See More and Know More: Zero-shot Point Cloud Segmentation via Multi-modal Visual Data

Yuhang Lu[1,*]**, Qi Jiang**[1,*]**, Runnan Chen**[2]**, Yuenan Hou**[3]**, Xinge Zhu**[4]**, Yuexin Ma**[1,†]

[1] ShanghaiTech University [2] The University of Hong Kong
[3] Shanghai AI Laboratory [4] The Chinese University of Hong Kong

{luyh2,jiangqi,mayuexin}@shanghaitech.edu.cn

Abstract

Zero-shot point cloud segmentation aims to make deep models capable of recognizing novel objects in point cloud that are unseen in the training phase. Recent trends favor the pipeline which transfers knowledge from seen classes with labels to unseen classes without labels. They typically align visual features with semantic features obtained from word embedding by the supervision of seen classes' annotations. However, point cloud contains limited information to fully match with semantic features. In fact, the rich appearance information of images is a natural complement to the textureless point cloud, which is not well explored in previous literature. Motivated by this, we propose a novel multi-modal zero-shot learning method to better utilize the complementary information of point clouds and images for more accurate visual-semantic alignment. Extensive experiments are performed in two popular benchmarks, i.e., SemanticKITTI and nuScenes, and our method outperforms current SOTA methods with 52% and 49% improvement on average for unseen class mIoU, respectively.

1. Introduction

Point cloud segmentation is a critical task for 3D scene understanding, which promotes the development of autonomous driving, assistive robots, digital urban, AR/VR, etc. Fully supervised methods [74, 64, 29, 34] have achieved impressive performance. However, there exist tremendous categories of objects in the real world, especially in large-scale outdoor scenes, bringing challenges for such methods to generalize to novel objects without labels in training data. Furthermore, manual annotations for 3D point clouds are extremely time-consuming and expensive. Zero-shot learning can recognize unseen objects by utilizing side information, especially the word embedding, to transfer the knowledge of seen categories to unseen
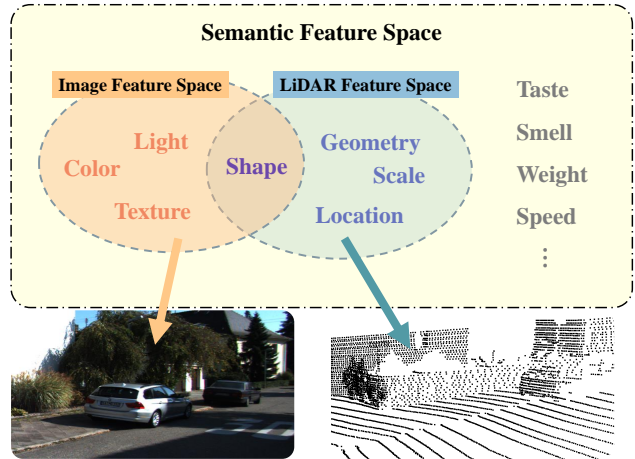
---

*Equal contribution. † Corresponding author.

Figure 1. Semantic features of objects obtained by word embedding contain rich and diverse information, including appearance characteristics existing in images(*i.e.*, color, light), geometry and location information contained in LiDAR point clouds(*i.e.*, scale, shape), and some other non-visual properties(*i.e.*, smell, weight). Previous image-based or point cloud-based zero-shot learning only considers the alignment between uni-modal visual features and semantic features, where the former can just match a small subset of the latter. We propose a more effective solution for zero-shot 3D segmentation by using multi-modal visual features.

ones, which is important for the point cloud segmentation in large-scale scenes.

Zero-shot semantic segmentation on 2D images has made promising progress in the past few years [61, 3, 8, 38, 17, 27]. There are two main streams of methods, including generative methods and projection-based methods, which inspire the following research works on 3D point clouds. For generative methods [45, 42], they usually train a fake feature generator supervised by seen classes and fine-tune a classifier for recognizing real seen-class features and synthesized unseen-class features. However, 3D features are more difficult to generate than 2D features due to higher dimensional information, making such strategies perform un-

satisfactorily on 3D point clouds. Moreover, these methods require additional training efforts when new unseen categories appear, which limits the generalization capability on real-world applications. For projection-based approaches [15], they target to align visual features to corresponding semantic features by the seen-class supervision, so that unseen class can be recognized by leveraging the similarity between its visual features and semantic features. Such methods can be easily generalized to novel classes without retraining. However, visual features extracted from the point cloud can only match a subset of semantic features and yield limited performance, as shown in Fig. 1.

In fact, current autonomous vehicles and robots are usually equipped with multiple sensors, where LiDAR and camera are the most common ones [9, 6]. Since point cloud contain accurate location and geometric information and images provide rich color and texture characteristics, many researchers focus on exploring sensor-fusion methods [16, 41, 59] for achieving more precise perception. Considering that see more and know more, we aim to make these two uni-modal visual data complement each other and generate more comprehensive visual features to better align with semantic features for more effective zero-shot learning. To our knowledge, we are the **first** to explore zero-shot learning based on multi-modal visual data.

In this paper, we focus on transductive generalized zero-shot learning for point cloud-based semantic segmentation, where both seen and unseen classes will appear in one scene but only objects of seen classes have labels during training. Based on the input of the synchronized point cloud and image, we propose a novel zero-shot point cloud segmentation method. Specifically, we propose an effective multi-modal feature fusion approach, termed **Semantic-Guided Visual Feature Fusion (SGVF)**, to obtain a more comprehensive visual feature representation, where valuable information from two uni-modal visual features are adaptively selected under the guidance of semantic features. As opposed to previous sensor-fusion methods, our strategy is more flexible and applicable for zero-shot learning by introducing semantic features to play an active role in the visual feature fusion stage. In this condition, exactly valid information can be utilized for the following semantic-visual feature alignment. Then, the knowledge of seen classes can be effectively transferred to unseen classes. Furthermore, to reduce the semantic-visual domain gap in advance, we propose **Semantic-Visual Feature Enhancement (SVFE)** to enhance both semantic features and visual features by transferring the domain knowledge, such as relationships among classes, to each other, which definitely benefits the following SGVF and the final semantic-visual alignment process. Actually, our method can be easily extended to more visual modalities.

We conduct extensive comparisons with current 2D and 3D zero-shot segmentation methods and our method outperforms others significantly on different datasets and settings. The effectiveness of each module of our method is also verified by ablation studies. In summary, our contributions are summarized as follows:

- We propose a novel multi-modal zero-shot approach for point cloud semantic segmentation.

