CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP

Runnan Chen[1,2] Youquan Liu[2,3] Lingdong Kong[2,4] Xinge Zhu[5] Yuexin Ma[6] Yikang Li[2]
Yuenan Hou[2,†] Yu Qiao[2] Wenping Wang[7,†]

[1]The University of Hong Kong [2]Shanghai AI Laboratory [3]Hochschule Bremerhaven [4]National University of Singapore
[5]The Chinese University of Hong Kong [6]ShanghaiTech University [7]Texas A&M University

Abstract

Contrastive Language-Image Pre-training (CLIP) achieves promising results in 2D zero-shot and few-shot learning. Despite the impressive performance in 2D, applying CLIP to help the learning in 3D scene understanding has yet to be explored. In this paper, we make the first attempt to investigate how CLIP knowledge benefits 3D scene understanding. We propose CLIP2Scene, a simple yet effective framework that transfers CLIP knowledge from 2D image-text pre-trained models to a 3D point cloud network. We show that the pre-trained 3D network yields impressive performance on various downstream tasks, i.e., annotation-free and fine-tuning with labelled data for semantic segmentation. Specifically, built upon CLIP, we design a Semantic-driven Cross-modal Contrastive Learning framework that pre-trains a 3D network via semantic and spatial-temporal consistency regularization. For the former, we first leverage CLIP's text semantics to select the positive and negative point samples and then employ the contrastive loss to train the 3D network. In terms of the latter, we force the consistency between the temporally coherent point cloud features and their corresponding image features. We conduct experiments on SemanticKITTI, nuScenes, and ScanNet. For the first time, our pre-trained network achieves annotation-free 3D semantic segmentation with 20.8% and 25.08% mIoU on nuScenes and ScanNet, respectively. When fine-tuned with 1% or 100% labelled data, our method significantly outperforms other self-supervised methods, with improvements of 8% and 1% mIoU, respectively. Furthermore, we demonstrate the generalizability for handling cross-domain datasets. Code is publicly available[1].

1. Introduction

3D scene understanding is fundamental in autonomous driving, robot navigation, etc [26, 28]. Current deep learning-based methods have shown inspirational performance on 3D point cloud data [15, 32, 33, 38, 47, 56, 62]. However, some drawbacks hinder their real-world applications. The first one comes from their heavy reliance on the large collection of annotated point clouds, especially when high-quality 3D annotations are expensive to acquire [39, 40, 44, 51]. Besides, they typically fail to recognize novel objects that are never seen in the training data [11, 45]. As a result, it may need extra annotation efforts to train the model on recognizing these novel objects, which is both tedious and time-consuming.

Contrastive Vision-Language Pre-training (CLIP) [48] provides a new perspective that mitigates the above issues in 2D vision. It was trained on large-scale free-available image-text pairs from websites and built vision-language

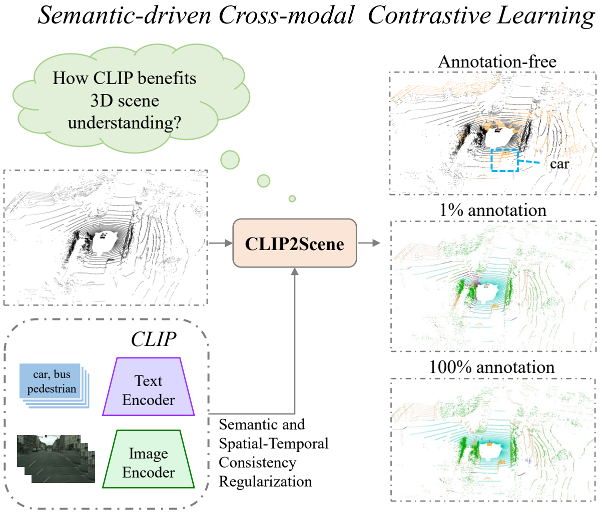
Semantic-driven Cross-modal Contrastive Learning
How CLIP benefits 3D scene understanding?
Annotation-free
car
1% annotation
100% annotation
CLIP2Scene
CLIP
car, bus pedestrian
Text Encoder
Image Encoder
Semantic and Spatial-Temporal Consistency Regularization

Figure 1. We explore how CLIP knowledge benefits 3D scene understanding. To this end, we propose CLIP2Scene, a Semantic-driven Cross-modal Contrastive Learning framework that leverages CLIP knowledge to pre-train a 3D point cloud segmentation network via semantic and spatial-temporal consistency regularization. CLIP2Scene yields impressive performance on annotation-free 3D semantic segmentation and significantly outperforms other self-supervised methods when fine-tuning on annotated data.

Symbol † denotes the corresponding authors.

[1] https://github.com/runnanchen/CLIP2Scene.

correlation to achieve promising open-vocabulary recognition. MaskCLIP [61] further explores semantic segmentation based on CLIP. With minimal modifications to the CLIP pre-trained network, MaskCLIP can be directly used for the semantic segmentation of novel objects without additional training efforts. PointCLIP [59] reveals that the zero-shot classification ability of CLIP can be generalized from the 2D image to the 3D point cloud. It perspectively projects a point cloud frame into different views of 2D depth maps that bridge the modal gap between the image and the point cloud. The above studies indicate the potential of CLIP on enhancing the 2D segmentation and 3D classification performance. However, whether and how CLIP knowledge benefits 3D scene understanding is still under-explored.

In this paper, we explore how to leverage CLIP's 2D image-text pre-learned knowledge for 3D scene understanding. Previous cross-modal knowledge distillation methods [44, 51] suffer from the optimization-conflict issue, *i.e.*, some of the positive pairs are regarded as negative samples for contrastive learning, leading to unsatisfactory representation learning and hammering the performance of downstream tasks. Besides, they also ignore the temporal coherence of the multi-sweep point cloud, failing to utilize the rich inter-sweep correspondence. To handle the mentioned problems, we propose a novel Semantic-driven Cross-modal Contrastive Learning framework that fully leverages CLIP's semantic and visual information to regularize a 3D network. Specifically, we propose Semantic Consistency Regularization and Spatial-Temporal Consistency Regularization. In semantic consistency regularization, we utilize CLIP's text semantics to select the positive and negative point samples for less-conflict contrastive learning. For spatial-temporal consistency regularization, we take CLIP's image pixel feature to impose a soft consistency constraint on the temporally coherent point features. Such an operation also alleviates the effects of imperfect image-to-point calibration.

