VLM2Scene: Self-Supervised Image-Text-LiDAR Learning with Foundation Models for Autonomous Driving Scene Understanding

Guibiao Liao[1, 2], **Jiankun Li**[3], **Xiaoqing Ye**[3*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Baidu Inc., China
gbliao269@gmail.com, lijiankun02@baidu.com, yxq@whu.edu.cn

Abstract

Vision and language foundation models (VLMs) have showcased impressive capabilities in 2D scene understanding. However, their latent potential in elevating the understanding of 3D autonomous driving scenes remains untapped. In this paper, we propose **VLM2Scene**, which exploits the potential of VLMs to enhance 3D self-supervised representation learning through our proposed image-text-LiDAR contrastive learning strategy. Specifically, in the realm of autonomous driving scenes, the inherent sparsity of LiDAR point clouds poses a notable challenge for *point-level* contrastive learning methods. This method often grapples with limitations tied to a restricted receptive field and the presence of noisy points. To tackle this challenge, our approach emphasizes *region-level* learning, leveraging regional masks without semantics derived from the vision foundation model. This approach capitalizes on valuable contextual information to enhance the learning of point cloud representations. **First**, we introduce Region Caption Prompts to generate fine-grained language descriptions for the corresponding regions, utilizing the language foundation model. These region prompts then facilitate the establishment of positive and negative text-point pairs within the contrastive loss framework. **Second**, we propose a Region Semantic Concordance Regularization, which involves a semantic-filtered region learning and a region semantic assignment strategy. The former aims to filter the false negative samples based on the semantic distance, and the latter mitigates potential inaccuracies in pixel semantics, thereby enhancing overall semantic consistency. Extensive experiments on representative autonomous driving datasets demonstrate that our self-supervised method significantly outperforms other counterparts. Codes are available at https://github.com/gbliao/VLM2Scene.

Introduction

3D scene understanding from LiDAR point clouds is a crucial perception component in autonomous driving. Most existing deep learning-based methods have yielded noteworthy progress with massive labeled data in this area (Qi et al. 2017a,b; Guo et al. 2020; Gao et al. 2021; Zhu et al. 2021; Hou et al. 2022). However, the manual annotation of point clouds is a resource-intensive endeavor, impeding its practicality within real-world autonomous driving scenarios. Consequently, self-supervised learning for 3D scene understanding, which harnesses the power of unlabeled data and defines the pretext task as semantic segmentation, arises as a promising and meaningful realm of study (Fei et al. 2023).

Vision and language foundation models (VLMs), including Contrastive Vision-Language Pre-training (CLIP) (Radford et al. 2021), Bootstrapping Language-Image Pre-training (BLIP-2) (Li et al. 2023), and Segment Anything (SAM) (Kirillov et al. 2023), have recently gained significant attention. CLIP, utilizing large-scale web-crawled image-text data, constructs powerful vision-language embeddings that show promising performance in zero-shot image classification. Effectively training on extensive image-text web data through image-text contrastive and matching losses, BLIP2 showcases an impressive ability for zero-shot image-to-text generation. In addition, SAM trains on a vast amount of image data (11 million images, 1 billion masks), revealing robust zero-shot segmentation capabilities. While these methods exhibit remarkable zero-shot image understanding abilities within open-world scenarios, their direct suitability for 3D tasks poses a challenge due to the scarcity of massive 3D annotations and large-scale 3D-text data.

To address this challenge, the recent effort (Chen et al. 2023) explores the utilization of CLIP for self-supervised 3D scene understanding. It begins by employing template text prompts alongside the annotation-free segmentation model MaskCLIP (Zhou, Loy, and Dai 2022) to generate image pixel predictions. Subsequently, a pixel-point contrastive learning scheme is proposed to transfer these image pixel predictions to 3D space through projection. Despite these advancements, some limitations may hinder its suitability for comprehensive 3D scene understanding. First, the mentioned textual embeddings are directly extracted from the dataset's categories (e.g., *car*, *bus*, *pedestrian*, etc.) using brief templates (e.g., *a photo of a* {}, *there is a* {} *in the scene*, etc.). Yet, these text descriptions may fail to provide a fine-grained depiction of the scene. Second, in outdoor autonomous driving scenes, the inherent sparsity within LiDAR-generated point clouds is a pronounced characteristic. This inherent sparsity magnifies the difficulty of employing the point-level contrastive learning method, because this mode of point-level learning might be susceptible to challenges arising from restricted receptive fields, limited

*Corresponding Author.

context, and the potential erroneous pixel semantics.

Building upon the impressive accomplishments of VLMs, we ask: *Can we harness the strengths of VLMs to overcome these limitations and advance 3D representations for scene understanding?* In pursuit of this goal, our core idea is to utilize VLMs for image-text-LiDAR self-supervised learning, aiming to capture useful contextual knowledge to enhance 3D representations. To this end, two critical factors come into play: (1) the model should learn more specific and fine-grained language descriptions, and (2) the model should fully explore informative region-level representations.

