Hierarchical Intra-modal Correlation Learning for Label-free 3D Semantic Segmentation

Xin Kang[1*] Lei Chu[2] Jiahao Li[2] Xuejin Chen[1] Yan Lu[2]
[1] University of Science and Technology of China [2] Microsoft Research Asia

Abstract

Recent methods for label-free 3D semantic segmentation aim to assist 3D model training by leveraging the open-world recognition ability of pre-trained vision language models. However, these methods usually suffer from inconsistent and noisy pseudo-labels provided by the vision language models. To address this issue, we present a hierarchical intra-modal correlation learning framework that captures visual and geometric correlations in 3D scenes at three levels: intra-set, intra-scene, and inter-scene, to help learn more compact 3D representations. We refine pseudo-labels using intra-set correlations within each geometric consistency set and align features of visually and geometrically similar points using intra-scene and inter-scene correlation learning. We also introduce a feedback mechanism to distill the correlation learning capability into the 3D model. Experiments on both indoor and outdoor datasets show the superiority of our method. We achieve a state-of-the-art 36.6% mIoU on the ScanNet dataset, and a 23.0% mIoU on the nuScenes dataset, with improvements of 7.8% mIoU and 2.2% mIoU compared with previous SOTA. We also provide theoretical analysis and qualitative visualization results to discuss the mechanism and conduct thorough ablation studies to support the effectiveness of our framework.

1. Introduction

Label-free 3D semantic segmentation, which aims to achieve scene understanding without reliance on labeled data, has recently emerged as a vital research topic. This task holds significant value for practical applications, including autonomous driving, robotic navigation, and augmented reality, where the collection of 3D annotations is expensive and novel objects may appear.

Existing methods [4,5,8,15,19,21,31,36,37,40] leverage the open-world recognition capability of pre-trained vision language models, such as CLIP [22] and MaskCLIP [39], to train 3D models through cross-modal transfer learning. These vision language models generate semantic features from texts and images to provide semantic guidance for 3D model training. However, pre-trained on image classification tasks, these vision language models struggle to generate consistent dense semantic predictions [4, 39]. For instance, as shown in Fig. 1, the predictions for pixels belonging to the same *chair* may differ within each view and across multiple views. The predictions for *chairs* in

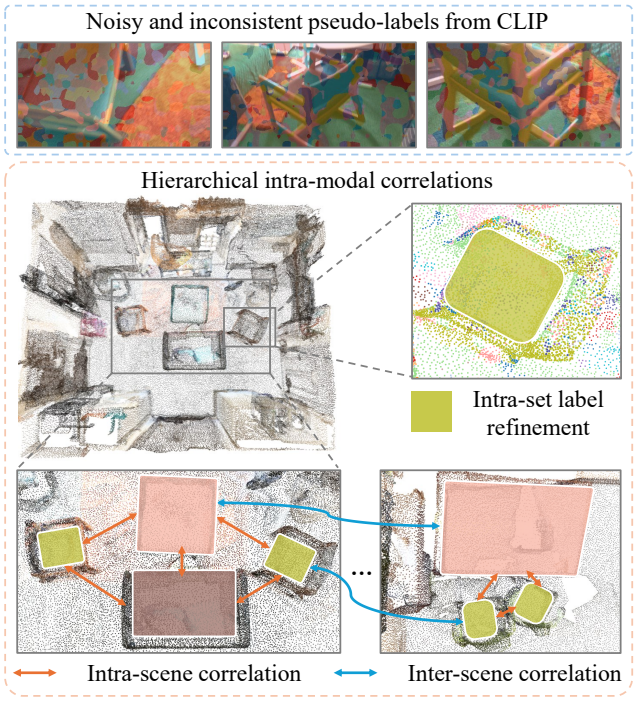


Figure 1. To address the noisy and inconsistent pseudo-label challenge, we design a hierarchical intra-modal correlation learning framework, including intra-set label refinement, intra-scene correlation learning, and inter-scene correlation learning. Intra-set label refinement reduces label inconsistency for points with similar geometric attributes in each local set. For intra-scene correlation learning, we model point correlations in each scene and constrain visually and geometrically similar points to be closer in feature space. For inter-scene correlation learning, we use cross-scene point correlations to help constrain the consistency of feature distribution in different scenes.