We conduct several downstream tasks on the indoor and outdoor datasets to verify how the pre-trained network benefits the 3D scene understanding. The first one is annotation-free semantic segmentation. Following MaskCLIP, we place class names into multiple hand-crafted templates as prompts and average the text embeddings generated by CLIP to conduct the annotation-free segmentation. For the first time, our method achieves 20.8% and 25.08% mIoU annotation-free 3D semantic segmentation on the nuScenes [24] and ScanNet [20] datasets without training on any labelled data. Secondly, we compare with other self-supervised methods in label-efficient learning. When fine-tuning the 3D network with 1% or 100% labelled data on the nuScenes dataset, our method significantly outperforms state-of-the-art self-supervised methods, with improvements of 8% and 1% mIoU, respectively. Besides, to verify the generalization capability, we pre-train the network on the nuScenes dataset and evaluate it on SemanticKITTI [3]. Our method still significantly outperforms state-of-the-art methods. The key contributions of our work are summarized as follows.

- The first work that distils CLIP knowledge to a 3D network for 3D scene understanding.
- We propose a novel Semantic-driven Cross-modal Contrastive Learning framework that pre-trains a 3D network via spatial-temporal and semantic consistency regularization.
- We propose a novel Semantic-guided Spatial-Temporal Consistency Regularization that forces the consistency between the temporally coherent point cloud features and their corresponding image features.
- For the first time, our method achieves promising results on annotation-free 3D scene segmentation. When fine-tuning with labelled data, our method significantly outperforms state-of-the-art self-supervised methods.

2. Related Work

Zero-shot Learning in 3D. The objective of zero-shot learning (ZSL) is to recognize objects that are unseen in the training set. Many efforts have been devoted to the 2D recognition tasks [1, 2, 4, 8, 21, 25, 34, 37, 41–43, 46, 54, 55, 58, 60], and few works concentrate on performing ZSL in the 3D domain [11, 16–18, 45]. [18] applies ZSL to 3D tasks, where they train PointNet [47] on "seen" samples and test on "unseen" samples. Subsequent work [16] addresses the hubness problem caused by the low-quality point cloud features. [17] proposes the triplet loss to boost the performance under the transductive setting, where the "unseen" class is observed and unlabeled in the training phase. [11] makes the first attempt to explore the transductive zero-shot segmentation for 3D scene understanding. Recently, some studies introduced CLIP into zero-shot learning. MaskCLIP [61] investigates the problem of utilizing CLIP to help the 2D dense prediction tasks and exhibits encouraging zero-shot semantic segmentation performance. PointCLIP [59] is the pioneering work that applies CLIP to 3D recognition and shows impressive performance on zero-shot and few-shot classification tasks. Our work takes a step further to investigate how the rich semantic and visual knowledge in CLIP can benefit the 3D semantic segmentation tasks.

Self-supervised Representation Learning. The purpose of self-supervised learning is to obtain a good representation that benefits the downstream tasks. The dominant approaches resort to contrastive learning to pre-train the network [7, 9, 9, 10, 12–14, 22, 23, 27, 30]. Recently, inspired by the success of CLIP, leveraging the pre-trained model of CLIP to the downstream tasks has raised the community's attention [35, 36, 49, 50, 57]. DenseCLIP [49] utilizes