In this paper, we present **VLM2Scene**, an innovative approach to *region-level* image-text-LiDAR contrastive learning from two perspectives. Specifically, we initiate the process by extracting class-agnostic regions (i.e., regional masks without semantics) from the image using SAM. Subsequently, 1) to enrich region-level text prompts, we introduce a strategy called Region Caption Prompts. This approach leverages the powerful image-to-text generation capabilities of BLIP-2, generating specific, fine-grained captions for each region derived from SAM. These captions encompass details like location relationships and color properties, thereby significantly diversifying the textual content. 2) To enhance consistent region-wise representations, we present Region Semantic Concordance Regularization, which involves semantic-filtered region learning and a region semantic assignment strategy, with the following considerations. A straightforward method is to employ region-wise image-LiDAR contrastive learning based on SAM's region masks. However, in this way, semantically identical parts of the same object are treated as negative samples and pushed away, making it challenging to learn. To mitigate this issue, we propose a semantic-filtered region learning strategy. This strategy leverages distances between regions in the text space to guide the contrastive loss calculation process, effectively filtering out false negative samples. Furthermore, considering potential incorrect pixel semantics from the vanilla CLIP visual encoder, a region semantic assignment strategy is introduced, that collaborates with SAM's masks to enhance regional semantic consistency.

Our contributions can be summarized as follows:

- We propose VLM2Scene, a novel approach aimed at harnessing the potential of VLMs to leverage text and image cues to enhance LiDAR representation in driving scenes.
- We propose Region Caption Prompts and Region Semantic Concordance Regularization, which exploit region-level image-text-lidar contrastive learning, elevating the quality of 3D context representation learning for scene understanding.
- Extensive experiments show that our method can improve 3D representation and achieve superior results for downstream automatic driving semantic segmentation.

Related Work

Autonomous Driving Scene Understanding

Scene understanding plays a crucial role in ensuring the safety and efficiency of autonomous driving systems. Current supervised methods for 3D scene understanding have made tremendous progress (Qi et al. 2017a,b; Thomas et al. 2019; Tang et al. 2020; Choy, Gwak, and Savarese 2019; Zhang et al. 2022; Gao et al. 2021; Zhu et al. 2021; Hou et al. 2022; Wu et al. 2022; Lai et al. 2023; Kong et al. 2023). However, these approaches exhibit a strong reliance on extensive annotation endeavors, hindering their usefulness in real-world autonomous driving scenarios. To overcome this restriction, some methods (Liu et al. 2021; Sautier et al. 2022; Chen et al. 2023; Mahmoud et al. 2023; Peng et al. 2023) made an attempt to mitigate the demands of 3D annotation by transferring 2D image knowledge. Importantly, how to transfer knowledge from a 2D network to a 3D network becomes critical in the learning of 3D representations. In this work, we aim to explore image and text knowledge from vision and language foundation models (VLMs), such as CLIP, BLIP-2, and SAM, to better advance 3D representations for autonomous driving scene understanding.

3D Self-supervised Representation Learning

Unlike 2D vision tasks (He et al. 2020; Chen et al. 2020b; Grill et al. 2020; Chen and He 2021; Chen et al. 2020a; He et al. 2022), which are often pre-trained on large-scale datasets like ImageNet (Deng et al. 2009), pre-training for 3D vision tasks faces distinct challenges. The high cost of data annotation and the inherent sparsity of outdoor LiDAR point clouds make it difficult to effective 3D pre-training. Recently, self-supervised representation learning for 3D scene understanding has gained a lot of attention (Fei et al. 2023). PointContrast (Xie et al. 2020) first shows that the contrastive learning paradigm helps 3D models learn valuable features from unlabeled RGB-D datasets. These pre-trained representations exhibit their efficacy in 3D scene segmentation. PPKT (Liu et al. 2021) exploits 2D-3D knowledge transferring to take advantage of the abundant semantic information learned from large-scale 2D datasets, which boosts the performances in 3D downstream tasks. However, in outdoor scenarios characterized by sparse LiDAR points, the point-level contrastive learning strategy in PointContrast and PPKT might fail to capture the features of sparse objects. To mitigate the sparse issue, SLidR (Sautier et al. 2022) proposes a 2D-3D representation distillation method based on super-pixels, and ST-SLidR (Mahmoud et al. 2023) further introduces a semantically tolerant loss to alleviate the local semantic ambiguity. Nevertheless, hand-crafted super-pixel may inherently fail to accurately provide semantically consistent segmentation regions, limiting accurate contextual understanding. The up-to-date work CLIP2Scene (Chen et al. 2023) transfers 2D pixel knowledge from CLIP (Radford et al. 2021) to the 3D network based on a point-level contrastive learning scheme. However, such a point-level method suffers from the potential noise point or erroneous pixel semantics due to the lack of region-level representations.

Overall, considering the sparsity of outdoor LiDAR point clouds and the limitations of the point-level learning scheme, we explore the potential of vision and language models (VLMs) to make the best of *region-wise* text and image knowledge, enhancing informative 3D context representation learning for scene understanding.

