem Formulation. Extended on the paradigm, our goal is to build a general 3D self-supervised learning framework, that affords a wide range of downstream 3D perception tasks like semantic segmentation, object detection, and panoptic segmentation. We seek to perform scene-level consistency of 2D-3D elements relying on coherent semantic cues provided by coupling the recent popular VFM with our multi-modality semantic prototypes. Thus, our

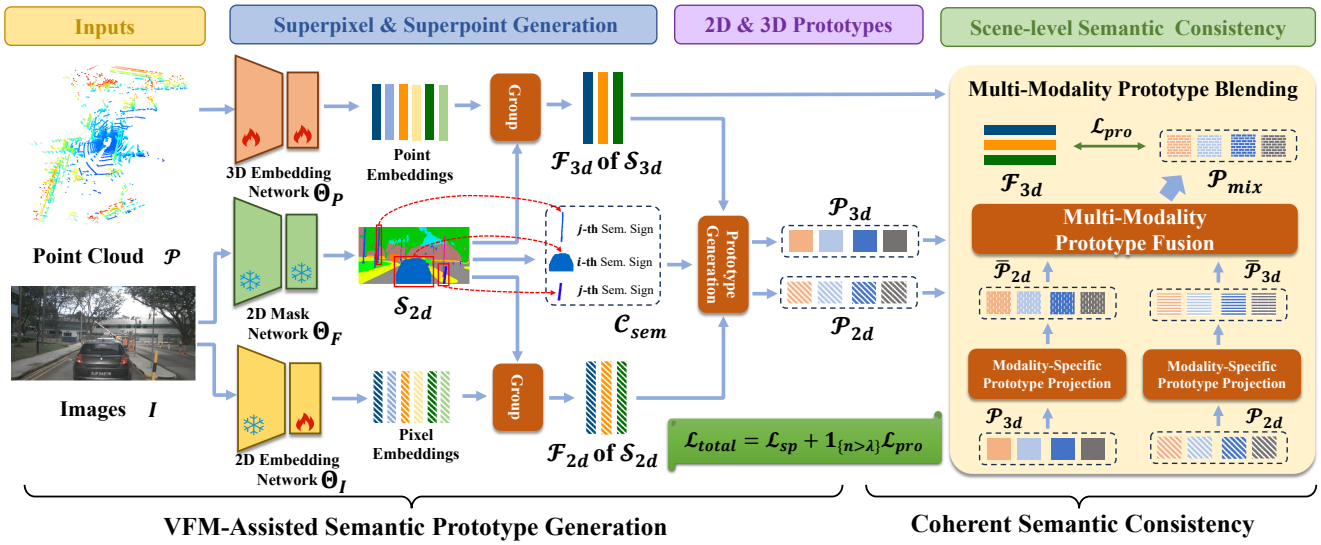

Figure 2. Overview of the CSC framework. CSC leverages the scene-level semantic consistency to obtain the universal 3D representations (Sec. 3), and then fine-tunes the pre-trained 3D backbone for three downstream perception tasks (Sec. 4). To achieve the scene-level semantic consistency, CSC consists of the VFM-assisted semantic prototype generation module (Sec. 3.2) and the coherent semantic consistency module (Sec. 3.3).

self-supervised objective is more explicit by replacing the frame-level consistency, which is prone to be to be influenced by spatio-temporal movement or scene changes.

Approach Overview. Our framework is outlined in Fig. 2. CSC framework consists of two key components: (1) a VFM-assisted semantic prototype generation covering semantic categories for large-scale cross-views/scenes (Sec. 3.2) and (2) a coherent semantic consistency between superpoints and prototypes for 3D representation learning (Sec. 3.3). The main differences from the current paradigm are that i) we leverage the VFM to obtain semantic-aware superpixels with cross-view/scene associations and ii) we propose multi-modality semantic prototypes to mine coherent semantic consistency for general 3D representation learning.

Formally, according to the VFM-assisted superpixels $\mathcal{S}_{2D}$, we obtain the semantic prototype features $\mathcal{F}_{2D}$ & $\mathcal{F}_{3D}$ for 2D and 3D data. Then, coupling these features with semantic cues $\mathcal{C}_{\text{sem}}$ from VFM, we maintain two separate modality prototypes $\mathcal{P}_{2D}$ & $\mathcal{P}_{3D}$, which represent coherent semantic representations for cross-scene objects with the same semantic. Moreover, to fully utilize two prototypes from heterogeneous space, we design a multi-modality prototype blending mechanism, which consists of modality-specific prototype projection and multi-modality prototype fusion modules, resulting in a mixed prototype $\mathcal{P}_{\text{mix}}$ containing rich multi-modality information. Upon on the $\mathcal{P}_{\text{mix}}$ and $\mathcal{F}_{3D}$, we derive a coherent semantic consistency loss $\mathcal{L}_{\text{pro}}$ to pushes close superpoint embeddings $\mathcal{F}_{3D}$ to the mixed prototypes $\mathcal{P}_{\text{mix}}$. Due to our cross-scene semantic prototypes, we achieve the scene-level semantic consistency for 3D representation learning. According to the observations of the universal improvement on three downstream tasks (Tab. 2), our scene-level semantic consistency constraint endows the generalization of 3D representations to various scene perception tasks.