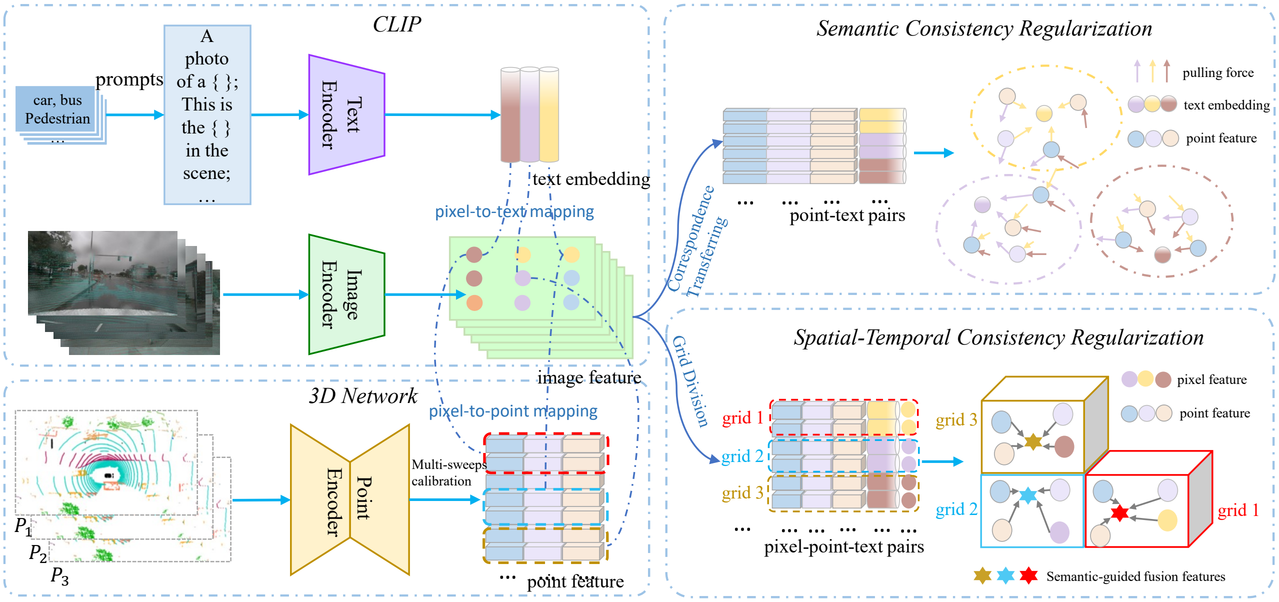

Figure 2. Illustration of the Semantic-driven Cross-modal Contrastive Learning. Firstly, we obtain the text embedding t_i, image pixel feature x_i, and point feature p_i by text encoder, image encoder, and point encoder, respectively. Secondly, we leverage CLIP knowledge to construct positive and negative samples for contrastive learning. Thus we obtain point-text pairs $\{x_i, t_i\}_{i=1}^{M}$ and all pixel-point-text pairs in a short temporal $\{\hat{x}_i^k, \hat{p}_i^k, t_i^k\}_{i=1,k=1}^{\hat{M},K}$. Here, $\{x_i, t_i\}_{i=1}^{M}$ and $\{\hat{x}_i^k, \hat{p}_i^k, t_i^k\}_{i=1,k=1}^{\hat{M},K}$ are used for Semantic Consistency Regularization and Spatial-Temporal Consistency Regularization, respectively. Lastly, we perform Semantic Consistency Regularization by pulling the point features to their corresponding text embedding and Spatial-Temporal Consistency Regularization by mimicking the temporally coherent point features to their corresponding pixel features.

the CLIP's pre-trained knowledge for dense image pixel prediction. DetCLIP [57] proposes a pre-training method equipped with CLIP for open-world detection. We leverage the image-text pre-trained CLIP knowledge to help 3D scene understanding.

Cross-modal Knowledge Distillation. Recently, increasing studies have focused on transferring knowledge from 2D images to 3D point clouds for self-supervised representation learning [44, 51]. PPKT [44] resorts to the InfoNCE loss to help the 3D network distil rich knowledge from the 2D image backbone. SLidR [51] further introduce the superpixel to boost the cross-modal knowledge distillation. In this paper, we first attempt to pre-train a 3D network with CLIP's knowledge.

3. Methodology

Considering the impressive open-vocabulary performance achieved by CLIP in image classification and segmentation, natural curiosities have been raised. Can CLIP endow the ability to a 3D network for annotation-free scene understanding? And further, will it promote the network performance when fine-tuned on labelled data? To answer the above questions, we study the cross-modal knowledge transfer of CLIP for 3D scene understanding, termed **CLIP2Scene**. Our work is a pioneer in exploiting CLIP knowledge for 3D scene understanding. In what follows, we revisit the CLIP applied in 2D open-vocabulary classification and semantic segmentation, then present our CLIP2Scene in detail. Our approach consists of three major components: Semantic Consistency Regularization, Semantic-Guided Spatial-Temporal Consistency Regularization, and Switchable Self-Training Strategy.

3.1. Revisiting CLIP

Contrastive Vision-Language Pre-training (CLIP) mitigates the following drawbacks that dominate the computer vision field: 1. Deep models need a large amount of formatted and labelled training data, which is expensive to acquire; 2. The model's generalization ability is weak, making it difficult to migrate to a new scenario with unseen objects. CLIP consists of an image encoder (ResNet [31] or ViT [6]) and a text encoder (Transformer [53]), both respectively project the image and text representation to a joint embedding space. During training, CLIP constructs positive and negative samples from 400 million image-text pairs to train both encoders with a contrastive loss, where the large-scale image-text pairs are free-available from the Internet and assumed to contain every class of images and most concepts of text. Therefore, CLIP can achieve promising open-vocabulary recognition.

For 2D zero-shot classification, CLIP first places the class name into a pre-defined template to generate the text embed-

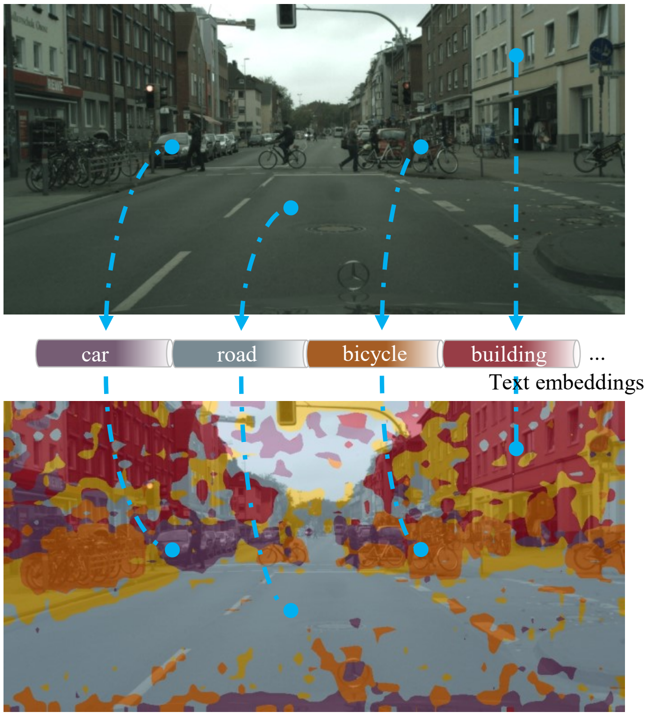

Figure 3. Illustration of the image pixel-to-text mapping. The dense pixel-text correspondence $\{x_i, t_i\}_{i=1}^{M}$ is extracted by the off-the-shelf method MaskCLIP [61].

dings and then encodes images to obtain image embeddings. Next, it calculates the similarities between images and text embeddings to determine the class. MaskCLIP further extends CLIP into 2D semantic segmentation. Specifically, MaskCLIP modifies the attention pooling layer of the CLIP's image encoder, thus performing pixel-level mask prediction instead of the global image-level prediction.