- We design an effective feature-fusion method with semantic-visual feature enhancement, which can better align visual features with semantic features to benefit the recognition of unseen classes.

- Our method achieves state-of-the-art performance on SemanticKITTI and nuScenes datasets.

2. Related Work

2.1. Zero-Shot Learning

Zero-shot learning aims at transferring the knowledge learned from seen categories to unseen ones. Many zero-shot learning studies[10, 33, 68, 48, 35, 1, 60, 36, 7, 23, 40, 26] leverage intermediate representations such as semantic embeddings and attributes to bridge the gap between seen and unseen classes. Early works of zero-shot learning (ZSL) [1] only recognize unseen classes of data during inference. While the recently discussed generalized zero-shot learning (GZSL) [55] requires model to recognize both seen and unseen classes, which is more challenging yet practical since real scenes usually contain both seen and unseen classes of objects at the same time. Apart from ZSL and GZSL, zero-shot learning tasks can also be classified as inductive[2, 11, 25, 49, 70, 54] and transductive[56, 71, 67, 28, 32]. The former excludes the occurrence of samples of unseen classes yet the latter permits. In this paper, we focus on the setting of transductive GZSL, which is more practical for real-world applications.

2.2. Zero-shot Segmentation on 2D Image

Zero-shot segmentation on 2D images has been widely explored [61, 3, 8, 38, 17, 27, 31, 30, 50, 44, 72, 63] in the past few years, which can be divided into projection-based methods and generative methods. Projection-based methods like SPNet[61] intend to align visual feature space with semantic feature space so as to generalize the model to unseen data by leveraging the structure of the semantic feature space. Since the training process only involves labels of seen classes, many methods [61, 12] try to alleviate the bias toward seen classes during training. By contrast, generative methods[8, 38, 17, 27] usually adopt a multi-stage training paradigm with a fake feature generator supervised with seen data and a classification layer fine-tuned by real seen-class features and synthesized unseen-class features. Following

the astonishing zero-shot transfer learning performance of CLIP [53], a series of works[65, 43, 24, 37, 73] begin to exploit its huge potential under ZS3 task and have made significant improvement. However, data leakage is a concern since unseen objects may already occur in the CLIP training data and it is also difficult to extend to 3D tasks due to the lack of huge 3D pre-training samples.

2.3. Zero-shot Segmentation on Point Cloud

The boom of autonomous driving and the expensive 3D manual annotation has led to zero-shot point cloud perception becoming an emerging research hotspot. Many works[21, 18, 20, 19, 42, 45, 13, 14, 15] focusing on zero-shot learning for 3D point clouds appear, especially for point cloud classification [21, 18, 20, 19]. Cheraghian et al.[21] uses PointNet[52] to extract point cloud features and leverages W2V[47] or Glove[51] as extra semantic features. Then, many researchers try to address the hubness problem [18, 20] and extend zero-shot learning to the transductive setting [19].

To the best of our knowledge, only three papers[42, 45, 15] propose solutions for zero-shot point cloud semantic segmentation. Among them, 3DGenZ[45] and SeCondPoint[42] are generative approaches. They all generate fake features of unseen classes with semantic features for training the classifier to achieve the zero-shot transfer. However, generative methods require extra training efforts when new unseen categories appear, which limits the generalization capability. In addition, since 3D features are more complex than 2D features, the generated feature distribution does not fit the original distribution well, resulting in poor results. Different from them, TGP[15] is a projection-based approach, which learns geometric primitives to facilitate the knowledge transfer from seen classes to unseen categories. However, only relying on the point cloud properties, the alignment between visual space and semantic space is difficult since there is a huge domain gap between these two spaces and the point cloud could only match a subset of semantic space, causing incomplete knowledge for unseen objects.

2.4. Multi-Sensor Fusion

Considering that the image contains rich appearance features and the point cloud possesses accurate location and geometry features, many works [62, 58, 59, 66, 75, 5, 41, 16, 39] explore effective fusion ways to make these two sensors complement each other for more precise 3D perception. PointPainting[58] and PointAugmenting[59] utilize the semantic label or feature at the projected image location as additional information to append to the corresponding point, while such point-level fusion strategy will lose dense appearance feature of images. PMF[75] performs perspective projection on the point cloud and performs fea-

ture fusion in the camera coordinate system. 2DPASS[66] leverages knowledge distillation for cross-modal knowledge transfer. Recently, transformer-based sensor-fusion methods [5, 16, 22, 39] achieve promising performance for 3D perception with learnable projection and the usage of the global context. However, these feature-fusion methods are designed for fully supervised tasks. For zero-shot segmentation, direct feature fusion operation results in a more complex fused visual feature, making the alignment to semantic features more difficult. We allow semantic features to adaptively select desired features from two visual modalities for matching, avoiding the interference of irrelevant information.

3. Methods

3.1. Problem Formulation

Point cloud semantic segmentation aims at classifying each point into a specified class. Similar to [15, 45], we divide all classes into seen and unseen ones. We focus on the generalized transductive zero-shot point cloud segmentation problem, which is a more realistic setting where the model needs to segment both the seen and unseen classes in the scene by seeing their features and supervised by the labels of only seen classes.

Let $P \in \mathbb{R}^{T \times 3}$ denote one frame of point cloud with T points represented by (x, y, z) coordinates, and $X \in \mathbb{R}^{3 \times \mathcal{H} \times \mathcal{W}}$ denote the corresponding image, where $\mathcal{H} \times \mathcal{W}$ means the image size. The set of seen and unseen classes are expressed as $C^s = \{c_i^s\}_{i=1}^{N^s}$ and $C^u = \{c_i^u\}_{i=1}^{N^u}$ ($C^s \cap C^u = \emptyset$), respectively, where s, u stand for seen and unseen categories and N^s and N^u denote the number of data samples involving seen and unseen categories. $W^s = \{w_i^s\}_{i=1}^{N^s}$ and $W^u = \{w_j^u\}_{j=1}^{N^u}$ are the word embedding [15] of seen and unseen class names from the word2vec [46] or glove [51], respectively. Since we focus on the transductive zero-shot segmentation setting, the training set is defined as $D_{train} = \left\{ (P_i^s, X_i^s, W_i^s, Y_i)_{i=1}^{N^s}, (P_j^u, X_j^u, W_j^u)_{j=1}^{N^u} \right\}$, where Y is the ground truth label for seen categories.