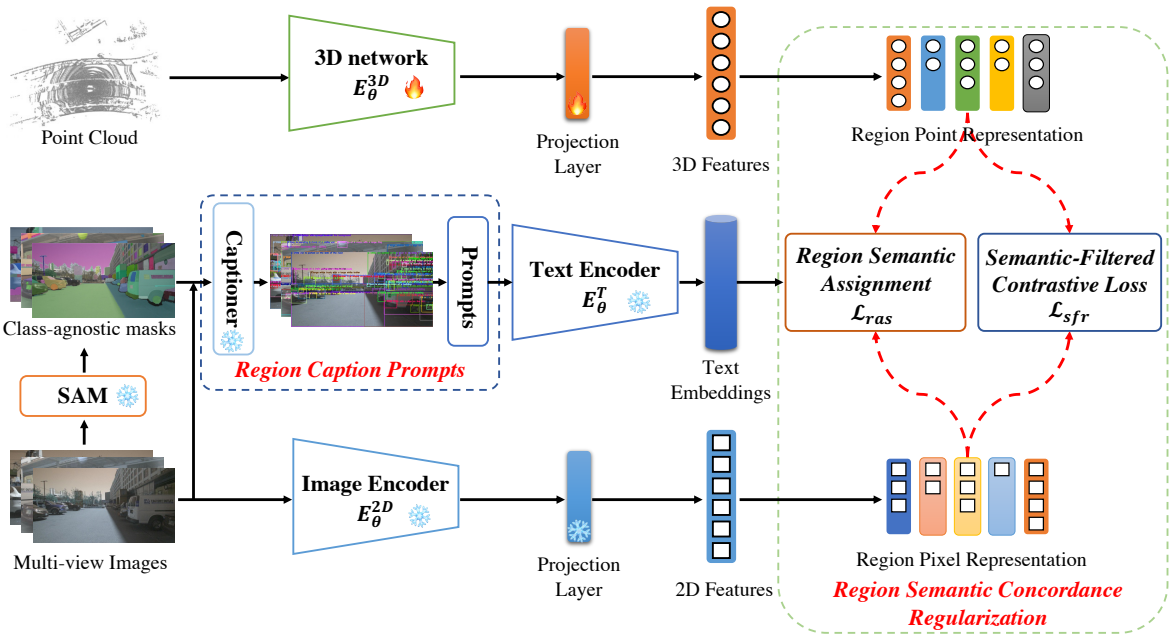


Figure 1: Illustration of our VLM2Scene. First, LiDAR point cloud and multi-view images are fed into a trainable 3D network and a frozen CLIP image encoder for feature extraction, respectively. Then, we use the vision foundation model SAM to generate class-agnostic masks over images. To enrich textual representation and better leverage CLIP knowledge, we propose Region Caption Prompts, that utilize an image caption foundation model BLIP-2 to produce fine-grained region-wise captions according to region inputs. These prompts are later used to extract text embedding by the CLIP text encoder. Moreover, the 3D features and 2D features are grouped into region representations, respectively, and then our proposed Semantic-Filtered Contrastive Loss and Region Semantic Assignment strategy are used to enhance 3D representations.

Methodology

Vision and language foundation models have shown impressive capabilities in 2D scene understanding (Radford et al. 2021; Kirillov et al. 2023). However, whether their capability can promote 3D autonomous driving scene understanding is a challenging question. Therefore, in this paper, we study the self-supervised image-text-LiDAR contrastive learning with vision and language foundation models (VLMs) for 3D autonomous driving scene understanding, namely VLM2Scene.

Reviewing and Motivations

Reviewing VLMs. VLM denotes a sophisticated deep learning model that is pre-trained on large amounts of data, whose capability empowers it to excel in diverse downstream tasks. For instance, CLIP (Radford et al. 2021), BLIP-2 (Li et al. 2023), and SAM (Kirillov et al. 2023) show impressive zero-shot image classification, image-to-text generation, and class-agnostic segmentation performance, respectively, and have attracted a lot of attention lately.

- CLIP contains a vision model and a text model and designs an image-text contrastive learning strategy. This enables CLIP to associate images and their corresponding text descriptions in a shared embedding space, resulting in powerful *open-world image classification*.
- BLIP-2 proposes a Querying Transformer to bridge the vision-language gap and bootstrap vision-to-language generative learning, which enables BLIP-2 to obtain impressive performance in *image-to-text generation*.
- As for the vision foundation model, SAM combines an image encoder, a prompt encoder, and a mask decoder, which is trained on a large amount of labeled data (11 million images, 1 billion masks). These massive amounts of training data allow it to obtain excellent *class-agnostic region proposals* across a wide range of distributions.

Recently, CLIP2Scene (Chen et al. 2023) attempts to leverage CLIP for 3D scene understanding. It consists of a 2D frozen annotation-free segmentation model MaskCLIP (Zhou, Loy, and Dai 2022) and a 3D network (Choy, Gwak, and Savarese 2019) to be trained. CLIP2Scene feeds fixed template texts into MaskCLIP to generate pixel-level image pseudo-labels. Then, it devises the pixel-point contrastive learning scheme at the point level. However, certain limitations may hinder their 3D representation learning. First, the aforementioned textual embeddings are straightforwardly generated from the category names (e.g., *car*, *pedestrian*, etc) of the dataset with fixed and short templates (e.g., *a photo of a* {}, etc). These text descriptions lack a realistic and detailed description of the scene. Second, the inherent sparsity in LiDAR point clouds poses the challenges of point-level contrastive learning, attributed to factors such as restricted receptive fields, limited context, and the susceptibility to noisy points or erroneous pixel semantics.