3.2. CLIP2Scene

As shown in Fig. 2, we first leverage CLIP and 3D network to respectively extract the text embeddings, image pixel feature, and point feature. Secondly, we construct positive and negative samples based on CLIP's knowledge. Lastly, we impose Semantic Consistency Regularization by pulling the point features to their corresponding text embedding. At the same time, we apply Spatial-Temporal Consistency Regularization by forcing the consistency between temporally coherent point features and their corresponding pixel features. In what follows, we present the details and insights.

3.2.1 Semantic Consistency Regularization

As CLIP is pre-trained on 2D images and text, our first concern is the domain gap between 2D images and the 3D point cloud. To this end, we build dense pixel-point correspondence and transfer image knowledge to the 3D point cloud via the pixel-point pairs. Specifically, we calibrate the LiDAR point cloud with corresponding images captured by six cameras. Therefore, the dense pixel-point correspondence $\{x_i, p_i\}_{i=1}^{M}$ can be obtained accordingly, where x_i and p_i indicates i-th paired image feature and point feature, which are respectively extracted by the CLIP's image encoder and the 3D network. M is the number of pairs. Note that it is an online operation and is irreverent to the image and point data augmentation.

Previous methods [44, 51] provide a promising solution to cross-modal knowledge transfer. They first construct positive pixel-point pairs $\{x_i, p_i\}_{i=1}^{M}$ and negative pairs $\{x_i, p_j\}(i \neq j)$, and then pull in the positive pairs while pushing away the negative pairs in the embedding space via the InfoNCE loss. Despite the encourageable performance of previous methods in transferring cross-modal knowledge, they are both confronted with the same optimization-conflict issue. For example, suppose i-th pixel x_i and j-th point p_j are in the different positions of the same instance with the same semantics. However, the InfoNCE loss will try to push them away, which is unreasonable and hammer the performance of the downstream tasks [51]. In light of this, we propose a Semantic Consistency Regularization that leverages the CLIP's semantic information to alleviate this issue. Specifically, we generate the dense pixel-text pairs $\{x_i, t_i\}_{i=1}^{M}$ by following the off-the-shelf method MaskCLIP [61] (Fig. 3), where t_i is the text embedding generated from the CLIP's text encoder. Note that the pixel-text mappings are free-available from CLIP without any additional training. We then transfer pixel-text pairs to point-text pairs $\{p_i, t_i\}_{i=1}^{M}$ and utilize the text semantics to select the positive and negative point samples for contrastive learning. The objective function is as follows:

$$\mathcal{L}_{S\_info} = -\sum_{c=1}^{C} \log \frac{\sum_{t_i \in c, p_i} \exp(D(t_i, p_i)/\tau)}{\sum_{t_i \in c, t_j \notin c, p_j} \exp(D(t_i, p_j)/\tau)}, \tag{1}$$

where $t_i \in c$ indicates that t_i is generated by c-th classes name, and C is the number of classes. D denotes the scalar product operation and τ is a temperature term ($\tau > 0$).

Since the text is composed of class names placed into pre-defined templates, the text embedding represents the semantic information of the corresponding class. Therefore, those points with the same semantics will be restricted near the same text embedding, and those with different semantics will be pushed away. To this end, our Semantic Consistency Regularization causes less conflict in contrastive learning.

3.2.2 Semantic-guided Spatial-temporal Consistency Regularization

Besides semantic consistency regularization, we consider how image pixel features help to regularize a 3D network.

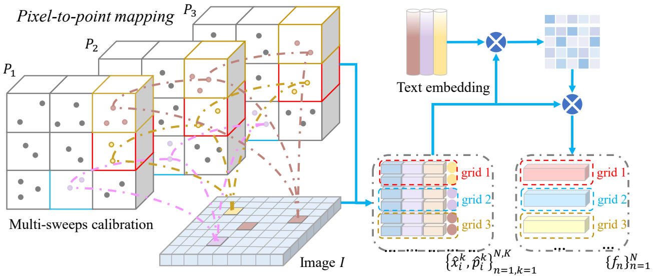

Figure 4. Illustration of the image pixel-to-point mapping (left) and semantic-guided fusion feature generation (right). We build the grid-wise correspondence between an image I and the temporally coherent LiDAR point cloud $\{P_k\}_{k=1}^{K}$ within S seconds and generate semantic-guided fusion features for individual grids. Both $\{\hat{x}_i^k, \hat{p}_i^k\}_{i=1,k=1}^{\hat{M},K}$ and $\{f_n\}_{n=1}^{N}$ are used to perform Spatial-Temporal Consistency Regularization.

The natural alternative directly pulls in the point feature with its corresponding pixel in the embedding space. However, the noise-assigned semantics of the image pixel and the imperfect pixel-point mapping hinder the downstream task's performance. To this end, we propose a novel semantic-guided Spatial-Temporal Consistency Regularization to alleviate the problem by imposing a soft constraint on points within local space and time.

Specifically, given an image I and temporally coherent LiDAR point cloud $\{P_k\}_{k=1}^{K}$, where K is the number of sweeps within S seconds. Note that the image is matched to the first frame of the point cloud P_1 with pixel-point pairs $\{\hat{x}_i^1, \hat{p}_i^1\}_{i=1}^{\hat{M}}$. We register the rest of the point cloud to the first frame via the calibration matrices and map them to the image (Fig. 4). Thus we obtain all pixel-point-text pairs in a short temporal $\{\hat{x}_i^k, \hat{p}_i^k, t_i^k\}_{i=1,k=1}^{\hat{M},K}$. Next, we divide the entire stitched point cloud into regular grids $\{g_n\}_{n=1}^{N}$, where the temporally coherent points are located in the same grid. We impose the spatial-temporal consistency constraint within individual grids by the following objective function:

$$\mathcal{L}_{SSR} = \sum_{g_n} \sum_{(\hat{i},\hat{k})\in g_n} (1 - \textbf{sigmoid}(D(\hat{p}_{\hat{i}}^{\hat{k}}, f_n)))/N, \quad (2)$$

