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

Runnan Chen1,2 Youquan Liu2,3 Lingdong Kong2,4 Xinge Zhu5 Yuexin Ma6 Yikang Li2 Yuenan Hou2,† Yu Qiao2 Wenping Wang7,†
1The University of Hong Kong 2Shanghai AI Laboratory 3Hochschule Bremerhaven 4National University of Singapore
5The Chinese University of Hong Kong 6ShanghaiTech University 7Texas 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 available1.

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
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 hampering 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 dataset [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 distills 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
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 super-pixel 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-
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 irrelevant 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 \( \{x_i, p_j\}_{i \neq j}^M \), 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 hammers 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:

\[
\mathcal{L}_{S, info} = - \frac{1}{C} \sum_{c=1}^{C} \log \frac{\sum_{t_i \in c, p_i} \exp(D(t_i, p_i)/\tau)}{\sum_{t_i \in c, p_i \neq p_j} \exp(D(t_i, p_j)/\tau)},
\]

(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 \( \tau \) 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.
Multi-sweeps calibration

\[ \text{Pixel-to-point mapping } \Phi \]

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 \( \{ \hat{x}^k_i, \hat{p}^k_i \}_{i=1,k=1}^{M,K} \) 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 \( \{ \hat{x}^1_i, \hat{p}^1_i \}_{i=1}^M \). 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 \( \{ \hat{x}^k_i, \hat{p}^k_i, t^k_i \}_{i=1,k=1}^{M,K} \). 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:

\[
\mathcal{L}_{SSR} = \sum_{g_n} \sum_{(i,k) \in g_n} (1 - \text{sigmoid}(D(\hat{p}^k_i, f_n))) / N, \tag{2}
\]

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

\[
f_n = \sum_{(i,k) \in g_n} a^k_i \hat{x}^k_i + b^k_i \hat{p}^k_i, \tag{3}
\]

where \( a^k_i \) and \( b^k_i \) are attention weight calculated by:

\[
a^k_i = \frac{\exp(D(\hat{x}^k_i, t^k_i)/\lambda)}{\sum_{(i,k) \in g_n} \exp(D(\hat{x}^k_i, t^k_i)/\lambda) + \exp(D(\hat{p}^k_i, t^k_i)/\lambda)},
\]

\[
b^k_i = \frac{\exp(D(\hat{p}^k_i, t^k_i)/\lambda)}{\sum_{(i,k) \in g_n} \exp(D(\hat{x}^k_i, t^k_i)/\lambda) + \exp(D(\hat{p}^k_i, t^k_i)/\lambda)},
\]

where \( \lambda \) 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 \( \mathcal{L}_{S,jn,fo} \) 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 \( \{ \hat{x}^k_i, \hat{p}^k_i \}_{i=1,k=1}^{M,K} \), which are used in \( \mathcal{L}_{S,jn,fo} \) 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
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 SPCVNN [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 $\lambda$ and $\tau$ 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 semantic 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

**Method.** 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.

**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 (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.

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

<table>
<thead>
<tr>
<th>Initialization</th>
<th>nuScenes 1%</th>
<th>100%</th>
<th>SemanticKITTI 1%</th>
<th>100%</th>
<th>ScanNet 5%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>42.2 69.1</td>
<td>32.5</td>
<td>52.1</td>
<td>46.1</td>
<td>63.3</td>
<td></td>
</tr>
<tr>
<td>PPkT [44]</td>
<td>48.0 70.1</td>
<td>39.1</td>
<td>53.1</td>
<td>47.5</td>
<td>64.2</td>
<td></td>
</tr>
<tr>
<td>SLidR [51]</td>
<td>48.2 70.4</td>
<td>39.6</td>
<td>54.3</td>
<td>47.9</td>
<td>64.9</td>
<td></td>
</tr>
<tr>
<td>PointContrast [55]</td>
<td>47.2 69.2</td>
<td>37.1</td>
<td>52.3</td>
<td>47.6</td>
<td>64.5</td>
<td></td>
</tr>
<tr>
<td>CLIP2Scene</td>
<td>56.3 71.5</td>
<td>42.6</td>
<td>55.0</td>
<td>48.4</td>
<td>65.1</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>nuScenes</th>
<th>ScanNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP2Scene</td>
<td>20.80</td>
<td>25.08</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Ablation Target</th>
<th>Setting</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Baseline</td>
<td>15.1</td>
</tr>
<tr>
<td>Prompts</td>
<td>nuScenes</td>
<td>15.1 (+0.6)</td>
</tr>
<tr>
<td></td>
<td>SemanticKITTI</td>
<td>13.9 (+1.2)</td>
</tr>
<tr>
<td></td>
<td>Cityscapes</td>
<td>11.3 (+3.8)</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>15.3 (+0.2)</td>
</tr>
<tr>
<td>Regularization</td>
<td>w/o StCR</td>
<td>19.8 (+4.7)</td>
</tr>
<tr>
<td></td>
<td>w/o SCR</td>
<td>16.8 (+1.7)</td>
</tr>
<tr>
<td></td>
<td>KL</td>
<td>0.00 (+15.3)</td>
</tr>
<tr>
<td>Training Strategy</td>
<td>w/o S3</td>
<td>18.8 (+3.7)</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>10.1 (+4.0)</td>
</tr>
<tr>
<td>Sweeps</td>
<td>1 sweep</td>
<td>18.7 (+3.6)</td>
</tr>
<tr>
<td></td>
<td>3 sweeps</td>
<td>20.8 (+5.7)</td>
</tr>
<tr>
<td></td>
<td>5 sweeps</td>
<td>20.6 (+5.5)</td>
</tr>
<tr>
<td></td>
<td>merged</td>
<td>18.6 (+5.5)</td>
</tr>
<tr>
<td>Full Configuration</td>
<td>CLIP2Scene</td>
<td>20.8 (+5.7)</td>
</tr>
</tbody>
</table>
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


[40] Rethinking range view representation for lidar segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 652–660, 2017. 1, 2