Motivations. Motivated by the above analysis, we propose

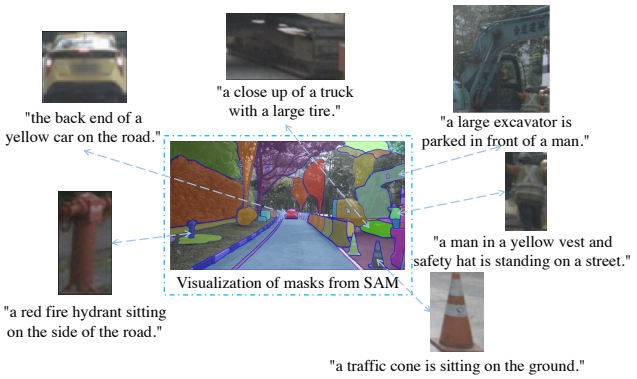


Figure 2: Visualization of the mask result from SAM (blue dotted box) and some generated region caption examples. In each example, the upper section shows the visual image region, while the lower section displays the corresponding fine-grained and specific text.

to make the most of the advantages of VLMs to promote 3D self-supervised representation learning. As depicted in Fig. 1, the image pixel features and point features are extracted by leveraging the CLIP visual encoder and a 3D network, respectively. Considering the bottlenecks of point-level mentioned above, we propose to exploit the foundation model SAM to produce class-agnostic regions over the image, and two *region-aware* strategies are introduced to make the best of region-level image and point features.

- First, unlike conventional methods that merely use class names with fixed templates, we propose Region Caption Prompts for region-wise textual embedding generation. Specifically, we combine the region masks derived from SAM with the input image, and leverage the BLIP-2 to generate fine-grained and rich caption descriptions over the regions. Then, these generated region captions are fed into the frozen text encoder of CLIP to produce region-wise textual embeddings.
- Second, given the impressive consistency of region masks from SAM, a Region Semantic Concordance Regularization strategy is proposed. It consists of a semantic-filtered region learning strategy and a region semantic assignment strategy, effectively enhancing the perception of contextual information and forcing the consistency of region representations.

Region Caption Prompt Strategy

Using short and brief template sentences with category names as text prompts for text embedding extraction potentially constrains the model's perception capacity. To overcome this constraint and improve the text quality, we propose a **Region Caption Prompt (RCP)** strategy that enriches region-level text prompts with more fine-grained and specific language descriptions.

For this purpose, we leverage the image-to-text generation strength of BLIP-2 (Li et al. 2023) to generate a series of raw language descriptions associated with the specified region based on the supplied SAM's masks and the original visual image. Specifically, we first crop the corresponding visual image region as foreground by region masks from SAM. To better understand the contextual information, we use a regular box to truncate the region that contains exactly the object and keep the background information of this region. Then, we input the above visual region into the BLIP-2 model and used a text prompt to guide raw language description generation, such as *'Question: what is the content of the image? Answer:'*. In particular, we set a minimum area of the mask to prevent the generation of overly localized or meaningless descriptions. Furthermore, we manually filter out descriptions that are irrelevant or meaningless to the categories of the pre-training dataset, such as *'a cloudy sky'*, to produce meaningful and descriptive captions that enhance region texts for each category.

Therefore, in each iteration of the pre-training, for each scene, we first generate rich and diverse regional text descriptions and categorize them by dataset categories. Then, we leverage the CLIP text encoder to derive the fine-grained region text embeddings $t_r \in \mathbb{R}^{B\times L}$. Meanwhile, we also employ the template text embeddings $t_c \in \mathbb{R}^{B\times L}$ obtained from the template-based text prompts used in CLIP2Scene for pre-training. Here, B and L denote the number of the pre-training dataset's class and the feature dimension, respectively. In this way, the generated region texts and the template texts can complement each other, resulting in better performance as the experiment shows.

As depicted in Fig. 2, we present the text enrichment example from our RCP. We can see that 1) Our RCP can generate realistic and specific region-level captions for each scenario. 2) Moreover, it also presents finer-grained and comprehensive descriptions, such as location relationships, color attributes, etc. 3) By applying our RCP, we exploit the potential of text modality to capture meaningful semantic information for the following contrastive learning process, enhancing 3D self-supervised representation learning.

Region Semantic Concordance Regularization

The CLIP-based pixel-point contrastive learning paradigm is crucial for achieving cross-modal alignment and enhancing 3D representation learning (Chen et al. 2023). A straightforward method is to use a frozen CLIP visual encoder to extract pixel-wise image features, and align it to corresponding point features by leveraging the 2D-3D calibration matrix. However, this point/pixel level contrastive learning method may lack sufficient receptive field that is incapable of perceiving context environments. Additionally, inaccurate semantic prediction from the CLIP visual encoder (e.g., (b) of Fig. 4) may mislead the learning of 3D feature representation, resulting in unsatisfactory performance.