3.2. VFM-Assisted Semantic Prototype Generation

Let us start by generating the semantic-aware superpixel $\mathcal{S}_{2D}$ from 2D vision foundation model, and then evolve two separate yet semantically aligned prototypes from complementary modalities.

Suerpixel & Superpoint Embeddings. Firstly, we utilize the pre-calibrated pose information to project each point $\mathbf{P}_k$ onto a camera image $\mathbf{I}_l$. Then, we leverage a VFM, DINOv2 [38] by default, to group visually similar regions into Q superpixels $\mathcal{S}_{2D} = \{S^q_{2D} \mid q = 1, \dots, Q\}$, where S^{2D}_q denotes the group of pixels belonging to the q-th superpixel. Combining the 2D-3D mapping and superpixels $\mathcal{S}_{2D}$, we can obtain the associated superpoints $\mathcal{S}_{3D} = \{S^q_{3D} \mid q = 1, \dots, Q\}$. Subsequently, pairing with pixel/point-wise features generated from the 2D/3D embedding networks, we are able to obtain superpixel and superpoint embeddings, $\mathcal{F}_{2D} = \{F^q_{2D} \mid q = 1, \dots, Q\}$ and $\mathcal{F}_{3D} = \{F^q_{3D} \mid q = 1, \dots, Q\}$, by averaging pooling the pixel and point features, where F^q_{2D} / F^q_{3D} is the q-th superpixel/superpoint embeddings.

Multi-Modality Prototype Generation. For all pairs of superpixels and superpoints from diverse scenes, we uniformly assign them with the semantic signs $\mathcal{C}_{\text{sem}}$ shared all scenes. The $\mathcal{C}_{\text{sem}}$ is obtained from the category-sensitive VFM, where we get the refined VFM on arbitrary se-

mantic segmentation benchmark. Subsequently, according to the $\mathcal{C}_{\text{sem}}$, we can group superpixel&superpoint embeddings $\mathcal{F}_{2D}$ & $\mathcal{F}_{3D}$ with the same semantic sign, to obtain the 2D&3D semantic prototype features, $\mathcal{P}_{2D} = \left\{ \boldsymbol{P}_{2D}^t \mid t = 1, \ldots, T \right\}$ & $\mathcal{P}_{3D} = \left\{ \boldsymbol{P}_{3D}^t \mid t = 1, \ldots, T \right\}$, by performing an averaging operation. The process of the multi-modality prototype generation can be expressed as follows:

$$\begin{aligned} \boldsymbol{P}_{2D}^t &= \frac{1}{|\mathcal{C}_{\text{sem}}^t|} \sum_{\boldsymbol{S}_{2D}^q = c^t} \boldsymbol{F}_{3D}^t, \\ \boldsymbol{P}_{3D}^t &= \frac{1}{|\mathcal{C}_{\text{sem}}^t|} \sum_{\boldsymbol{S}_{3D}^q = c^t} \boldsymbol{F}_{2D}^t, \end{aligned} \tag{1}$$

where $|\mathcal{C}_{\text{sem}}^t|$ is the count of the superpixels with the same semantic sign t. The total number of semantic signs is T. In our experiments, we use mask2former-based DINOV2 [38] as the mask network to generate $\mathcal{C}_{\text{sem}}$, where the network is fine-tuned on the ADE20K dataset [55] including $T = 150$ semantic classes.

Dicussion. From the pioneering work SLidR [42] to the amazing study Seal [32], the superpixels generation has transitioned from the non-learning segmentation algorithm (*i.e.*, SLIC [1] in [42]) to the category-insensitive VFM (*i.e.*, SAM [18] in [32]). The driving force behind improving segmentation algorithms stems from the annoying self-conflict challenge within contrastive-based self-supervised frameworks. To circumvent this impediment, Seal [32] first introduces category-insensitive VFMs (such as SAM [18] and SEEM [59]) to improve the quality of superpixels, significantly reducing the self-conflict issue between over segmentation and semantic consistency in each image. Attracted by the dramatic performance improvement from incorporating VFMs, we also adopt the powerful VFM and further devise our scene-level consistency pre-training framework. Compared to Seal, we exploit believable and consistent semantic cues provided by category-sensitive VFMs to alleviate self-conflict across all scenes.