*Xin Kang did this work during his internship at Microsoft Research Asia.

different scenes may also be inconsistent. Such ambiguous guidance poses significant challenges for the 3D model to learn stable visual representations. Recently, Segment Anything (SAM) [16] has been proposed to pretrain models on dense prediction tasks and can obtain accurate object masks. Leveraging SAM, concurrent work Chen et al. [4] use a label refinement strategy to mitigate noisy supervision in each mask, resulting in significant improvements. However, the use of SAM comes with a great increase in training costs. More importantly, the supervision inconsistency across multiple views and different scenes still exists.

This inconsistency issue is caused by the unstable features learned by the vision language model from images with occlusions and a lack of geometric information. We observe that point cloud data is free of occlusion and rich in geometric clues, thus facilitating the learning of stable point features. This allows us to establish reliable correlations between different points, thus providing strong guidance to help maintain feature compactness under inconsistent supervision. As shown in Fig.1, intra-modal correlations in 3D scenes can be categorized into three aspects: **Intra-set correlation**: Local point sets sharing similar geometric attributes usually can help generate cleaner and sharper segmentation boundaries when compared to the boundaries generated by the vision language model. This can be used to reduce the noise and inconsistency of pseudo-labels and thus improve local feature coherence. **Intra-scene correlation**: Besides aligning 3D-2D and 3D-text features, aligning the features of objects with similar appearance and geometry during training can assist the 3D model in mitigating disruptions caused by ambiguous supervision from the vision language model. This can help learn a more focused and concise feature space. **Inter-scene correlation**: Aligning the features of objects with similar appearance and geometry in different scenes can further address the inter-scene contradictory semantic guidance, leading to consistent feature distributions in various scenes.

In this paper, we present a *hierarchical* intra-modal correlation learning framework to leverage the three aforementioned correlations. **(1)** We use an *intra-set label refinement* scheme that statistically analyzes the pseudo-labels within each geometric consistency set and refines the pseudo-labels to encourage fewer label conflicts. **(2)** We propose the *intra-scene correlation learning* module to capture point feature correlations between different objects and thus constrain visually and geometrically similar points to be closer in feature space. **(3)** We introduce the *inter-scene correlation learning* module that leverages cross-scene attention to model correlations among objects in different scenes, promoting the 3D model to learn stable feature distributions. Finally, we design a *feedback mechanism* that aligns the output features of the 3D model with the final aggregated point features, thereby distilling the correlation learning capability into the 3D model. Experiments on both indoor and outdoor datasets demonstrate the superiority of our method. We achieve a state-of-the-art (SOTA) mIoU of 36.6% on the ScanNet dataset, surpassing the previous SOTA method CLIP2Scene [5] by 7.8% mIoU. On the nuScenes dataset, we achieve 23.0% mIoU and surpass CLIP2Scene by 2.2% mIoU. Theoretical analysis, qualitative visualization, and extensive ablation studies further support the effectiveness of our framework.

The key contributions can be summarized as follows.

- We propose a novel *hierarchical intra-modal correlation learning* framework for label-free 3D semantic segmentation that leverages intra-modal correlations at three levels: intra-set, intra-scene, and inter-scene, to capture visual and geometric correlations hierarchically and thus assist in learning compact 3D features.
- We present a comprehensive theoretical analysis, qualitative visualization, extensive ablation studies, and a thorough discussion of our framework's mechanism.
- Our method achieves promising results on both indoor and outdoor datasets, showing significant improvement over previous SOTA methods.

2. Related Work

2.1. Label-free 3D Semantic Segmentation

Label-free 3D semantic segmentation, with the goal of achieving scene understanding independent of labeled data, has attracted significant attention in recent years. Existing works [4, 5, 8, 15, 18, 19, 21, 23, 31, 32, 36, 37, 40] mainly utilize powerful pre-trained vision language models [16, 22, 39] to extract semantic knowledge from image and text modalities and distill these semantic knowledge into a 3D model through cross-modal transfer learning. Specifically, these methods first construct dense correspondences between 3D points, 2D images, and text descriptions, and then optimize the similarities between point features and text embeddings according to 2D pseudo-labels. However, since vision language models are typically pre-trained on image classification tasks and lack dense prediction constraints, the pseudo-labels generated from dense visual features are usually noisy and inconsistent. This results in ambiguous semantic guidance for the 3D model training and hinders its performance. To address this issue, concurrent work by Chen et al. [4] use segmentation masks from Segment Anything [16] to reduce the noise in 2D pseudo-labels and improve the 3D model's performance. Nonetheless, SAM brings considerable costs in the training process. A comparison of training costs is provided in Section 5 of the supplementary material. Besides, the inconsistency of pseudo-labels still exists in different views in each scene