where $(\hat{i}, \hat{k}) \in g_n$ indicates the pixel-point pair $\{\hat{x}_i^k, \hat{p}_i^k\}$ is located in the n-th grid. $\{f_n\}_{n=1}^{N}$ is a semantic-guided cross-modal fusion feature formulated by:

$$f_n = \sum_{(\hat{i},\hat{k})\in g_n} a_{\hat{i}}^{\hat{k}} * \hat{x}_{\hat{i}}^{\hat{k}} + b_{\hat{i}}^{\hat{k}} * \hat{p}_{\hat{i}}^{\hat{k}}, \quad (3)$$

where $a_{\hat{i}}^{\hat{k}}$ and $b_{\hat{i}}^{\hat{k}}$ are attention weight calculated by:

$$\begin{aligned} a_{\hat{i}}^{\hat{k}} &= \frac{\exp(D(\hat{x}_{\hat{i}}^{\hat{k}}, t_{\hat{i}}^{1})/\lambda)}{\sum_{(\hat{i},\hat{k})\in g_n} \exp(D(\hat{x}_{\hat{i}}^{\hat{k}}, t_{\hat{i}}^{1})/\lambda) + \exp(D(\hat{p}_{\hat{i}}^{\hat{k}}, t_{\hat{i}}^{1})/\lambda)}, \\ b_{\hat{i}}^{\hat{k}} &= \frac{\exp(D(\hat{p}_{\hat{i}}^{\hat{k}}, t_{\hat{i}}^{1})/\lambda)}{\sum_{(\hat{i},\hat{k})\in g_n} \exp(D(\hat{x}_{\hat{i}}^{\hat{k}}, t_{\hat{i}}^{1})/\lambda) + \exp(D(\hat{p}_{\hat{i}}^{\hat{k}}, t_{\hat{i}}^{1})/\lambda)}, \end{aligned} \quad (4)$$

where λ is the temperature term.

Actually, those pixel and point features within the local grid g_n are restricted near a dynamic centre f_n. Thus, such a soft constraint alleviates the noisy prediction and calibration error issues. At the same time, it imposes Spatio-Temporal Regularization on the temporally coherent point features.

3.2.3 Switchable Self-training Strategy

We combine the loss function $\mathcal{L}_{S\_info}$ and $\mathcal{L}_{SSR}$ to end-to-end train the whole network, where the CLIP's image and text encoder backbone are frozen during training. We find that method worked only when the pixel-point feature $\{x_i, p_i\}_{i=1}^{M}$ and $\{\hat{x}_i^k, \hat{p}_i^k\}_{i=1,k=1}^{\hat{M},K}$, which are used in $\mathcal{L}_{S\_info}$ and $\mathcal{L}_{SSR}$, are generated from different learnable linear layer. On top of that, we further put forward an effective strategy to promote performance. Specifically, after contrastive learning of the 3D network for a few epochs, we randomly switch the point pseudo label between the paired image pixel's pseudo label and the point's predicted label. Since different modality networks learn different feature representations, they can filter different types of error introduced by noisy pseudo labels. By this switchable operation, the error flows can be reduced by mutually [29].

4. Experiments

Datasets. We conduct extensive experiments on two large-scale outdoor LiDAR semantic segmentation datasets, *i.e.*, SemanticKITTI [3] and nuScenes [5, 24], and one indoor dataset ScanNet [20]. The nuScenes dataset contains 700 scenes for training, 150 scenes for validation, and 150 scenes for testing, where 16 classes are utilized for LiDAR semantic segmentation. As for SemanticKITTI, it contains 19 classes for training and evaluation. It has 22 sequences, where sequences 00 to 10, 08, and 11 to 21 are used for training, validation, and testing, respectively. ScanNet [20] contains 1603 scans with 20 classes, where 1201 scans are for training, 312 scans are for validation, and 100 scans are for testing.

Implementation Details. We follow SLidR [51] to pre-train the network on the nuScenes [5, 24] dataset. The network is pre-trained on all keyframes from 600 scenes. Besides, the pre-trained network is fine-tuned on SemanticKITTI [3] to verify the generalization ability. We

Table 1. Comparisons (mIoU) among self-supervised methods on the nuScenes [24], SemanticKITTI [3], and ScanNet [20] *val* sets.

Initialization	nuScenes		SemanticKITTI		ScanNet	
	1%	100%	1%	100%	5%	100%
Random	42.2	69.1	32.5	52.1	46.1	63.3
PPKT [44]	48.0	70.1	39.1	53.1	47.5	64.2
SLidR [51]	48.2	70.4	39.6	54.3	47.9	64.9
PointContrast [55]	47.2	69.2	37.1	52.3	47.6	64.5
CLIP2Scene	**56.3**	**71.5**	**42.6**	**55.0**	**48.4**	**65.1**

Table 2. Annotation-free 3D semantic segmentation performance (mIoU) on the nuScenes [24] and ScanNet [20] *val* sets.

Method	nuScenes	ScanNet
CLIP2Scene	20.80	25.08

leverage the CLIP model to generate image features and text embedding. Following MaskCLIP, we modify the attention pooling layer of the CLIP's image encoder, thus extracting the dense pixel-text correspondences. We take SPVCNN [52] as the 3D network to produce the point-wise feature. The framework is developed on PyTorch, where the CLIP model is frozen during training. The training time is about 40 hours for 20 epochs on two NVIDIA Tesla A100 GPUs. The optimizer is SGD with a cosine scheduler. We set the temperature λ and τ to be 1 and 0.5, respectively. The sweep number is set to be 3 empirically. Besides, We adopt MinkowskiNet14 [19] as the backbone for evaluation on the ScanNet dataset, where the number of sweeps is set to be 1 and the training epochs is 30. As for the Switchable Self-Training Strategy, we randomly switch the point supervision signal after 10 epochs. We apply several data augmentations in contrastive learning, including random rotation along the z-axis and random flip on the point cloud, random horizontal flip, and random crop-resize on the image.