3.2. Overview

As illustrated in Fig. 2, our method consists of four main modules, including Feature Extraction, Semantic-Visual Feature Enhancement (SVFE), Semantic-Guided Visual Feature Fusion (SGVF), and Semantic-Visual Alignment. Firstly, following [15], we use a 3D backbone network to produce the point visual representation F_l, and utilize the 2D backbone network to extract image visual representation F_i. Meanwhile, a Multi-Layer Perception (MLP) $G(\cdot)$ is used to project the word embedding W into F_s as the semantic feature of specific categories. Secondly, to reduce huge domain gaps between visual and semantic features, we design SVFE to make these two feature space interact
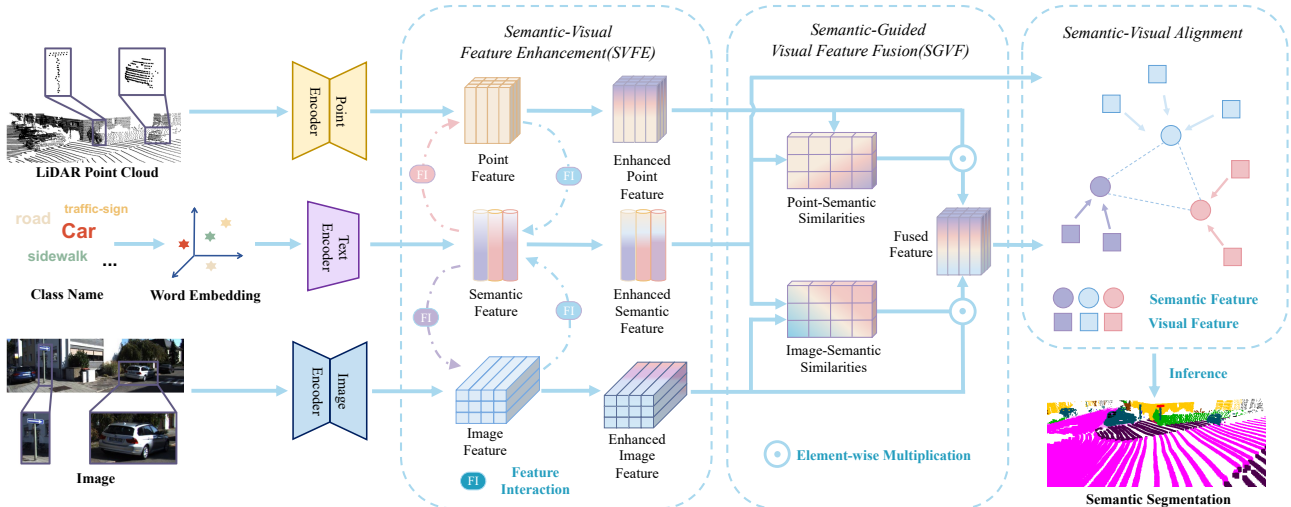
Figure 2. Method overview. Firstly, 3D and 2D backbones extract visual features from LiDAR point cloud and image, while MLP extracts semantic features. Secondly, for reducing the semantic-visual gap, visual features and semantic features interact with each other by learnable projection in the SVFE module. Then, we make semantic features adaptively select valuable visual features from two modalities for effective feature fusion in the SGVF module. Finally, we perform semantic-visual feature alignment for zero-shot learning.

knowledge with each other to enhance the feature representation by the cross-attention mechanism. Then, we propose SGVF to make semantic features automatically select valuable information from two visual modalities for better feature alignment. In the end, we perform semantic-visual feature alignment for transferring the knowledge from seen objects to unseen ones for zero-shot point cloud segmentation.

In the following sections, we will introduce the technical details of modules according to the following order: SGVF, SVFE, and Semantic-Visual Alignment.

3.3. Semantic Guided Visual Feature Fusion

Point cloud contains precise location and geometry information, while images provide rich texture and color information. The combination of both visual features can better match semantic features extracted from language descriptions, which may contain the information of diverse properties of the category. Therefore, we propose to leverage multi-modal visual data for semantic-visual alignment to solve zero-shot learning problems. However, direct fusing point cloud feature and image feature in the data level [58] or feature level [59] will make the fused feature more complex, resulting in difficulty in aligning with semantic features. We design an adaptive selection mechanism for semantic features, where the network can learn valuable information from two visual modalities automatically under the semantic guidance and integrate them together as the richer visual feature.

Based on the point cloud feature F_{el}, image feature F_{ei}, and semantic feature F_{es} gained from the last module (Section. 3.4), we search valuable visual features for semantic

features from the 3D point cloud and 2D image by calculating the weight matrix w_{3D} and w_{2D}, respectively. It is conducted by utilizing the multi-head attention [57]:

$$w_{3D} = \text{MultiHeadAttention}(F_{es}, F_{el}),$$
$$w_{2D} = \text{MultiHeadAttention}(F_{es}, F_{ei}). \tag{1}$$

The weights stand for the significance of two uni-modal visual features to semantic features. And the fused visual feature can be obtained by applying element-wise multiplication between the weight matrix and visual features. Then, we obtain the final fused visual feature by employing an MLP:

$$F_{fusion} = \text{MLP}(\text{softmax}(\text{stack}(w_{3D} \odot F_{el}, w_{2D} \odot F_{ei}))). \tag{2}$$

In this way, the network utilizes multi-modal visual data effectively by selecting valuable information to match semantic features for different categories of objects, which can benefit the following alignment of semantic and visual spaces, thus improving the recognition ability of unseen objects.

3.4. Semantic-Visual Feature Enhancement

During the selection step in SGVF, the huge domain gap between visual features and semantic features will hinder the learning process for fusing effective visual features. Therefore, we consider narrowing the semantic-visual gap in advance by transferring the knowledge, such as relationships among various categories, between semantic and visual space. We conduct the knowledge interaction by the cross-attention mechanism, which can learn the semantic-

visual projection automatically and enhance each feature with valuable knowledge of the other.