To tackle the above challenges, we propose leveraging well-grouped region proposals from vision foundation model SAM, and utilizing region-level contrastive learning with visually similar features to enhance 3D feature representations. Specifically, we denote E_θ^{3D} as the trainable 3D network and E_θ^{2D} as the frozen CLIP image encoder, that takes the LiDAR point cloud and corresponding multi-view images as input, respectively. Through feature extraction,

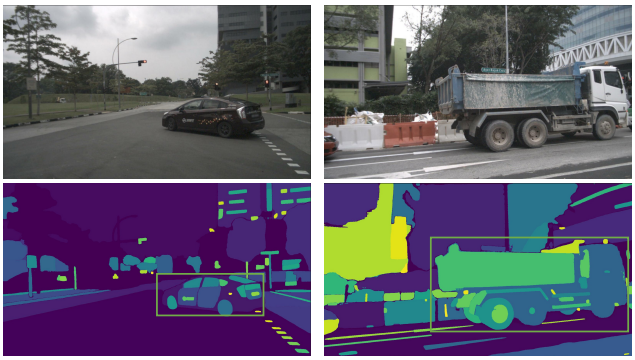


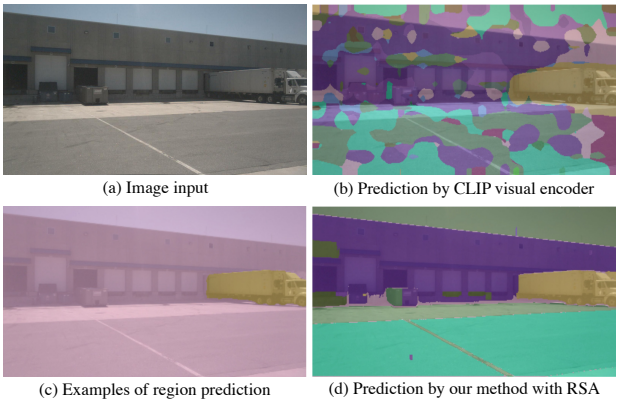
Figure 3: First row: Camera Images. Second row: Visualization of the corresponding mask result from SAM. However, the vanilla per-mask-driven contrastive learning method treats other semantically identical parts of the same object as negative samples (e.g. the case in the green box), resulting in false negative samples. Accordingly, such false negative samples will interfere with the regional semantic structure for accurate 3D representation learning.

the 3D point cloud features $F_P \in \mathbb{R}^{N\times D}$ and 2D image pixel features $F_I \in \mathbb{R}^{h\times w\times C}$ can be produced. Here, N and D denote the number and the feature dimension of the point cloud feature, respectively. h, w, and C indicate the size and the feature dimension of the image feature, respectively. To pre-train the 3D point cloud network without any annotated labels, we transfer the image knowledge to the 3D network via cross-modal alignment. To achieve this, we first use a trainable point cloud projection layer to map the F_P to the share contrastive loss embedding space, resulting in $\hat{F}_P \in \mathbb{R}^{N\times L}$. The image pixel features F_I is mapped to $\hat{F}_I \in \mathbb{R}^{h\times w\times L}$ with a frozen projection layer from the CLIP image encoder. Then, we leverage known sensor calibration parameters to build point-pixel correspondence $\{p_i, x_i\}_{i=1}^{K}$, where p_i and x_i denote the i-th paired point feature and image pixel feature, respectively. K indicates the number of pairs. Notably, different from previous point-level contrastive learning, we conduct a region-level solution to promote informative feature learning. Concretely, based on the point-pixel pair correspondence and the region masks from SAM, we compute the region-wise point and pixel representation and then group them with mean pooling operation into region point representation $P \in \mathbb{R}^{M\times L}$ and region pixel representation $Q \in \mathbb{R}^{M\times L}$, respectively. In this way, vanilla region-level feature contrastive learning can be calculated:

$$\mathcal{L}(\mathbf{P},\mathbf{Q}) = -\frac{1}{M}\sum_{i=0}^{M}\log\left[\frac{e^{((\mathbf{p}_i\cdot\mathbf{q}_i)/\tau)}}{\sum_{j\neq i}e^{((\mathbf{p}_i\cdot\mathbf{q}_j)/\tau)}+e^{((\mathbf{p}_i\cdot\mathbf{q}_i)/\tau)}}\right], \tag{1}$$

where $\cdot$ indicates the scalar product operation for similarity measurement. τ denotes the temperature term.