4.1. Annotation-free Semantic Segmentation

After pre-training the network, we show the performance of the 3D network when it is not fine-tuned on any annotations (Table 2). As no previous method reports the 3D annotation-free segmentation performance, we compare our method with different setups (Table 3). In what follows, we describe the experimental settings and give insights into our method and the different settings.

Settings. We conduct experiments on the nuScenes and ScanNet datasets to evaluate the annotation-free semantic segmentation performance. Following MaskCLIP [61], we place the class name into 85 hand-craft prompts and feed it into the CLIP's text encoder to produce multiple text features. We then average the text features and feed the averaged features to the classifier for point-wise prediction. Besides, to explore how to effectively transfer CLIP's knowledge to the 3D network for annotation-free segmentation, We conduct the following experiments to highlight the effectiveness of different modules in our framework.

Table 3. Ablation study on the nuScenes [24] *val* set for annotation-free 3D semantic segmentation.

Ablation Target	Setting	mIoU (%)
-	Baseline	15.1
Prompts	nuScenes	15.1 (+0.0)
	SemanticKITTI	13.9 (−1.2)
	Cityscapes	11.3 (−3.8)
	All	15.3 (+0.2)
Regularization	w/o StCR	19.8 (+4.7)
	w/o SCR	16.8 (+1.7)
	KL	0.0 (−15.1)
Training Strategy	w/o S3	18.8 (+3.7)
	ST	10.1 (−4.0)
Sweeps	1 sweep	18.7 (+3.6)
	3 sweeps	20.8 (+5.7)
	5 sweeps	20.6 (+5.5)
	merged	18.6 (+3.5)
Full Configuration	CLIP2Scene	**20.8** (+5.7)

Baseline. The input of the 3D semantic segmentation network is only one sweep, and we pre-train the framework via semantic consistency regularization.

Prompts (nuScenes, SemanticKITTI, Cityscapes, All). Based on the baseline, we respectively replace the nuScenes, SemanticKITTI, Cityscapes, and all class names into the prompts to produce the text embedding.

Regularization (w/o StCR, w/o SCR, KL). Based on the full method, we remove the Spatial-temporal Consistency Regularization (w/o StCR) and remove the Semantic Consistency Regularization (w/o SCR). Besides, we abuse both StCR and SCR and distill the image feature to the point cloud by Kullback–Leibler (KL) divergence loss.

Training Strategies (w/o S3, ST). We abuse the Switchable Self-Training Strategy (w/o S3) in the full method. Besides, we show the performance of only training the 3D network by their own predictions after ten epochs (ST).

Sweeps Number (1 sweep, 3 sweeps, 5 sweeps, and merged). We set the sweep number K to be 1, 3, and 5, respectively. Besides, we also take three sweeps of the point cloud as the input to pre-train the network (merged).

Effect of Different Prompts. To verify how text embedding affects the performance, we generate various text embedding by the class name from different datasets (nuScenes, SemanticKITTI, and Cityscapes) and all classes for pre-training the framework. As shown in Table 3, we find that even learning with other datasets' text embedding (SemanticKITTI and Cityscapes), the 3D network could still recognize the nuScenes's objects with decent performance (13.9% and 11.3% mIoU, respectively). The result shows that the 3D network is capable of open-vocabulary recognition ability.

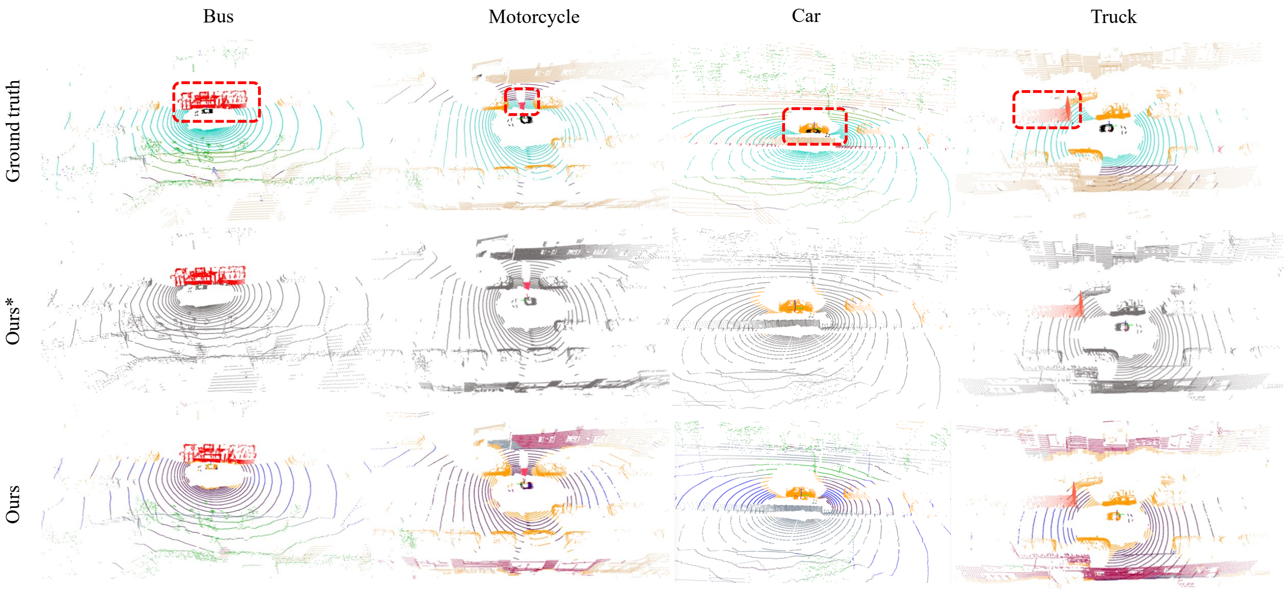

Figure 5. Qualitative results of annotation-free semantic segmentation on nuScenes dataset. Note that we show the results of individual classes. From the left to the right column are *bus*, *motorcycle*, *car*, and *truck*, respectively. The first row [ground truth] is the annotated semantic label. The second row [ours*] is our prediction of the highlighted target. The third row [ours] is our prediction of full classes.