3.3. Coherent Semantic Consistency

Based on the VFM-assisted semantic prototype, we propose a coherent semantic consistency to alleviate the challenge of cross-scenes self-conflict and conduct scene-level semantic regularization for 3D representation learning. To achieve an ideal 3D backbone, it is imperative to explore the information-rich multi-modality prototypes. However, although the two prototypes of different modalities have been semantically aligned, they do not lie in a uniform feature space. To reduce this gap, we design a multi-modality prototype blending module consisting of modality-specific prototype projection and multi-modality prototype fusion sub-modules. By parallel performing feature projection on each modality prototype followed by the fusion of multi-modality prototypes, the blending module will generate information-rich hybrid prototypes $\mathcal{P}_{\text{mix}}$. Considering both $\mathcal{P}_{\text{mix}}$ and $\mathcal{P}_{3D}$, we could achieve the scene-level semantic contrastive loss $\mathcal{L}_{\text{pro}}$ to endow the pre-trained 3D backbone with the ability to stable semantic discrimination on complex and dynamic large-scale autonomous driving scenes.In the next, we would illustrate each component in detail.

Multi-Modality Prototype Blending The MMPB module sequentially achieves feature alignment and fusion of heterogeneous modality prototypes via the modality-specific prototype projection and multi-modality prototype fusion modules. The two modules are defined as follows:

1. **Modality-Specific Prototype Projection.** Given the $\mathcal{P}_{2D}$ and $\mathcal{P}_{3D}$, several linear layers are employed in parallel to implicitly project the feature space of different modality to uniform one, resulting the updated 2D prototypes $\bar{\mathcal{P}}_{2D} = \left\{ \bar{\boldsymbol{P}}_{2D}^t \mid t = 1, \ldots, T \right\}$ and 3D prototypes $\bar{\mathcal{P}}_{3D} = \left\{ \bar{\boldsymbol{P}}_{3D}^t \mid t = 1, \ldots, T \right\}$. This computation can be expressed as:

$$\begin{aligned} \{\boldsymbol{P}_{2D}^1, \ldots, \boldsymbol{P}_{2D}^T\} &\xrightarrow{\text{Linear Layers}} \{\bar{\boldsymbol{P}}_{2D}^1, \ldots, \bar{\boldsymbol{P}}_{2D}^T\}, \\ \{\boldsymbol{P}_{3D}^1, \ldots, \boldsymbol{P}_{3D}^T\} &\xrightarrow{\text{Linear Layers}} \{\bar{\boldsymbol{P}}_{3D}^1, \ldots, \bar{\boldsymbol{P}}_{3D}^T\}. \end{aligned} \tag{2}$$

2. **Multi-Modality Prototype Fusion.** Then, we fuse the prototypes from two modality prototypes $\bar{\mathcal{P}}_{2D}$ & $\bar{\mathcal{P}}_{3D}$ that are both semantic category and feature space aligned, resulting in mixed prototypes $\bar{\mathcal{P}}_{\text{mix}} = \left\{ \bar{\boldsymbol{P}}_{\text{mix}}^t \mid t = 1, \ldots, T \right\}$. The computation is:

$$\{\bar{\boldsymbol{P}}_{2D}^1, \ldots, \bar{\boldsymbol{P}}_{2D}^T, \bar{\boldsymbol{P}}_{3D}^1, \ldots, \bar{\boldsymbol{P}}_{3D}^T\} \xrightarrow{\text{Linear Layers}} \left\{ \boldsymbol{P}_{\text{mix}}^1, \ldots, \boldsymbol{P}_{\text{mix}}^T \right\}. \tag{3}$$

Following the MMPB, we obtain the blended prototypes $\mathcal{P}_{\text{mix}}$ consist of the complementary multi-modality information. Thus, if using $\mathcal{P}_{\text{mix}}$, the resulting 3D backbone will equip the comprehensive discrimination from both 2D image modality and 3D lidar modality.