Semantic Feature Enhancement. To enhance the semantic feature F_s by visual features, we take F_s as the query q and visual feature as key k and value v to feed into a Transformer Decoder [57] as follows.

$$
\begin{aligned}
\text{TD}(q, k, v) &= \text{Linear}(\text{LN}(\text{MLP}(Q) + Q)), \\
Q &= \text{LN}(\text{CrossAttention}(q, k, v) + q),
\end{aligned}
\tag{3}
$$

where Linear indicates linear mapping layer and LN denotes layer normalization. Because we have two modalities of visual features, we make them interact with the semantic features in order. Considering that we target for point cloud segmentation, we first enhance feature F_s by point feature F_l to pull the representation of two spaces closer, and then conduct the same operation on image feature F_i for further semantic feature enhancement. Formula. 4 illustrates the process.

$$
F_{es} = \text{TD}(\text{TD}(F_s, F_l, F_l), F_i, F_i).
\tag{4}
$$

Visual Feature Enhancement. Similarly, we enhance the visual feature by querying to semantic feature and fetching knowledge from semantic space. Then, we obtain enhanced point feature by $F_{el} = \text{TD}(F_l, F_s, F_s)$, and enhanced image feature by $F_{ei} = \text{TD}(F_i, F_s, F_s)$.

In this way, we reduce the difference between semantic and visual rpresentations by feature interaction and the enhanced features can further facilitate the visual feature selection process in SGVF and the final alignment between semantic and visual spaces.

To demonstrate the effectiveness of SVFE and SGVF intuitively, we select one scene from SemanticKITTI validation set and visualize semantic and visual features of all classes occurred in Fig. 3. It is obvious to see that our model pulls visual features to corresponding semantic features gradually by effective semantic-visual feature enhancement and multi-modal visual feature fusion.

3.5. Semantic-Visual Alignment

Through the feature enhancement of SVFE and multi-modal visual feature fusion of SGVF, we obtain a comprehensive representation of visual features, which can match more content and represent similarly with semantic features. We align visual and semantic feature spaces by the supervision of seen classes. Therefore, the knowledge of seen class can be transferred to unseen class with the aid of side information, e.g., semantic features from word embedding.

Loss function. Following TGP[15], we adopt a cross entropy loss and an unknown-aware InfoNCE loss to distinguish various seen classes and allow the model to identify whether an object is a seen class or an unseen class.
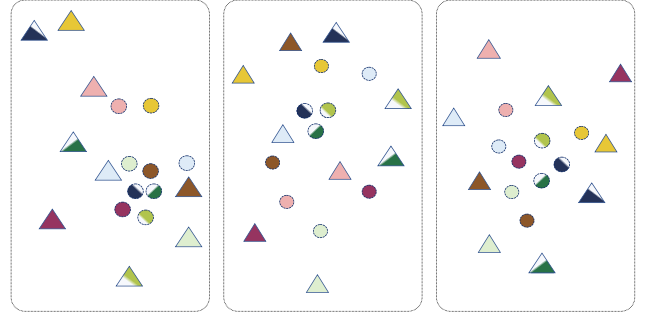

(a) Before SVFE   (b) After SVFE   (c) After SGVF

Figure 3. Visualization of semantic-visual feature relationships of various classes in one scene of SemanticKITTI by t-SNE. Triangles indicate semantic features, circles denote visual features, pure colors stand for seen classes, and gradient colors mean unseen classes. The same color represents the same class. Visual features in (a) and (b) come from the point cloud branch.

In order to obtain a better semantic-visual space alignment, we have to ensure that the distribution of each class is compact within classes and distinguishable between classes. To this end, we use the following objective function.

$$
L_s = -\log \sum_i^{N^s} \sum_t^{T_i} \frac{\exp(D(f_i^t, e_{y_i^t})/\tau)}{\sum_{c=1}^{C^s + C^u} \exp(D(f_i^t, e_c)/\tau)},
\tag{5}
$$

where f_i^t denotes the visual features of the t-th point in the i-th sample, $e_{y_i^t}$ is the corresponding ground truth semantic representation. τ is the inversed temperature term. C^s and C^u are the number of seen and unseen classes, respectively. $D(\cdot)$ is the similarity function between visual and semantic features. In this paper, we choose the dot product similarity.

Since only seen classes have annotations during training, the zero-shot model is naturally biased towards the seen classes. To avoid this, we push the features of the seen and unseen classes apart by the following loss formula.

$$
L_u = \log \sum_j^{N^u} \sum_t^{T_j} \frac{\sum_{c=1}^{C^s} \exp(D(f_j^t, e_c)/\tau)}{\sum_{\hat{c}=1}^{C^s + C^u} \exp(D(f_j^t, e_{\hat{c}})/\tau)}.
\tag{6}
$$

The overall loss function is the combination of two losses.

$$
L = L_s + L_u.
\tag{7}
$$

Inference. We infer a new scene with fused visual feature $F_{fusion} = \left\{ f_{fusion}^t \right\}_{t=1}^{T}$ and semantic feature $F_{es} = \{e_c\}_{c=1}^{C^s + C^u}$, where T is the number of points in this scene. The class of the tth point is determined as follows.

$$
C_t = \arg\max_c \frac{\exp\left(D(f_{fusion}^t, e_c)\right)}{\sum_{\hat{c}=1}^{C^s + C^u} \exp\left(D(f_{fusion}^t, e_{\hat{c}})\right)}.
\tag{8}
$$

Table 1. Comparison with state-of-the-art methods on SemanticKITTI and nuScenes datasets. We show the performance of diverse unseen-class settings introduced in Section. 4.1. Setting "0" indicates fully supervised manner. "Improvement" means the percentage improvement in the metric unseen mIoU for our method relative to the previous SOTA method. "Supervised" denotes that both seen and unseen classes have labels during the training of our method, which stands for the upper bound for zero-shot learning performance.