Although the aforementioned vanilla region-level contrastive learning can promote 3D representation learning, it fails to consider the *self semantic conflict* issue as shown in Fig. 3. In other words, different parts belonging to the same object may be labeled as different mask regions, due to the random point prompts of the original SAM model. The



(a) Image input (b) Prediction by CLIP visual encoder

(c) Examples of region prediction (d) Prediction by our method with RSA

Figure 4: Illustration of different image predictions. From (b), we can see that the prediction by the CLIP visual encoder can only roughly identify the localization of objects, but lacks precise edges and may even generate incorrect regional predictions. From (c), our RSA strategy can impose semantic consistency by leveraging region-level masks compared to pixel-level, thus producing more precise representations (d) for self-supervised learning.

above process will treat such semantically identical parts of the same object as negative samples, pushing away these false negative samples and making it challenging to learn. To mitigate this issue, we propose a **Semantic-Filtered Region Learning (SFR)** strategy, that leverages the image region feature and text embeddings to filter out regions sharing similar texts. In other words, we can use the distance between regions in the text space as a guiding signal for contrastive loss calculation. Specifically, for two regions i and j present in the image, if $\epsilon > \Phi(t_i, x_i) - \Phi(t_i, x_j)$, where ϵ is a small threshold, and Φ is the cosine-similarity measurement, the semantic of i and j will be inferred to very close, and j will be removed from the negative samples of i during contrastive loss computing. Thus, for each positive sample $\mathbf{p}_i$, we use $\mathbf{T}_{ij}$ denotes the indicator function that determines whether negative sample $\mathbf{q}_j$ will be filtered based on the above calculation process. Concretely, the negative sample will be removed when $\mathbf{T}_{ij}$ is equal to 0, and retained when 1 otherwise. This process can be formulated as:

$$\mathcal{L}_{sfr}(\mathbf{P},\mathbf{Q}) = -\frac{1}{M}\sum_{i=0}^{M}\log\left[\frac{e^{((\mathbf{p}_i\cdot\mathbf{q}_i)/\tau)}}{\sum_{j\neq i}\mathbf{T}_{ij}\cdot e^{((\mathbf{p}_i\cdot\mathbf{q}_j)/\tau)}+e^{((\mathbf{p}_i\cdot\mathbf{q}_i)/\tau)}}\right]. \tag{2}$$

In addition to semantic feature learning via image-LiDAR pairs, we also consider the semantic assignment via image-text-LiDAR pairs for better 3D representations. The recent CLIP2Scene uses CLIP to produce 2D pixel labels via image-text pairs and assigns these pixel labels to 3D points. However, it may fall short of elaborate object localization and even generate noisy assignments as shown in (b) of Fig. 4, hindering accurate 3D representation learning. To alleviate this issue, we propose a **Region Semantic Assignment (RSA)** strategy, that imposes a per-mask constraint on semantic assignment. Concretely, for each region mask assign-

Method	Reference	nuScenes						KITTI
		LP	1%	5%	10%	25%	100%	1%
Random	N/A	8.10	30.30	47.84	56.15	65.48	74.20	39.50
PointContrast (Xie et al. 2020)	ECCV20	21.90	32.50	-	-	-	-	41.10
DepthContrast (Zhang et al. 2021)	ICCV21	22.10	31.70	-	-	-	-	41.50
PPKT (Liu et al. 2021)	arXiv21	35.90	37.80	53.74	60.25	67.14	74.52	44.00
SLidR (Sautier et al. 2022)	CVPR22	38.80	38.30	52.49	59.84	66.91	74.79	44.60
CLIP2Scene (Chen et al. 2023)	CVPR23	-	33.05	52.18	59.87	66.87	74.63	43.10
ST-SLidR (Mahmoud et al. 2023)	CVPR23	40.48	40.75	54.69	60.75	67.70	75.14	44.72
VLM2Scene (Ours)		**51.54**	**47.59**	**58.08**	**63.08**	**68.39**	**75.42**	**47.37**

Table 1: Performance comparison with other methods pre-trained on nuScenes and fine-tuned on nuScenes, and SemanticKITTI. LP indicates linear probing with frozen backbones. We report the mIoU scores for evaluation.

ment, we count the number of categories appearing in the region. Then, we sort them and select the category with the most occurrences to assign as the semantic of this region. It aims at reducing errors and inconsistencies by selecting the most semantically similar factors per region, improving the quality of semantic assignment as shown in (c) and (d) of Fig. 4. Then, we use the same cross-entropy loss calculation as CLIP2Scene with our improved semantic assignment to optimize the 3D network, and this process can be denoted as $\mathcal{L}_{ras}$. In summary, the total pre-training loss consists of $\mathcal{L}_{sfr}$ and $\mathcal{L}_{ras}$, which can be formulated: $\mathcal{L} = \mathcal{L}_{sfr} + \mathcal{L}_{ras}$. More details are presented in the appendix.

Experiments

Experiments Setup

Datasets. To evaluate the performance of our proposed methods, we conduct experiments on two large-scale autonomous driving datasets, i.e., SemanticKITTI (Behley et al. 2019) and nuScenes (Caesar et al. 2020). The SemanticKITTI dataset, collected in urban street scenes by a Velodyne HDL-64E LiDAR sensor, is a comprehensive dataset for LiDAR autonomous driving semantic scene understanding. It contains 22 point cloud sequences and 19 classes for training and evaluation. The large autonomous driving dataset nuScenes provides a substantial number of samples for scene understanding in urban scenes. It contains 700 training scenes, 150 validation scenes, and 150 testing scenes, with a total of 1000 driving scenes and 16 classes.