Prototype-based Loss. We propose a scene-level semantic contrastive loss $\mathcal{L}_{\text{proto}}$ between $\mathcal{P}_{3D}$ and $\mathcal{P}_{\text{mix}}$, to endow the pre-trained 3D backbone with the ability to coherence semantic discrimination on complex and dynamic large-scale autonomous driving scenes. Formally, the prototype-based contrastive loss $\mathcal{L}_{\text{pro}}$ is defined as follows:

$$\mathcal{L}_{\text{pro}} = -\log \frac{\exp\left(\left\langle \boldsymbol{F}_{3D}, \boldsymbol{P}_{\text{mix}}^{+} \right\rangle / \tau_{pro}\right)}{\sum_{i=0}^{|\mathcal{C}_{\text{sem}}|} \exp\left(\left\langle \boldsymbol{F}_{3D}, \boldsymbol{P}_{\text{mix}}^{i} \right\rangle / \tau_{pro}\right)}, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar production. The sign $\boldsymbol{P}_{\text{mix}}^{+}$ is the positive prototype of superpoint embedding $\boldsymbol{F}_{3D}$. The symbol τ_{pro} is a temperature hyper-parameter.

Discussion. Here, we argue that using coherent semantic cues from VFM is much better than the commonly adopted traditional cluster algorithm in the benefits for multi-modality prototypes fusion. Mostly self/unsupervised methods [6, 53] leverage an unsupervised clustering algorithm, such as K-Means and DBSCAN, to produce the their

prototypes. However, these methods all face a common challenge that requires manually adjusting clustering parameters based on the distribution of a specific dataset. In addition to the problem of hand-crafted parameters, there existing an other tricky challenge in the multi-modality self-supervised pre-training task, which is the alignment problem across different modalities. Fortunately, the introduction of class-sensitive VFM is able to bypass the above trouble challenges without any additional effort.

3.4. Loss Functions

Overall, our pre-training framework consists of two losses. **(1)** We utilize the common superpixel-superpoint contrastive loss $\mathcal{L}_{\text{sp}}$ [42] to optimize our backbones:

$$\mathcal{L}_{\text{sp}} = -\sum_{i=0}^{Q} \log \frac{\exp\left(\left\langle \boldsymbol{F}_{\text{3D}}^{i}, \boldsymbol{F}_{\text{2D}}^{i} \right\rangle / \tau_{sp}\right)}{\sum_{j=0}^{Q} \exp\left(\left\langle \boldsymbol{F}_{\text{3D}}^{i}, \boldsymbol{F}_{\text{2D}}^{j} \right\rangle / \tau_{sp}\right)}, \quad (5)$$

where i-th superpoint feature $\boldsymbol{F}_{\text{3D}}^{i}$ and i-th superpixel feature $\boldsymbol{F}_{\text{2D}}^{i}$ are matched according to the calibration between point cloud frames and the related surround images. **(2)** We leverage the proposed prototype-based loss $\mathcal{L}_{\text{pro}}$ to provide the global semantic consistency for 3D representation learning. Our total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sp}} + \mathbf{1}_{\{n>\lambda\}} \mathcal{L}_{\text{pro}}, \quad (6)$$

where the indicator of $\mathbf{1}_{\{n>\lambda\}}$ takes the value 1 if $n > \lambda$ and 0 otherwise, where n is the current training epoch and λ is the hyper-parameter that controls the starting epoch of using $\mathcal{L}_{\text{pro}}$. By default, $\lambda = 5$, $\tau_{sp} = 0.07$, and $\tau_{pro} = 1.0$.

4. Experiments

In this section, we present the experimental results of three different 3D perception tasks, each implemented by the popular 3D backbone of its domain. Specifically, semantic segmentation implemented by MinkUNet [7] in Sec. 4.1, object detection implemented by VoxelNet [56] in Sec. 4.2, and panoptic segmentation implemented by Cylinder3D [58] in Sec. 4.3. Conveniently, we draw Tab. 1 to show the comprehensive comparison of CSC with existing methods on three perception tasks with limited labeling. In addition, we study the role of each component in Sec. 4.4. Due to the limited space, visualization and other experiments would be shown in the supplementary materials.

Datasets. We pre-train all three models on nuScenes dataset, which is a large-scale autonomous driving dataset including 1,400,000 camera images as well as 90,00 Lidar sweeps across 1000 scenes. On nuScenes dataset, each point cloud keyframe is equipped with six calibrated surround images. During pre-training phase, we use the unlabeled RGB images and point clouds from 600 scenes to update backbones in our CSC, same as SLidR. About fine-tuning on three 3D perception tasks, we all conduct experiments on nuScenes, to evaluate the quality of pre-trained