Setting	Model	SemanticKITTI					nuScenes				
		Seen mIoU	Uneen mIoU	Improvement	Overall mIoU	Overall hIoU	Seen mIoU	Uneen mIoU	Improvement	Overall mIoU	Overall hIoU
0	TGP[15]	-	-	-	59.1	-	-	-	-	67.9	-
	Ours	-	-	-	**62.6**	-	-	-	-	**69.1**	-
2	3DGenZ[45]	40.9	12.4	-	37.9	19.0	**67.8**	4.2	-	**59.9**	7.9
	TGP[15]	58.3	28.8	+3.5%	55.2	38.6	58.9	26.9	+25.3%	54.9	36.9
	Ours	**59.5**	**29.8**	-	**56.4**	**39.7**	59.4	**33.7**	-	56.2	**43.0**
	Supervised	61.5	71.8	-	62.6	66.3	70.1	61.9	-	69.1	65.7
4	3DGenZ[45]	41.4	10.8	-	35.0	17.1	**67.2**	3.1	-	51.2	5.9
	TGP[15]	54.6	17.3	+54.9%	46.7	26.3	65.7	14.8	+56.1%	53.0	24.2
	Ours	**58.8**	**26.8**	-	**52.1**	**36.8**	66.4	**23.1**	-	**55.6**	**34.3**
	Supervised	60.3	71.2	-	62.6	65.3	71.9	60.6	-	69.1	65.8
6	3DGenZ[45]	40.3	6.5	-	29.6	11.2	53.8	3.2	-	34.8	6.0
	TGP[15]	53.6	13.3	+79.7%	40.9	21.3	**68.8**	14.1	+56.7%	48.3	23.4
	Ours	**56.6**	**23.9**	-	**46.3**	**33.6**	66.8	**22.1**	-	**50.0**	**33.2**
	Supervised	56.8	75.3	-	62.6	64.8	74.5	60.1	-	69.1	66.5
8	3DGenZ[45]	38.3	1.3	-	22.7	2.5	36.5	2.1	-	19.3	4.0
	TGP[15]	**53.2**	8.6	+70.9%	**34.4**	14.8	**68.4**	13.7	+56.9%	41.1	22.8
	Ours	46.0	**14.7**	-	32.8	**22.3**	68.2	**21.5**	-	**44.9**	**32.7**
	Supervised	52.1	77.1	-	62.6	62.2	73.5	64.7	-	69.1	68.8

4. Experiments

We first introduce the datasets, evaluation metrics, and implementation details. Then we show results and analysis of extensive comparison experiments and ablation studies to verify the effectiveness and superiority of our method.

4.1. Dataset and Category Division

SemanticKITTI [6] contains 22 sequences, where ten sequences are for training, sequence 08 for validation, and the remaining sequences are used for testing. It has annotations for 20 classes in total. For a full evaluation, we conduct diverse zero-shot settings with different numbers of unseen classes, including **2**-motorcycle/truck, **4**-bicyclist/traffic-sign, **6**-car/terrain, **8**-vegetation/sidewalk. The classes in the unseen set increase incrementally for different settings. Especially, the setting **4** with motorcycle, truck, bicyclist, and traffic-sign is following [15] and is taken as the main setting for ablation study.

nuScenes [9] contains 40157 annotated samples with 6 monocular camera images with $360°$ FoV and a 32-beam LiDAR scan. It has annotations for 17 classes in total. We conduct several zero-shot settings with different numbers of unseen classes, including the **2**-Motorcycle/trailer, **4**-terrain/traffic-cone, **6**-bicycle/car, **8**-vegetation/sidewalk. The classes in the unseen set increase incrementally for different settings. The rest classes are taken as seen classes.

4.2. Evaluation Metrics

We report the mean-intersection-over-union(mIoU) of seen classes, unseen classes, and all classes, respectively. Following [15], we utilize the harmonic mean IoU (hIoU) to demonstrate the overall performance of methods.

$$\text{hIoU} = \frac{2 \times \text{mIoU}_{seen} \times \text{mIoU}_{unseen}}{\text{mIoU}_{seen} + \text{mIoU}_{unseen}}, \quad (9)$$

where mIoU_{seen} and mIoU_{unseen} represents the mIoU of seen classes and unseen classes, respectively.

4.3. Implementation Details

Following [15], we use Cylinder3D [74] to extract LiDAR point cloud features and ResUnet to extract image features. The visual feature dimension is 128. We adopt W2V [47] and Glove [51] to embed the class name and obtain the word embedding features (600-dimensional vector) as the auxiliary information for zero-shot segmentation. The $G(\cdot)$ is a two-layer MLP with the dimension of 96 and 128, then we obtain a 128-dimensional semantic feature. For SVFE, the transformer decoder is comprised of cross-attention and MLP. We use one decoder and the number of heads is 4. For SGVF, we utilize multi-head attention (the number of heads is 4) to compute the similarity between semantic features and visual features. Our method is built on the Pytorch platform, optimized by Adam. The learning rate for the backbone is 0.001, while the learning rate for SVFE and
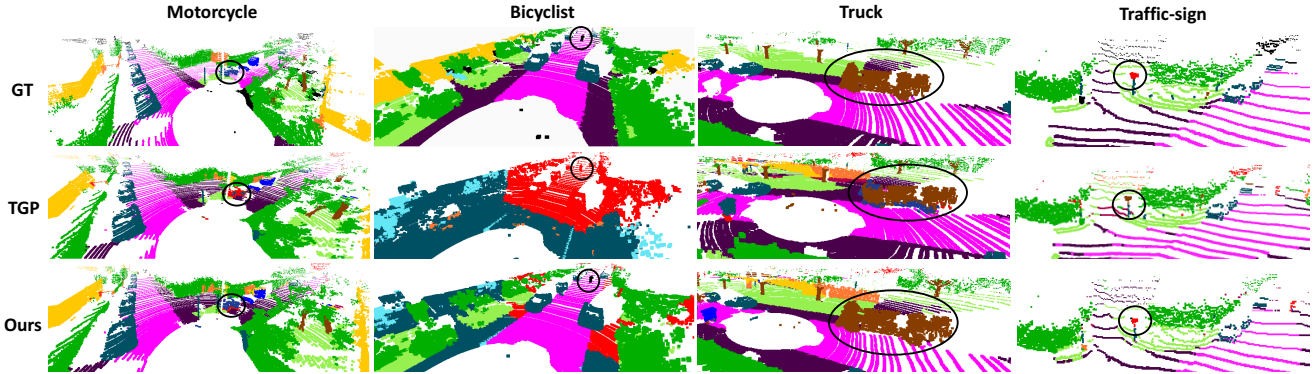
Figure 4. Visualization of results on SemanticKITTI. We show the ground truth, segmentation results of TGP[15], and segmentation results of our method in rows. Objects highlighted by black circles are unseen classes, including motorcycle, bicyclist, truck, and traffic-sign. It is obvious that our model classifies unseen classes more accurately and is closer to ground truth.

Table 2. Comparison with state-of-the-art 2D methods on SemanticKITTI dataset with 4-unseen-class setting. We extend those 2D zero-shot methods to solve 3D zero-shot point cloud segmentation, where "⋆" indicates results reported in [15].

Model	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
SPNet[61]	57.0	0.0	45.0	0.0
ZS5Net⋆ [8]	53.2	5.1	43.1	9.3
PMOSR⋆ [69]	55.1	8.7	45.3	15.0
JoEm [4]	56.7	2.8	45.4	5.3
Ours	**58.8**	**26.8**	**52.1**	**36.8**

SGVF is 0.0003. The batch sizes for both SemanticKITTI and nuScenes are 4. It costs 80 hours to train 40 epochs on four RTX 3090 GPUs for the SemanticKITTI dataset and costs 45 hours to train 20 epochs for the nuScenes dataset.