and across different scenes and impedes the 3D model to learn stable visual representations. In this work, we design a hierarchical intra-modal correlation learning framework that models visual and geometric correlations between points in each local neighborhood, in each scene, and across different scenes. Based on the learned multi-scale correlations, we draw the deep features of points with high correlations closer, leading to a concise 3D feature space with fewer conflicts.

2.2. Scene Context Learning

Scene context learning utilizes semantic and spatial correlations among visual elements to facilitate various scene-understanding tasks, such as semantic segmentation and object detection. Existing methods can be mainly categorized into three directions: convolutional-based, attention-based, and graph-based methods. Convolutional-based methods aim to design various sparse convolutional kernels [6, 12–14, 27] to progressively integrate multi-scale point features and help the 3D model capture both local details and global context. Another research line focuses on attention mechanisms [9, 10, 17, 26, 29, 35, 38] to model semantic and spatial correlations between different points, enabling the 3D model to learn robust contextual point representations. Graph-based methods [1, 25, 28, 30, 33, 34] treat scenes as graphs and utilize graph convolutional networks to propagate information between different visual items. By leveraging the inherent structure of the data, these methods effectively capture the visual correlations within each scene. In this work, we propose using flexible and adaptive attention mechanisms on point clouds to capture visual correlations in 3D scenes for label-free 3D semantic segmentation.

3. Method

Following the cross-modal transfer learning framework in CLIP2Scene [5], we optimize a 3D model using 2D pseudo-labels generated by MaskCLIP [5, 39]. In Sec. 3.1, we formulate the cross-modal transfer learning framework, while in Sec. 3.2, we present our hierarchical scene correlation learning framework that leverages visual and geometric correlations to build a compact feature space. We also introduce a feedback distillation module in Sec. 3.3 to incorporate the correlation learning ability into the 3D model. The overall training objective is explained in Sec. 3.4.

3.1. Cross-modal Transfer Learning

Given a scene S, its projected image I under certain camera position, and class prompts $\{T_c\}_{c\in[1,C]}$, the pre-trained vision language model maps each image pixel I_{ij} to feature space $\boldsymbol{f}_{ij}^I \in \mathbb{R}^d$, and each class prompt T_c to text embedding $\hat{\boldsymbol{T}}_c \in \mathbb{R}^d$, where C is the number of categories. Subsequently, with given similarity measurement ψ, the pixel-wise pseudo label can be defined as $l_{ij} = \arg\max_c \psi(\boldsymbol{f}_{ij}^I, \hat{\boldsymbol{T}}_c)$. In the meanwhile, a point cloud P can be sampled from the scene and be processed with a 3D encoder mapping each point p_m to $\boldsymbol{f}_m^P \in \mathbb{R}^d$. Then we project point p_m back to the image plane of I filtered by depth and transfer the pseudo label of I_{ij} to p_m, resulting in $(\boldsymbol{f}_m^P, \hat{\boldsymbol{T}}_{l_{ij}})$, which we call a *pair*. For simplicity of expression, we denote $\boldsymbol{f}_m$ as the m-th point feature, and l_m as the m-th paired pseudo label. With the variation of S under different projection camera positions, we can collect a set of such pairs denoted as $\mathcal{M}$. In the end, a cross-entropy loss is used for cross-modal feature alignment as follows,

$$\mathcal{L}_{ce} = -\sum_{\mathcal{M}} \log \frac{\exp(\psi(\boldsymbol{f}_m, \hat{\boldsymbol{T}}_{l_m}))}{\sum_c \exp(\psi(\boldsymbol{f}_m, \hat{\boldsymbol{T}}_c))}. \tag{1}$$

Denote $h_{m,c} = \psi(\boldsymbol{f}_m, \hat{\boldsymbol{T}}_c)$ and $y_{m,l_m} = \frac{\exp(h_{m,l_m})}{\sum_c \exp(h_{m,c})}$. The gradient of m-th item of loss function with respect to $h_{m,c}$ can be expressed as

$$\frac{\partial \mathcal{L}_m}{\partial h_{m,c}} = \begin{cases} y_{m,c} - 1, & c = l_m; \\ y_{m,c}, & c \neq l_m. \end{cases} \tag{2}$$

During the optimization process, this gradient forces h_{m,l_m} to increase and $h_{m,c}$ $(c \neq l_m)$ to decrease, driving $\boldsymbol{f}_m$ towards $\hat{\boldsymbol{T}}_{l_m}$ and away from $\hat{\boldsymbol{T}}_c$ $(c \neq l_m)$. This results in an alignment between the point features and text embeddings.