Effect of Semantic and Spatial-temporal Consistency Regularization. We remove Spatial-temporal Consistency Regularization (w/o SCR) from our method. Experiments show that the performance is dramatically decreased, indicating the effectiveness of our design. Besides, we also distill the image feature to the point cloud by KL divergence loss, where the text embeddings calculate the logits. However, such a method fails to transfer the semantic information from the image. The main reason is the noise-assigned semantics of the image pixel and the imperfect pixel-point correspondence due to the calibration error.

Effect of Switchable Self-training Strategy. To examine the effect of the Switchable Self-Training Strategy, we either train the network with image supervision (w/o S3) or train the 3D network by their own predictions. Both trials witness a performance drop, indicating Switchable Self-Training Strategy is efficient in cross-modal self-supervised learning.

Effect of Sweep Numbers. Intuitively, the performance of our method benefits from more sweeps information. Therefore, we also show the performance when restricting sweep size to 1, 3, and 5, respectively. However, we observe that the performance of 5 sweeps is similar to 3 sweeps but is more computationally expensive. Thus, we empirically set the sweep number to be 3.

Qualitative Evaluation. The qualitative evaluations of individual classes (bus, motorcycle, car, and truck) are in Fig. 5, indicating that our method is able to perceive the objects even without training on any annotated data. However, we also observe the false positive predictions around the ground truth objects. We will resolve this issue in future work.

4.2. Annotation-efficient Semantic Segmentation

The pre-trained 3D network also boosts the performance when few labeled data are available for training. We directly compare SLidR [51], the only published method for image-to-Lidar self-supervised representation distillation. Besides, we also compared PPKT [44] and PointContrast [55]. In the following, we introduce SLidR and PPKT and compare them in detail.

PPKT. PPKT is a cross-modal self-supervised method for the RGB-D dataset. It performs 2D-to-3D knowledge distillation via pixel-to-point contrastive loss. For a fair comparison, we use the same 3D network and training protocol but replace our semantic and Spatio-Temporal Regularization with InfoNCE loss. The framework is trained 50 epochs on 4096 randomly selected image-to-point pairs.

SLidR. SLidR is an image-to-Lidar self-supervised method for autonomous driving data. Compared with PPKT, it introduces image super-pixel into cross-modal self-supervised learning. For a fair comparison, we replace our loss function with their superpixel-driven contrastive loss.

Performance. As shown in Table 1, our method significantly outperforms the state-of-the-art methods when fine-tuned on 1% and 100% nuScenes dataset, with the improvement of 8.1% and 1.1%, respectively. Compared with the random initialization, the improvement is 14.1% and 2.4%, respectively, indicating the efficiency of our Semantic-driven Cross-modal Contrastive Learning framework. The qualitative results are shown in Fig. 6. Besides, we also verify the cross-domain generalization ability of our method. When pre-training the

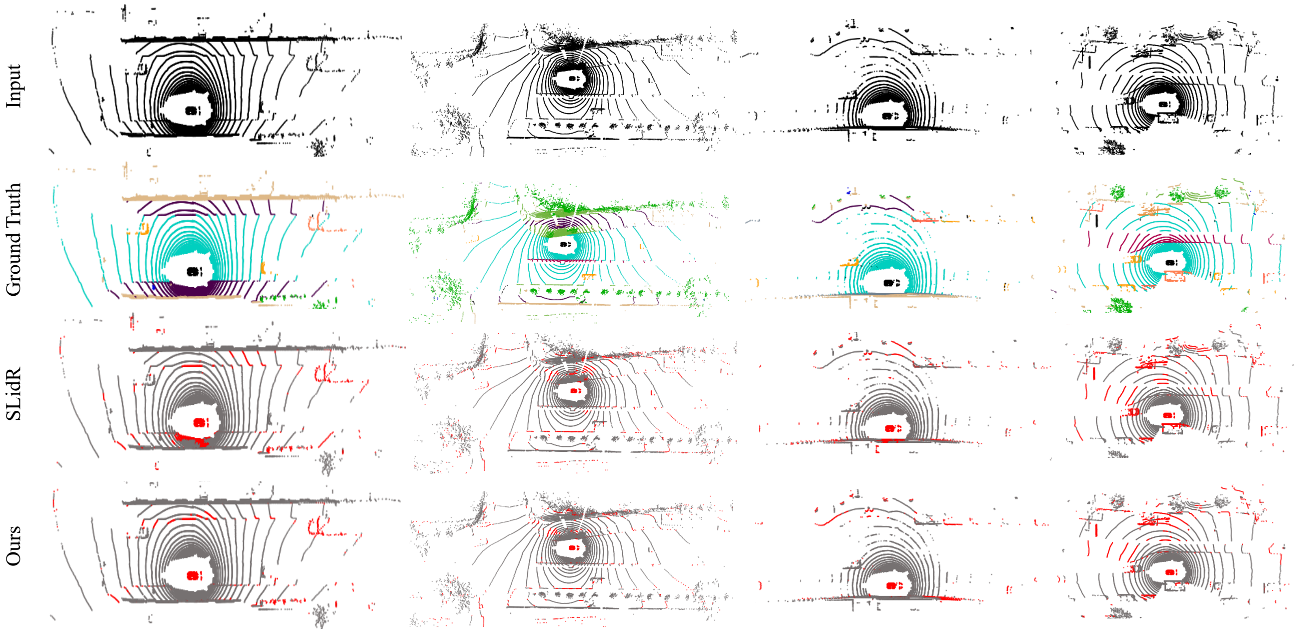

Figure 6. Qualitative results of fine-tuning on 1% nuScenes dataset. From the first row to the last row are the input LiDAR scan, ground truth, prediction of SLidR, and our prediction, respectively. Note that we show the results by error map, where the red point indicates the wrong prediction. Apparently, our method achieves decent performance.