4.4. Comparison Results

In this section, we show results of our method and compare it with current state-of-the-art 3D zero-shot segmentation methods, including 3DGenZ[45] and TGP[15]. Extensive experiments with different unseen class settings are conducted on SemanticKITTI and nuScenes datasets to comprehensively evaluate the methods' performance. Qualitative analysis is also provided. In particular, due to limited works on 3D zero-shot segmentation, we also adapt 2D SOTA methods to solve 3D tasks for further verification. In addition, since our method is based on multimodal fusion, we also compare with current popular multimodal fusion strategies used for full supervised tasks. Our method outperforms others in recognizing unseen classes of objects by a large margin. Detailed analysis is as follows.

Comparision with 3D methods. As shown in Table. 1, our method achieves SOTA performance on 2-, 4-, 6- and 8-unseen-class settings on both SemanticKITTI and nuScenes datasets, outperforming previous SOTA methods with im-

provement rates of 52% and 49% in average for unseen class mIoU, respectively. For the comprehensive evaluation metric hIoU concerning both seen and unseen classes, our method is also superior to others by a large margin. Qualitative results are provided in Figure 4, where we visualize the segmentation results of our method and TGP[15] on SemanticKITTI dataset alongside the ground truth annotations. Our method makes more accurate predictions on unseen classes. For example, TGP identifies a motorcycle as a traffic sign (first column) and takes parts of a truck as cyclist (third column), while our model accurately segments corresponding categories. It illustrates that multi-modal visual feature in our method really benefits the semantic-visual matching and further boost the unseen class recognition.

Comparision with extensions of 2D methods. Because 3D zero-shot segmentation just gets noticed recently, there are limited research works for comparison. Meanwhile, 2D zero-shot segmentation has been well explored and they can also provide essential inspirations for solving 3D tasks. Thus, for more adequate validation, we compare our method with four representative 2D methods by using their released source code, namely SPNet[61], ZS5Net[8], PMOSR[69] and JoEm[4], on SemanticKITTI dataset. Some modifications are imposed on the source code for adapting to 3D point cloud data. We replace all 2D segmentation backbones with Cylinder3D[74], the same as ours. Stacked calibration is applied in JoEm with $\gamma = 0.08$ on softmax scores but is not used in SPNet due to terrible performance. The center loss is adopted instead of the BAR loss in JoEm since the feature interpolation is hard to implement in 3D sparse convolution. The results are shown in Table. 2. Notably, these 2D methods have limited performance when adapting to 3D tasks with more complex 3D features. We also try some generative methods [17, 27] but fail for the similar reason that they usually rely on high-quality 2D feature map to train their generators, but 3D feature is challenging for generation. Our method is superior due to the

Table 3. Comparision with other fusion methods on SemanticKITTI dataset with the 4-unseen-class setting.

Model	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
PointPainting[58]	**58.9**	16.3	49.9	25.5
PointAugmenting[59]	57.9	15.0	48.9	23.8
2DPASS[66]	57.4	13.0	48.1	21.2
PMF[75]	56.9	14.1	47.9	22.6
Deepfusion[39]	54.9	15.8	46.7	24.5
TransFuser [22]	54.1	11.9	45.2	19.5
Ours	58.8	**26.8**	**52.1**	**36.8**

Table 4. Ablation experiments of the module of our framework on SemanticKITTI dataset with the 4-unseen-class setting.

Model	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
Ours	58.8	**26.8**	**52.1**	**36.8**
Ours w/o SGVF	58.8	23.4	51.3	33.5
Ours w/o SVFE	**59.0**	19.9	50.8	29.8
Ours w/o Image	58.3	20.0	50.2	29.8

rational exploration and utilization of each modal feature.

Comparison with popular multi-modal fusion methods. While it is intuitive that multi-sensor-based methods naturally outperform single-sensor-based methods since extra visual information is exploited, designing effective sensor-fusion methods for zero-shot tasks is non-trivial because we have to consider the complex projection relationship between semantic information and visual features. To verify the superiority of our multi-modal fusion approach, we apply two data-level fusion methods[59, 58] and four feature-level fusion methods[66, 75, 39, 22], including two transformer-based methods[39, 22], to the 3D zero-shot segmentation task by using the same baseline as our proposed method, which is a TGP[15] models trained by ourselves for fair comparisons. As the results in Table. 3 shown, all previous camera-LiDAR-fusion methods gain limited or no improvements compared with the baseline because they fuse visual features directly without considering the semantic guidance, which is not suitable for zero-shot learning. In contrast, our method outperforms others by around 10% since it allows semantic features to adaptively select valid LiDAR and image features for fusion, avoiding unnecessary information, and benefiting the knowledge transfer from seen classes to unseen classes.

4.5. Ablation Studies

In this section, we conduct ablation studies on the SemanticKITTI dataset to verify the effectiveness of proposed modules in our network. Additionally, further analysis of the internal design of the SVFE and SGVF modules can be found in Appendix A and B.

Effect of SGVF. To verify the effectiveness of our feature-fusion strategy, we keep the backbone network and SVFE module and adopt a simple concatenation fusion instead of SGVF. We concatenate F_{el} and F_{ei} and utilize an MLP to compress the features to 128 dimensions. Then we perform visual-semantic alignment for the fused visual feature and F_{es}. As shown in Table. 4, compared with using SGVF, the unseen mIoU drops about 3.5%, illustrating that SGVF fuses valid information and effectively transfers knowledge from seen classes to unseen ones.

Effect of SVFE. To demonstrate the advantage of SVFE, we maintain the backbone network and the features extracted from the backbone are directly fed into SGVF. As Table. 4 shows, without SVFE, the unseen mIoU drops about 7%, showing that the huge semantic-visual gap leads to the difficulty of feature alignment in the joint space, while SVFE reduces the gap by feature enhancement.

Effect of image modality. We also conduct ablation study for the image modality in Table. 4 by keeping the LiDAR backbone and its branch in the SVFE module. For semantic-visual alignment, we only utilize the single modal point cloud feature. We can see that compared with the multimodal setting, the unseen mIoU drops about 7% without the appearance feature. Our method takes advantage of both sensors to match semantic feature space, which achieves significant improvement. It is worth noting that even with the LiDRA-only setting, our method is still superior to TGP (Table. 1) due to our effective semantic-visual feature enhancement.