Implementation Details. Following previous works (Sautier et al. 2022; Mahmoud et al. 2023), only keyframes from the 600 nuScenes training sequences are used for self-supervised pre-training without labelled data. Then, the pre-trained network is fine-tuned on nuScenes to evaluate the pre-training performance and on SemanticKITTI to validate the generalization ability. We follow previous works (Sautier et al. 2022; Mahmoud et al. 2023) to use the Minkowski U-Net (Choy, Gwak, and Savarese 2019) as the 3D network to generate the point cloud feature, which uses $3 \times 3 \times 3$ kernels for all sparse convolutional layers. Based on the 2D MaskCLIP network (Zhou, Loy, and Dai 2022), the attention pooling operation of the CLIP ViT-B (Radford et al. 2021) image encoder is modified to extract image features, text embedding, and pixel-text correspondences. Particularly, the CLIP model is frozen during training, and the generations of class-agnostic masks from SAM ViT-H (Kirillov et al. 2023) and fine-grained region captions from BLIP-2 (Li et al. 2023) are implemented offline. For pre-training, we use the SGD optimizer with a cosine scheduler to pre-train our 3D network for 50 epochs on eight NVIDIA Tesla A100 GPUs with a total batch size of 16. For fine-tuning, we follow the evaluation protocol of (Sautier et al. 2022; Mahmoud et al. 2023) to finetune our pre-trained 3D network on SemanticKITTI and nuScenes by using different proportions of annotated data.

Metric. We adopt the recognized mean Intersection-over-Union (mIoU) score across all classes for 3D autonomous driving semantic segmentation evaluation.

Comparison Results

We evaluate the pre-training effectiveness of our proposed approach against state-of-the-art techniques on two popular autonomous driving datasets. Specifically, we present the results of random initialization and six prior methods in Table 1. We can observe that 1) our proposed VLM2Scene pre-training strategy can achieve impressive performance gain on downstream tasks, especially with limited annotated data for fine-tuning. For instance, our approach provides a 6.84% mIoU improvement for the 1% few-shot fine-tuning task over the second-best ST-SLidR on the nuScenes dataset. 2) Pre-training with our method provides a significant gain of 11.06% for linear probing, exhibiting the better quality of our pre-trained representations. 3) The superior performance on SemanticKITTI demonstrates the strong generalization ability of our method.

Moreover, we present the per-class IoU of different methods when using 1% labelled data for fine-tuning in Table 2. Our approach yields much better results than others for each class, especially for bus, construction vehicle, and truck. We attribute this to the fact that our approach can enhance the semantics of these classes for better representation learning.

Ablation Study

Component Analysis. To verify the effectiveness of our proposed strategies, we conduct the ablation experiments and report in Table 3. The baseline model leverages template-based text prompts and point-level contrastive learning for 3D self-supervised learning. Compared to the baseline model, our region caption prompt (RCP) strategy improves the performance by 4.6% and 3.0% when using 1% and 5% annotated data for fine-tuning, respectively.

Method	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
Random	30.3	0.0	0.0	8.1	65.0	0.1	6.6	21.0	9.0	9.3	25.8	89.5	14.8	41.7	48.7	72.4	73.3
PointContrast	32.5	0.0	1.0	5.6	67.4	0.0	3.3	31.6	5.6	12.1	30.8	91.7	21.9	48.4	50.8	75.0	74.6
DepthContrast	31.7	0.0	0.6	6.5	64.7	0.2	5.1	29.0	9.5	12.1	29.9	90.3	17.8	44.4	49.5	73.5	74.0
PPKT	37.8	0.0	2.2	20.7	75.4	1.2	13.2	45.6	8.5	17.5	38.4	92.5	19.2	52.3	56.8	80.1	80.9
SLidR	38.3	0.0	1.8	15.4	73.1	1.9	19.9	47.2	17.1	14.5	34.5	92.0	27.1	53.6	61.0	79.8	82.3
CLIP2Scene	33.1	0.0	1.9	10.4	70.2	1.5	9.1	41.3	0.0	20.0	28.3	87.8	15.6	37.1	52.7	74.8	77.6
ST-SLidR	40.8	0.0	2.7	16.0	74.5	3.2	25.4	50.9	20.0	17.7	40.2	92.0	30.7	54.2	61.1	80.5	82.9
Ours	**47.6**	0.0	**7.3**	**49.0**	**77.7**	**17.1**	**30.3**	**53.2**	**40.7**	**20.2**	**51.9**	**92.5**	**36.2**	**57.6**	**62.3**	**82.2**	**83.0**

Table 2: Per-class 3D semantic segmentation IoU performance on the nuScenes vaild set when fine-tuning with 1 % labels.

Methods	Components			nuScenes	
	RCP	SFR	RSA	1%	5%
Baseline				38.8	51.6
Ours	✓			43.4	54.6
		✓		43.8	55.0
			✓	42.1	53.6
	✓	✓		46.5	56.9
		✓	✓	45.4	56.3
	✓	✓	✓	**47.6**	**58.1**

Table 3: Ablation Study of each component.