Following CLIP2Scene [5], we use MaskCLIP [39] as the pre-trained vision language model and an inner product function $\psi = \langle \cdot, \cdot \rangle$ as the similarity measurement. During training, the pre-trained text embeddings remain fixed, serving as feature anchors for different categories. The loss function $\mathcal{L}_{ce}$ optimizes the 3D model by pulling point features towards corresponding text embeddings according to the 2D pseudo-labels. However, due to the inherent noise and inconsistency in pseudo-labels, point features within the same category may be directed towards different text embeddings, resulting in a confusing 3D feature space. To address this, we propose a hierarchical intra-modal correlation learning framework that captures visual and geometric correlations in 3D scenes hierarchically and helps learn more consistent point representations in various environments.

3.2. Hierarchical Scene Correlation Learning

To leverage intra-modal correlations for learning consistent visual features, we design a hierarchical scene correlation learning framework, as shown in Fig. 2, containing three parts, intra-set label refinement, intra-scene correlation learning, and inter-scene correlation learning.

Intra-set label refinement. Pre-trained on image classification tasks, CLIP potentially produces noisy dense predictions and thus hinders the model's performance. Concurrent work Chen et al. [4] address this inconsistency issue

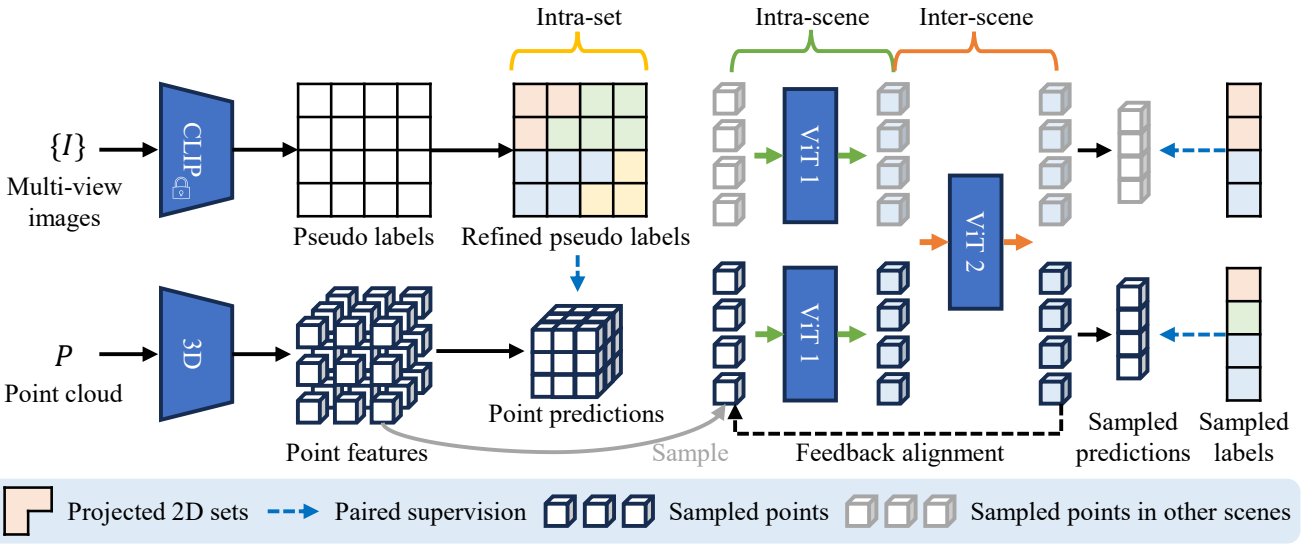


Figure 2. Illustration of the hierarchical intra-modal correlation learning framework. First, we use a CLIP model and a 3D model to generate pixel-wise pseudo-labels and point features. Then we perform intra-modal correlation learning hierarchically. For intra-set label refinement, we cluster points into geometric consistency sets and adopt a voting mechanism to refine pseudo-labels in each set. We extract intra-scene correlation by sampling points in each set and processing them via a vision transformer. For inter-scene correlation learning, we use a cross-scene vision transformer that processes points from multiple scenes. Finally, we align the output features of the 3D model with the final aggregated point features. Note that only the 3D model is retained for inference.