3D network on the nuScenes dataset and fine-tuning on 1% and 100% SemanticKITTI dataset, our method significantly outperforms other state-of-the-art self-supervised methods.

Discussions. PPKT and SLidR reveal that contrastive loss is promising for transferring knowledge from image to point cloud. Like self-supervised learning, constructing the positive and negative samples is vital to unsupervised cross-modal knowledge distillation. However, previous methods suffer from optimization-conflict issues, *i.e.*, some negative paired samples are actually positive pairs. For example, the *road* occupies a large proportion of the point cloud in a scene and is supposed to have the same semantics in the semantic segmentation task. When randomly selecting training samples, most negatively defined road-road points are actually positive. When feedforwarding such samples into contrastive learning, the contrastive loss will push them away in the embedding space, leading to unsatisfactory representation learning and hammering the downstream tasks' performance. SLidR introduces superpixel-driven contrastive learning to alleviate such issues. The motivation is that the visual representation of the image pixel and the projected points are consistent intra-superpixel. Although avoiding selecting negative image-point pairs from the same superpixel, the conflict still exists inter-superpixel. In our CLIP2Scene, we introduce the free-available dense pixel-text correspondence to alleviate the optimization conflicts. The text embedding represents the semantic information and can be used to select more reasonable training samples for contrastive learning.

Besides training sample selection, the previous method also ignores the temporal coherence of the multi-sweep point cloud. That is, for LiDAR points mapping to the same image pixel, their feature is restricted to be consistent. Besides, considering the calibration error between the LiDAR scan and the camera image. We relax the pixel-to-point mapping to image grid-to-point grid mapping for consistency regularization. To this end, our Spatial-temporal consistency regularization leads to a more rational point representation.

Last but not least, we find that randomly switching the supervision signal benefits self-supervised learning. Essentially, different modality networks learn different feature representations. They can filter different types of errors introduced by noisy pseudo labels. By this switchable operation, the error flows can be reduced mutually.

5. Conclusion

We explored how CLIP knowledge benefits 3D scene understanding in this work, termed CLIP2Scene. To efficiently transfer CLIP's image and text features to a 3D network, we propose a novel Semantic-driven Cross-modal Contrastive Learning framework including Semantic Regularization and Spatial-Temporal Regularization. For the first time, our pre-trained 3D network achieves annotation-free 3D semantic segmentation with decent performance. Besides, our method significantly outperforms state-of-the-art self-supervised methods when fine-tuning with labelled data.

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013. 2

[2] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 2

[3] 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 *IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 2, 5, 6

[4] Max Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *IEEE/CVF International Conference on Computer Vision Workshop*, pages 2666–2673, 2017. 2

[5] 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. 5

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020. 3

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924, 2020. 2

[8] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016. 2

[9] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9121–9130, 2020. 2

[10] Runnan Chen, Zhu Xinge, Nenglun Chen, Dawei Wang, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Referring self-supervised learning on 3d point cloud. 2021. 2

[11] Runnan Chen, Xinge Zhu, Nenglun Chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Zero-shot point cloud segmentation by transferring geometric primitives. *arXiv preprint arXiv:2210.09923*, 2022. 1, 2

[12] Runnan Chen, Xinge Zhu, Nenglun Chen, Dawei Wang, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Towards 3d scene understanding by referring synthetic models. *arXiv preprint arXiv:2203.10546*, 2022. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 2

[14] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[15] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021. 1

[16] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the Hubness Problem for Zero-Shot Learning of 3D Objects. *arXiv preprint arXiv:1907.06371*, 2019. 2

[17] Ali Cheraghian, Shafin Rahman, Townim F Chowdhury, Dylan Campbell, and Lars Petersson. Zero-shot learning on 3d point cloud objects and beyond. *International Journal of Computer Vision*, 130(10):2364–2384, 2022. 2

[18] Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-Shot Learning of 3D Point Cloud Objects. In *IEEE International Conference on Machine Vision Applications*, pages 1–6, 2019. 2

[19] 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. 6

[20] Angela Dai, Angel Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 2, 5, 6

[21] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *IEEE/CVF International Conference on Computer Vision*, pages 1241–1250, 2017. 2

[22] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 1422–1430, 2015. 2

[23] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014. 2

[24] 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, 5, 6

[25] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *AAAI Conference on Artificial Intelligence*, 2015. 2

[26] 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

[27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020. 2

[28] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2020. 1

[29] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 5

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[32] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting network. *arXiv preprint arXiv:2203.07186*, 2022. 1

[33] 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. 1

[34] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 2

[35] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[36] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *arXiv preprint arXiv:2205.15585*, 2022. 2

[37] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4447–4456, 2017. 2

[38] 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. *arXiv preprint arXiv:2303.05367*, 2023. 1

[39] 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*, 2023. 1

[40] 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*, 2023. 1

[41] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 2

[42] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:453–465, 2014. 2

[43] Yan Li, Zhen Jia, Junge Zhang, Kaiqi Huang, and Tieniu Tan. Deep semantic structural constraints for zero-shot learning. In *AAAI Conference on Artificial Intelligence*, 2018. 2

[44] 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. 1, 2, 3, 4, 6, 7

[45] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *International Conference on 3D Vision*, pages 992–1002, 2021. 1, 2

[46] Ashish Mishra, M. Shiva Krishna Reddy, Anurag Mittal, and Hema A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 2269–22698, 2018. 2

[47] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 1, 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[49] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2

[50] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141, 2022. 2

[51] 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. 1, 2, 3, 4, 5, 6, 7

[52] 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. 6

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 3

[54] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016. 2

[55] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pretraining for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591, 2020. 2, 6, 7

[56] 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. 1

[57] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 2, 3

[58] Éloi Zablocki, Patrick Bordes, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Context-aware zero-shot learning for object recognition. In *International Conference on Machine Learning*, pages 7292–7303, 2019. 2

[59] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 2

[60] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2

[61] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712, 2022. 2, 4, 6

[62] 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