Strategies	Methods	nuScenes	
		1%	5%
RCP	only template prompts	45.4	56.3
	only RCP	46.5	57.1
	Ours	**47.6**	**58.1**
RSC	w point-level	43.4	54.6
	w super-pixel	44.1	55.1
	w/o semantic filtering	45.0	55.9
	Ours	**47.6**	**58.1**

Table 4: Experimental results for different strategies.

These results demonstrate that the fine-grained and specific region caption prompts can provide more meaningful semantic information from the text modality. The semantic-filtered region learning (SFR) and region semantic assignment (RSA) also contribute to a performance gain of 5.0% and 3.3% when using 1% annotated data for fine-tuning, respectively. This can be attributed to the fact that our region-level contrastive learning can capture informative contextual representations, leading to better results. Moreover, the combination of different strategies leads to further gain. Finally, the integration of all our components achieves the best performance, which is 8.8% and 6.5% better than the baseline, respectively. Overall, these results demonstrate the effectiveness of each component and also show the availability of VLMs for 3D scene representation learning.

The Region Caption Prompt Strategy. In Table 4, we conduct a comparative analysis to verify how text prompt affects performance. When only performing template-based text prompts for per-training, the model achieves 45.4% and 56.3% using 1% and 5% labelled data for fine-tuning, respectively. Notably, a significant performance drop can be seen when our region caption prompts (RCP) strategy is omitted during pre-training. This suggests that our RCP is beneficial for rich and informative region semantic understanding via text prompts, promoting the model's capacity. Moreover, we noticed that the template-based text prompts and our region caption prompts can complement each other, further improving our model's performance.

The Region Semantic Concordance Regularization. From Table 4, we utilize different main contrastive learning manners to evaluate our approach. 1) w point-level: we use the pixel-point contrastive learning from CLIP2Scene for pre-training. 2) w super-pixel: we replace the region masks from VLMs with the super-pixel from SLIC (Achanta et al. 2012). 3) w/o semantic filtering: we remove the process of false negative samples as stated in Equation (2). We can see that leveraging region-wise learning manner is better than point-level due to the informative context cues. However, super-pixel-wise and region-mask-wise (i.e., w/o semantic filtering) may suffer from the self semantic conflict issue, resulting in the bottleneck of performance. The experimental results reveal that our region masks generated from our method yield more representative and highly accurate features, and the filtering of false negative samples achieves effective improvement for 3D scene understanding.

Conclusion

In this paper, we explore a VLMs-assisted self-supervised approach for 3D scene understanding in the wild, namely VLM2Scene. To imbue richer semantic information, we propose Region Caption Prompts for more meaningful and fine-grained region-level semantic understanding. Besides, to overcome the limitations of point-level contrastive learning, we propose Region Semantic Concordance Regularization for informative region-level representation enhancement. Extensive experiments verify the superiority of our approach for 3D representation enhancement. We hope that our VLM2Scene will inspire more exciting research in harnessing VLMs for 3D pre-training in the future.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fei, B.; Yang, W.; Liu, L.; Luo, T.; Zhang, R.; Li, Y.; and He, Y. 2023. Self-supervised Learning for Pre-Training 3D Point Clouds: A Survey. *arXiv preprint arXiv:2305.04691*.

Gao, B.; Pan, Y.; Li, C.; Geng, S.; and Zhao, H. 2021. Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods. *IEEE Transactions on Intelligent Transportation Systems*, 23(7): 6063–6081.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12): 4338–4364.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hou, Y.; Zhu, X.; Ma, Y.; Loy, C. C.; and Li, Y. 2022. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8479–8488.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kong, L.; Liu, Y.; Chen, R.; Ma, Y.; Zhu, X.; Li, Y.; Hou, Y.; Qiao, Y.; and Liu, Z. 2023. Rethinking range view representation for lidar segmentation. *arXiv preprint arXiv:2303.05367*.

Lai, X.; Chen, Y.; Lu, F.; Liu, J.; and Jia, J. 2023. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17545–17555.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Liu, Y.-C.; Huang, Y.-K.; Chiang, H.-Y.; Su, H.-T.; Liu, Z.-Y.; Chen, C.-T.; Tseng, C.-Y.; and Hsu, W. H. 2021. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*.

Mahmoud, A.; Hu, J. S.; Kuai, T.; Harakeh, A.; Paull, L.; and Waslander, S. L. 2023. Self-Supervised Image-to-Point Distillation via Semantically Tolerant Contrastive Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7102–7110.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–824.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sautier, C.; Puy, G.; Gidaris, S.; Boulch, A.; Bursuc, A.; and Marlet, R. 2022. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9891–9901.

Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; and Han, S. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, 685–702. Springer.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; and Zhao, H. 2022. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35: 33330–33342.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. 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*, 574–591. Springer.

Zhang, X.; Liao, G.; Gao, W.; and Li, G. 2022. TDRNet: Transformer-Based Dual-Branch Restoration Network for Geometry Based Point Cloud Compression Artifacts. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Zhang, Z.; Girdhar, R.; Joulin, A.; and Misra, I. 2021. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10252–10263.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*, 696–712. Springer.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; and Lin, D. 2021. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9939–9948.