Method & Year	**Semantic Segmentation**	
	1% (mIoU)	5% (mIoU)
MinkUNet		
Random Init.	30.3	47.7
SLidR, 22 [42]	38.2	52.2
ST-SLidR, 23 [36]	40.7	54.6
TriCC, 23 [39]	41.2	54.1
Seal, 23 [32]	45.8	55.6
Ours	**47.0** (+1.2 mIoU)	**57.0** (+1.4 mIoU)
Method	**Object Detection**	
	5% (mAP / NDS)	20% (mAP / NDS)
VoxelNet + CenterPoint		
Random Init.	38.0 / 44.3	50.2 / 59.7
SLidR, 22 [42]	43.3 / 52.4	50.4 / 59.9
TriCC, 23 [39]	44.6 / **54.4**	50.9 / **61.3**
Ours	**45.3 / 54.2** (+0.9 mAP)	**51.9 / 61.3** (+1.0 mAP)
Method	**Panoptic Segmentation**	
	1% (PQ / SQ / RQ)	5% (PQ / SQ / RQ)
Cylinder3D + Panoptic-PolarNet		
Random Init.	15.3 / 62.6 / 20.4	20.9 / 73.4/26.5
SLidR, 22 [42]	16.3 / 65.7 / 21.4	21.6 / 73.5 / 27.1
Ours	**19.3 / 74.5 / 24.6** (+3.0 PQ)	**23.1 / 76.9 / 28.5** (+1.5 PQ)

Table 1. On nuScenes, CSC is compared with current state-of-the-art methods in three downstream tasks with limited annotation. Obvious improvement in term of semantic segmentation, object detection, and panoptic segmentation could be found.

3D backbone with various percentage annotations. The nuScenes dataset is also used in our fine-tuning for three perception task, to evaluate the annotation-efficient of the various pre-training methods under different percentages of labeling.

Pre-training Details. Due to variances in network architectures, various 3D networks require different configurations in pre-training. For MinkUNet, we use the SGD optimizer with the 2.0 initial learning rate and a cosine annealing learning rate scheduler with a total of 50 epochs. The pre-training configuration of VoxelNet is similar to that of MinkUNet, the difference is that the initial learning rate is 0.01. As for Cylinder3D, we use the Adam optimizer of 0.001 initial learning rate and also employ the cosine annealing learning rate scheduler with a total of 15 epochs. All pre-trained 3D backbones are done with 2 RTX A6000 with a batch size 16.

4.1. Annotation-Efficient Semantic Segmentation

In this section, we measure the information of semantics learned by the 3D representations using various self-supervised frameworks. Overall, we compared CSC with the state-of-the-art methods on two benchmark datasets. In details, we evaluate the fine-tuned semantic segmentation performance of pre-trained 3D backbone on nuScenes [4] and SemanticKITTI [2] datasets.

Following SLidR [42], we fine-tune the pre-trained 3D backbone using various percentage point cloud subsets with 1%, 5%, 10%, 25%, and 100% of annotations for nuScenes and 1% for SemanticKITTI. Meanwhile, we conduct a linear evaluation using 100% annotations, which trains only

Method	Venue	nuScenes						KITTI
		LP	1%	5%	10%	25%	100%	1%
Random Init.	N/A	8.1	30.3	47.7	56.6	64.8	74.2	39.5
Point Con. [50]	ECCV 2020	21.9	32.5	-	57.1	-	74.3	41.1
Depth Con. [52]	ICCV 2021	22.1	31.7	-	57.3	-	74.1	41.5
PPKT [33]	arXiV 2021	35.9	37.8	51.7	59.2	66.8	73.8	44.0
SLidR [42]	CVPR 2022	38.0	38.2	52.2	58.8	66.2	74.6	44.6
ST-SLidR [36]	CVPR 2023	40.4	40.7	54.6	60.7	67.7	75.1	44.7
TriCC [39]	CVPR 2023	38.0	41.2	54.1	60.4	67.6	75.6	45.9
Seal [32]	NeurIPS 2023	44.9	45.8	55.6	62.9	68.4	75.6	46.6
Ours	-	**46.0** (+1.1)	**47.0** (+1.2)	**57.0** (+1.4)	**63.3** (+0.4)	**68.6** (+0.2)	**75.7** (+0.1)	**47.2** (+0.6)

Table 2. Results (mIoU) of different pre-training methods on semantic segmentation fine-tuning. On nuScenes, we use 100% annotated scans for linear probing and 1%, 5%, 10%, 25%, 100% annotation for fine-tuning. In addition, we use 1% labels for fine-tuning on SemanticKITTI.

a linear head and freezes other layers of the 3D backbone, to investigate the generalizability of representations learned via self-supervised learning without task-specific fine-tuning. We report the metric of mean Iou (mIoU) to evaluate various methods.