5. Conclusions

We make the first attempt to investigate the potential of multi-modal visual data in solving the transductive generalized zero-shot point cloud semantic segmentation problem. We have designed an effective multi-modal fusion method with mutual feature enhancement, which can adaptively determine what information should be taken from each modality under the semantic guidance for better semantic-visual alignment. Our method achieves SOTA performance on two large-scale datasets under diverse zero-shot settings.

6. Acknowledgements

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. *CVPR*, pages 819–826, 2013. 2

[2] Zeynep Akata, Scott E. Reed, and etc. Evaluation of output embeddings for fine-grained image classification. *CVPR*, pages 2927–2936, 2015. 2

[3] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. *ICCV*, pages 9516–9525, 2021. 1, 2

[4] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *ICCV*, pages 9536–9545, October 2021. 7

[5] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, pages 1090–1099, June 2022. 3

[6] J. Behley, M. Garbade, and etc. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *ICCV*, 2019. 2, 6

[7] Max Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. *ICCVW*, pages 2666–2673, 2017. 2

[8] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, volume 32. Curran Associates, Inc., 2019. 1, 2, 7

[9] 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. *arXiv preprint arXiv:1903.11027*, 2019. 2, 6

[10] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. *CVPR*, pages 5327–5336, 2016. 2

[11] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. *CVPR*, pages 5327–5336, 2016. 2

[12] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. *ArXiv*, abs/1605.04253, 2016. 2

[13] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models, 2023. 3

[14] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip, 2023. 3

[15] 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. 2, 3, 5, 6, 7, 8

[16] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022. 2, 3

[17] Jiaxin Cheng and etc. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *ICCV*, pages 9556–9566, October 2021. 1, 2, 7

[18] 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. 3

[19] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *WACV*, pages 923–933, 2020. 3

[20] Ali Cheraghian, Shafin Rahman, Townim F Chowdhury, and etc. Zero-shot learning on 3d point cloud objects and beyond. *IJCV*, 130(10):2364–2384, 2022. 3

[21] Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-shot learning of 3d point cloud objects. In *MVA*, pages 1–6. IEEE, 2019. 3

[22] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE TPAMI*, pages 1–18, 2022. 3, 8

[23] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. *ICCV*, pages 1241–1250, 2017. 2

[24] Jian Ding, Nan Xue, Guisong Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. *CVPR*, pages 11573–11582, 2022. 3, 12

[25] Andrea Frome, Gregory S. Corrado, and etc. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 2

[26] Chuang Gan, Ming Lin, and etc. Exploring semantic interclass relationships (sir) for zero-shot action recognition. In *AAAI*, 2015. 2

[27] Zhangxuan Gu, Siyuan Zhou, and etc. Context-aware feature generation for zero-shot semantic segmentation. In *ACM MM*. ACM, oct 2020. 1, 2, 7

[28] Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. Transductive zero-shot recognition via shared model space learning. In *AAAI*, 2016. 2

[29] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. *CVPR*, pages 13085–13094, 2021. 1

[30] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. In *NeurIPS*, 2020. 2

[31] Naoki Kato, T. Yamasaki, and Kiyoharu Aizawa. Zero-shot semantic segmentation via variational mapping. *ICCVW*, pages 1363–1370, 2019. 2

[32] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. *ICCV*, pages 2452–2460, 2015. 2

[33] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. *CVPR*, pages 4447–4456, 2017. 2

[34] 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, 2023. 1

[35] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *CVPR*, pages 951–958, 2009. 2

[36] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36:453–465, 2014. 2

[37] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ArXiv*, abs/2201.03546, 2022. 3, 12

[38] Peike Li and etc. Consistent structural relation learning for zero-shot segmentation. In *NeurIPS*, 2020. 1, 2

[39] Yingwei Li and etc. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, pages 17182–17191, 2022. 3, 8

[40] Yan Li, Zhen Jia, and etc. Deep semantic structural constraints for zero-shot learning. In *AAAI*, 2018. 2

[41] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*, 2022. 2, 3

[42] Bo Liu, Shuang Deng, Qiulei Dong, and Zhanyi Hu. Language-level semantics conditioned 3d point cloud segmentation. *arXiv preprint arXiv:2107.00430*, 2021. 1, 3

[43] Timo Lüddecke and etc. Image segmentation using text and image prompts. *CVPR*, pages 7076–7086, 2022. 3, 12

[44] Fengmao Lv, Haiyang Liu, Yichen Wang, Jiayi Zhao, and Guowu Yang. Learning unbiased zero-shot semantic segmentation networks via transductive transfer. *IEEE SPL*, 27:1640–1644, 2020. 2

[45] Björn Michele, Alexandre Boulch, and etc. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *3DV*, pages 992–1002, 2021. 1, 3, 6, 7

[46] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 3

[47] Tomas Mikolov, Ilya Sutskever, and etc. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 26, 2013. 3, 6

[48] Ashish Mishra, M. Shiva Krishna Reddy, and etc. A generative model for zero shot learning using conditional variational autoencoders. *CVPRW*, pages 2269–22698, 2018. 2

[49] Mohammad Norouzi and etc. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 2

[50] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. *CVPRW*, pages 2687–2696, 2021. 2

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 3, 6

[52] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, July 2017. 3

[53] Alec Radford, Jong Wook Kim, and etc. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 12

[54] Bernardino Romera-Paredes and etc. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 2

[55] Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE TPAMI*, 35:1757–1772, 2013. 2

[56] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. *CVPR*, pages 1024–1033, 2018. 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, volume 30. Curran Associates, Inc., 2017. 4, 5

[58] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *CVPR*, June 2020. 3, 4, 8

[59] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, pages 11794–11803, 2021. 2, 3, 4, 8

[60] Yongqin Xian, Zeynep Akata, and etc. Latent embeddings for zero-shot classification. *CVPR*, pages 69–77, 2016. 2

[61] Yongqin Xian, Subhabrata Choudhury, and etc. Semantic projection network for zero- and few-label semantic segmentation. In *CVPR*, June 2019. 1, 2, 7

[62] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, June 2018. 3

[63] Jiarui Xu, Shalini De Mello, and etc. Groupvit: Semantic segmentation emerges from text supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18123, 2022. 2

[64] Jianyun Xu, Ruixiang Zhang, and etc. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. *ICCV*, pages 16004–16013, 2021. 1