In Tab. 2, we show the comparison of the previous methods with CSC. It is evident that the 3D backbone with pre-trained parameters derived from arbitrary 3D self-supervised pre-training framework substantially outperforms the random initialized one. Compared with the current state-of-the-art method Seal [32] on nuScenes, our CSC provides significant mIoU improvements of +1.1% for linear probing, +1.2% and +1.4% for 1% and 5% few-shot fine-tuning settings, respectively. In addition, CSC also achieves better generalization of +0.6% boosting on the out-of-distribution annotation-efficient semantic segmentation in the SemanticKITTI. Compared to Seal, who only utilizes a class-insensitive VFM for each individual image to alleviate the self-conflict problem, our CSC presents a better 3D network with strong discriminative power. This suggests the importance of embracing the coherent semantic cues from the class-sensitive VFM and the scene-level semantic consistency in the pre-training phase. Due to page limitations, we show the average per-class performance of 1% annotation for detailed analysis in supplementary materials.

4.2. Annotation-Efficient Object Detection

In the vision system of autonomous driving, 3D object detection is a common and challenging task. Thus, we further evaluate the quality of our pre-trained lidar representation on this object-level task on the nuScenes. Following the previous works, we fine-tune the pre-trained 3D backbone with 5%, 10%, and 20% of the labeled data, respectively. Moreover, we embed the pre-trained Cylinder3D into two detection models, CenterPoint and SECOND. We refer to the evaluation protocol of nuScenes [4] and report the mean average precision (mAP) and nuScenes detection score (NDS), where NDS is a weighted average of mAP that measures the quality of the detection in various terms.

Method	nuScenes					
	5%		10%		20%	
	mAP	NDS	mAP	NDS	mAP	NDS
VoxelNet + CenterPoint						
Random Init.	38.0	44.3	46.9	55.5	50.2	59.7
Point Con. [50]	39.8	45.1	47.7	56.0	-	-
GCC-3D [30]	41.1	46.8	48.4	56.7	-	-
SLidR [42]	43.3	52.4	47.5	56.8	50.4	59.9
TriCC [39]	44.6	**54.4**	48.9	58.1	50.9	60.3
Ours	**45.3**	54.2	**49.3**	**58.3**	**51.9**	**61.3**
VoxelNet + SECOND						
Random Init.	35.8	45.9	39.0	51.2	43.1	55.7
SLidR [42]	36.6	48.1	39.8	52.1	44.2	56.3
TriCC [39]	37.8	**50.0**	41.4	53.5	45.5	57.7
Ours	**38.2**	49.4	**42.5**	**54.8**	**45.6**	**58.1**

Table 3. Results (mAP and NDS) when fine-tuning the pre-trained backbones to object detection using two models (CenterPoint and SECOND) with 5%, 10%, and 20% labels on nuScenes.

In Tab. 3, we compare the existing methods with our CSC. Compared to SLidR, the current SOTA method TriCC has gain the significantly improvement of 1.3% and 1.2% mAP in 5% mAP annotations by using the temporal consistency loops on both detection models. Taking the excellent work TriCC as the comparison, our CSC achieves the improvement of 0.7% mAP and 0.4% mAP without explicit temporal consistency. Surprisingly, the growth in performance by our CSC is steady, even using more annotation.

4.3. Annotation-Efficient Panoptic Segmentation

In this experiment, we compare the various pre-training methods for panoptic segmentation, which evaluates both semantic and instance recognition ability of the learned 3D backbone. To our knowledge, CSC is the first pre-training framework that transferring the pre-trained 3D backbone, which absorbs the prior knowledge from 2D realm, to the more challenging and annotation-intensive panoptic segmentation. Considering fair comparisons, we refer to both the setting from the previous lidar-only pre-training method

Method	nuScenes					
	1%			5%		
	PQ	SQ	RQ	PQ	SQ	RQ
Random Init.	15.3	62.6	20.4	20.9	73.4	26.5
SLidR + SLIC	16.3	65.7	21.4	21.6	73.5	27.1
SLidR + DINOV2	17.6	70.7	22.7	22.3	75.1	27.8
Ours (default)	19.3	74.5	24.6	23.1	**76.9**	28.5
Ours + OneFormer	**19.5**	**78.3**	**25.0**	**23.4**	75.7	**28.8**

Table 4. Results (PQ, SQ, and RQ) when fine-tuning the pre-trained models to panoptic segmentation with 1% and 5% labels on nuScenes. Considering the absence of existing pre-training methods, besides random initialization and the default CSC settings, we additionally establish three sets of experiments: original SLidR with SLIC, SLidR with DINOV2, and CSC with OneFormer.

[37] and the current state of development in the panoptic segmentation. Specifically, we employ the Panoptic-PolarNet [57] with Cylinder3D [58], just like the current supervised SOTA method [54]. About utilization rate of labels, we select the percentages of 1%, 5%, and 10% to fine-tune the pre-trained network. Given the absence of existing methods conducted on panoptic segmentation, we treat SLidR as the primary comparison approach and using different 2D segmentation methods for superpixel generation. Overal, there are five experiments: random initialization, pre-training by SLidR and SLIC, pre-training by SLidR and Dinov2, pre-training by CSC and Dinov2, and pre-training by CSC and OneFormer. About the evaluation metrics, we report segmentation quality (SQ), recognition quality (RQ), and panoptic quality (PQ).

Tab. 4 shows that multi-modality pre-training method is consistently better than the random initialization. With 1% annotation, replacing the superpixel generation method from SLIC to Dinov2 and further adopting our CSC can improve the PQ metric by 1.3% and 3.0%, respectively, compared to the original SLidR. On top of CSC, replacing DINOv2 by other semantic segmentation networks, such as OneFormer, will result in a considerable performance gain of 3.2%. Observing the changes in the SQ and RQ metrics of panorama segmentation across 1% and 5% ratios, we can find that the improvement brought from SLidR to our CSC is mainly in the SQ metrics (*i.e.*, $65.7 \rightarrow 74.5$ and $73.5 \rightarrow 76.9$) while the improvement in the RQ metrics is relatively slight (*i.e.*, $21.4 \rightarrow 24.7$ and $27.1 \rightarrow 28.5$). This phenomenon is consistent with previous results on semantic segmentation and object detection, *i.e.*, our approach significantly improves 3D network's ability to recognize semantic categories while providing a limited increase in the ability to discriminate different instances with the same semantics.

4.4. Ablation Study

We perform ablation experiments to examine the contribution of VFM for superpixel generation and multi-modality prototype blending (MMPB). We also investigate the impact of discarding MMPB to directly employ raw 3D prototypes for 3D backbone learning. All the ablation studies are conducted by 1%, 5%, and 10% semantic segmentation fine-tuning and 100% linear evaluation on nuScenes dataset.

#	SP	VFM	3D Pro.	MMPB	nuScene			
					LP	1%	5%	10%
(1)	✓				38.0	38.2	52.2	58.8
(2)	✓	✓			44.0	41.1	52.1	60.9
(3)	✓	✓	✓		44.0	40.3	53.3	60.5
(4)	✓	✓	✓	✓	**46.0**	**47.0**	**57.0**	**63.3**

Table 5. Ablation study of each component pre-trained and fine-tuned on nuScenes. **SP**: Superpixel-Superpoint contrastive loss $\mathcal{L}_{\text{sp}}$. **VFM**: Vision foundation models. **3D Pro.**: The mixed prototypes is replace by the raw 3D prototypes in the $\mathcal{L}_{\text{pro}}$. **MMPB**: Multi-Modality Prototype blending.

Compared to the original SLidR #(1) at 1% labeling, introducing our VFM #(2) and MMPB #(4) in turn, resulting in the steep rise of mIoU, from $38.2 \rightarrow 41.4 \rightarrow 47.0$. This upward trend persists across various annotation ratios, however, it tends to decelerate with the increase in the number of labels. Interestingly, when comparing #(3) with #(4), we can observe that directly employing the 3D semantic prototypes assisted by the VFM for 3D representation learning results in a performance deterioration of 0.8% mIoU, yet incorporating our proposed MPPB reverses this degradation and gains the promotion of 6.7%.

5. Conclusion

In this paper, we study the multi-modality 3D pre-training task from the drawbacks of the existing methods, including the self-conflict of cross-scene semantic segments and the absence of building the global semantic units. Our CSC performs scene-level semantic consistency via the combination VFM-assisted semantic cues and multi-modality semantic prototypes. Firstly, obtain the coherent semantic superpixels based on the VFM and use the semantics to generate prototypes for two modalities. Then, compute the unified prototypes by modality-specific prototype projection with multi-modality prototype blending. Thereby, we can achieve the cluster contrastive loss between 3D superpoint features and mixed prototypes for learning universal 3D representation. Extensive experiments show that our method delivers state-of-the-art results on semantic segmentation, object detection, and panoptic segmentation.

Acknowledgment. This work is supported by the National Natural Science Foundation of China No.62302167, U23A20343, 62222602, 62206293, 62176224, 62106075, and 62006139, Shanghai Sailing Program (23YF1410500), Natural Science Foundation of Shanghai (23ZR1420400), Science and Technology Commission of Shanghai No.21511100700, Natural Science Foundation of Chongqing, China (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137), CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (RAGR20230121), CAAI-Huawei MindSpore Open Fund, and Development Project of Ministry of Industry and Information Technology (Grant Number: ZTZB.23-990-016).