[65] Mengde Xu, Zheng Zhang, and etc. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *ArXiv*, abs/2112.14757, 2021. 3, 12

[66] Xu Yan, Jiantao Gao, Chaoda Zheng, Chaoda Zheng, Ruimao Zhang, Shenghui Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *ECCV*, 2022. 3, 8

[67] Yunlong Yu, Zhong Ji, Xi Li, Jichang Guo, Zhongfei Zhang, Haibin Ling, and Fei Wu. Transductive zero-shot learning with a self-training dictionary approach. *TCYB*, 48:2908–2919, 2018. 2

[68] Éloi Zablocki and etc. Context-aware zero-shot learning for object recognition. *ArXiv*, abs/1904.12638, 2019. 2

[69] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. *ICCV*, pages 6954–6963, 2021. 7

[70] Ziming Zhang and etc. Zero-shot learning via semantic similarity embedding. *ICCV*, pages 4166–4174, 2015. 2

[71] An Zhao, Mingyu Ding, and etc. Domain-invariant projection learning for zero-shot recognition. *ArXiv*, abs/1810.08326, 2018. 2

[72] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. *CVPR*, pages 2593–2602, 2021. 2

[73] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3, 12

[74] 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. *arXiv preprint arXiv:2011.10033*, 2020. 1, 6, 7

[75] Zhuangwei Zhuang and etc. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *ICCV*, pages 16280–16290, October 2021. 3, 8

Appendix A. More Details of SVFE

Why SVFE improves the performance? The main function of the SVFE module is to narrow the semantic-visual gap and facilitate early knowledge transfer between semantic and visual spaces, rather than simply scaling up the model. To demonstrate the importance of the semantic-visual interaction, we conduct an experiment where we replace it with self-attention operation with the same parameter scale for each single modality. The results in Table.5 show the performance drops sharply without the SVFE module.

Table 5. Ablation experiments of the design of SVFE module on SemanticKITTI dataset with the 4-unseen-class setting,

Model	Seen mIoU	Unseen mIoU	Overall mIoU	hIoU
baseline	54.6	17.3	46.7	26.3
baseline + self attention	57.3	19.4	49.3	29.0
baseline + SVFE	**58.8**	**23.4**	**51.3**	**33.5**
image features first(F'_{es})	58.3	16.1	49.4	25.2
point cloud features first(F_{es})	**58.8**	**26.8**	**52.1**	**36.8**

Does fusion order in SVFE matter? As mentioned in Sec 3.4, semantic feature enhancement is implemented as: $F_{es} = \text{TD}(\text{TD}(F_s, F_l, F_l), F_i, F_i)$. We provide the result of fusing image visual features first and then point cloud visual features: $F'_{es} = \text{TD}(\text{TD}(F_s, F_i, F_i), F_l, F_l)$. As shown in Table.5, The ordering of feature fusion presented in the paper is superior because visual features extracted from point clouds are more central to 3D point cloud segmentation. By fusing these visual features with semantic features first, we are able to provide better guidance for the segmentation process.

Appendix B. More Details of SGVF

Are there any better fusion methods than SGVF module? As SGVF adopts an attention-based design, to further validate the effectiveness of the SGVF module, we design experiments to compare our method with two variants of transformer-based multimodal fusion methods, as shown in Table.6. We find that the performance of "w/o SGVF, w/ cross attention", which uses LiDAR to query image features for fusion without considering semantic features, is not as good as our SGVF module. This is consistent with our intuition that simply fusing the visual features without considering the semantic information is not sufficient for zero-shot tasks. However, the result of "w/ SGVF, w/ self attention" is unexpected. The performance of the method with the added self-attention mechanism for the fused features is lower than that of SGVF, even though the parameter quantity is increased. This suggests that simply increasing the model complexity does not necessarily lead to better performance. In fact, the additional self-attention mechanism

may have introduced noise and decreased the discriminative power of the fused features.

Table 6. Ablation experiments of the design of SGVF module on SemanticKITTI dataset with the 4-unseen-class setting,

Model	Seen mIoU	Unseen mIoU	Overall mIoU	hIoU
w/o SGVF, w/ cross attention	56.6	21.9	49.3	31.6
w/ SGVF, w/ self attention	50.4	21.2	44.3	29.8
Ours	**58.8**	**26.8**	**52.1**	**36.8**

Appendix C. Model inference time

With the addition of an extra image modality, our model's inference time is **0.097 seconds per frame**, slightly larger than **0.087s/f** of the SOTA single-modal method TGP[13]. But our model outperforms it with more than 50% improvement of unseen category mIOU. Furthermore, it yields real-time performance (All of the results are tested on 1 NVIDIA GTX3090 GPU).

Appendix D. The impact of various image encoders on performance

We employed ResUnet-34 as our image encoder (L591). To show the impact of various image encoders, we replace the encoder with ResUnet-18 and ResUnet-50 and get comparable performance, as shown in the below table.

Model	Seen mIoU	Unseen mIoU	Overall mIoU	hIoU
ResUnet-18	57.3	24.7	50.4	34.5
ResUnet-50	**58.9**	**27.1**	**52.2**	**37.1**
Ours(ResUnet-34)	58.8	26.8	52.1	36.8

Appendix E. Discussion on the CLIP Model

Given the success of the CLIP [53] model in 2D zero-shot segmentation [65, 43, 24, 37, 73], we aim to investigate its potential for 3D point cloud semantic segmentation by incorporating the CLIP model into our method. We follow the approach used in MaskCLIP [73], where the class name is inserted into 85 hand-crafted prompts and they are fed into CLIP's text encoder to generate multiple text features. Additionally, we replace the 2D ResUNet backbone with MaskCLIP+. As shown in Table 7, even though unseen objects may already occur in the CLIP training data, causing data leakage, the incorporation of the CLIP model still performs worse than our **pure zero-shot method**. It is mainly because CLIP is based on the contrastive learning between image and text pairs and the significant disparity between

Table 7. CLIP model experiment on SemanticKITTI dataset with the 4-unseen-class setting.

Model	Seen mIoU	Unseen mIoU	Overall mIoU	hIoU
Ours ← CLIP model	56.6	14.1	47.7	22.6
Ours	**58.8**	**26.8**	**52.1**	**36.8**

point cloud features and image features makes point cloud visual features difficult to align with semantic features extracted by CLIP. However, it is interesting to explore the projection between point cloud and images and transfer the knowledge learnt by CLIP to solve 3D zero-shot problems in large scenarios.