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3, 5

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 2, 6

[3] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. Also: Automotive lidar self-supervision by occupancy estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13455–13465, 2023. 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6, 7

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 2, 3

[6] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 3667–3675, New York, NY, USA, 2023. Association for Computing Machinery. 3, 5

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 6

[8] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, pages 1142–1160, 2022. 3

[9] Chengjian Feng, Zequn Jie, Yujie Zhong, Xiangxiang Chu, and Lin Ma. Aedet: Azimuth-invariant multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21580–21588, 2023. 2

[10] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17131–17141, 2023. 3

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[12] Jingyu Gong, Jiachen Xu, Xin Tan, Haichuan Song, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11673–11682, 2021. 2

[13] Jingyu Gong, Jiachen Xu, Xin Tan, Jie Zhou, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Boundary-aware geometric encoding for semantic segmentation of point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1424–1432, 2021. 2

[14] Jingyu Gong, Fengqi Liu, Jiachen Xu, Min Wang, Xin Tan, Zhizhong Zhang, Ran Yi, Haichuan Song, Yuan Xie, and Lizhuang Ma. Optimization over disentangled encoding: Unsupervised cross-domain point cloud completion via occlusion factor manipulation. In *European Conference on Computer Vision*, pages 517–533. Springer, 2022. 2

[15] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 2

[16] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. 2023. 3

[17] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21643–21652, 2023. 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 3, 5

[19] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13343–13353, 2023. 2

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023. 2

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

[22] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu. En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9315, 2022. 3

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

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11809–11818, 2022. 2

[25] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21694–21704, 2023. 2

[26] Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Chang Liu, Li Yuan, and Jie Chen. Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2023. 3

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

[28] Xiaofan Li, Yachao Zhang, Shiran Bian, Yanyun Qu, Yuan Xie, Zhongchao Shi, and Jianping Fan. Vs-boost: boosting visual-semantic association for generalized zero-shot learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1107–1115, 2023. 3

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

[30] Hanxue Liang, Chenhan Jiang, Dapeng Feng, Xin Chen, Hang Xu, Xiaodan Liang, Wei Zhang, Zhenguo Li, and Luc Van Gool. Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3293–3302, 2021. 7

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

[32] 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. *arXiv preprint arXiv:2306.09347*, 2023. 2, 3, 5, 6, 7

[33] 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:2104.04687*, 2021. 7

[34] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 525–534, 2021. 3

[35] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *ICML*, 2023. 3

[36] Anas Mahmoud, Jordan SK Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, and Steven L Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7102–7110, 2023. 2, 3, 6, 7

[37] Lucas Nunes, Louis Wiesmann, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Temporal consistent 3d lidar representation learning for semantic perception in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5228, 2023. 3, 8

[38] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3, 4, 5

[39] Bo Pang, Hongchi Xia, and Cewu Lu. Unsupervised 3d point cloud representation learning by triangle constrained contrast for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5229–5239, 2023. 2, 3, 6, 7

[40] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 2, 3

[41] Luigi Riz, Cristiano Saltori, Elisa Ricci, and Fabio Poiesi. Novel class discovery for 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9393–9402, 2023. 2

[42] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, pages 9891–9901, 2022. 2, 3, 5, 6, 7

[43] Corentin Sautier, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Vincent Lepetit. Bevcontrast: Self-supervision in bev space for automotive lidar point clouds. *arXiv preprint arXiv:2310.17281*, 2023. 3

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

[45] Xin Tan, Qihang Ma, Jingyu Gong, Jiachen Xu, Zhizhong Zhang, Haichuan Song, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Positive-negative receptive field reasoning for omni-supervised 3d segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15328–15344, 2023. 2

[46] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022. 3

[47] Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Frustumformer: Adaptive instance-aware resampling for multi-view 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5096–5105, 2023. 2

[48] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *Advances in Neural Information Processing Systems*, 35:16423–16438, 2022. 3

[49] Yanhao Wu, Tong Zhang, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Spatiotemporal self-supervised learning for point clouds in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5251–5260, 2023. 3

[50] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 7

[51] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 2

[52] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 7

[53] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17619–17629, 2023. 5

[54] Zhiwei Zhang, Zhizhong Zhang, Qian Yu, Ran Yi, Yuan Xie, and Lizhuang Ma. Lidar-camera panoptic segmentation via geometry-consistent and semantic-aware alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3662–3671, 2023. 2, 8

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 5

[56] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 6

[57] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13194–13203, 2021. 2, 8

[58] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021. 6, 8

[59] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 3, 5