by employing the Segment Anything model [16]. However, it introduces considerable additional training costs. In this work, we choose a more efficient way for label refinement. Our key insight is that geometric clues like normal smoothness can help identify the same semantics for points in the same object part. Specifically, following Chen et al. [3], with a given equivalence relation $\sim$ between points in all the scenes $\cup S_i$, we can obtain the quotient set $\{G_k\} = \cup S_i / \sim$, denoted as *geometric consistency sets*. Then we project points within each set G_k to the image plane of I and denote the projected set as $\hat{G}_k$. In each projected set $\hat{G}_k$, we use a voting mechanism formulated as $\hat{l}_{ij} = \text{Mode}(\{l_{ij}\}_{I_{ij} \in \hat{G}_k})$, and replace the original pseudo-labels with this most frequent label, resulting in refined pseudo-labels $\{\hat{l}_{ij}\}$. In this work, we generate geometric consistency sets using a normal-based over-segmentation algorithm [11, 20]. Other complex equivalences, such as topological, semantic, and functional equivalences, can also be used to generate point sets. We show the refined pseudo-labels for the point cloud in Fig. 3. Through intra-set label refinement, we can efficiently reduce the noise in pseudo-labels and provide locally consistent supervision during training.

Intra-scene correlation learning. Previous cross-modal transfer learning methods typically optimize the 3D model by aligning point features with text embeddings according to 2D pseudo-labels, as formulated in Eq. 1. Given that the 2D pseudo-labels tend to be noisy and inconsistent, the resulting 3D feature space is often ambiguous. To this end, we propose to investigate intra-scene correlations to help constrain features of points with strong correlations to be closer and construct a more focused feature space. Initially, we sample a set of points $\{p_i\}_k$ in each geometry consistency set G_k and concatenate all the sampled points to formulate a subset of the input point cloud, noted as P^S. The number of sampled points M_i is determined by the ratio of the number of points N_i in each set to the total number of points N, i.e., $M_i = M \times N_i / N$. In our experiments, we limit the total number of sampling points to $M = 1024$ to manage GPU memory consumption.

The sampled point features in i-th scene, $\{\boldsymbol{f}_m\}_{p_m \in P_i^S}$, are processed through a transformer block. The m-th output feature of the transformer block can be formulated as

$$\begin{aligned} \tilde{\boldsymbol{f}}_m &= \boldsymbol{f}_m + \sum\nolimits_n w_{mn} \boldsymbol{v}_n, \\ \hat{\boldsymbol{f}}_m &= \tilde{\boldsymbol{f}}_m + \text{Linear}(\tilde{\boldsymbol{f}}_m) \\ &= (\boldsymbol{W} + \boldsymbol{I})(\boldsymbol{f}_m + \sum\nolimits_n w_{mn} \boldsymbol{v}_n) + \boldsymbol{b} \\ &= \boldsymbol{f}'_m + \sum\nolimits_n w_{mn} \boldsymbol{v}'_n + \boldsymbol{b}, \end{aligned} \tag{3}$$

where w_{mn} denotes the attention weight between the m-th and n-th visual features, $\boldsymbol{v}_n$ is the n-th value vector, which is derived by passing $\boldsymbol{f}_n$ through a linear projection layer and $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ are the weight matrix and the bias vector of the linear layer. Since $\boldsymbol{b}$ is unrelated to the gradient for other items, it can be omitted from the loss function. The cross-entropy loss function is formulated as,

$$\hat{\mathcal{L}}_{ce} = -\sum_{\mathcal{M}} \log \frac{\exp(\psi(\boldsymbol{f}'_m + \sum_n w_{mn} \boldsymbol{v}'_n, \hat{\boldsymbol{T}}_{\hat{l}_m}))}{\sum_c \exp(\psi(\boldsymbol{f}'_m + \sum_n w_{mn} \boldsymbol{v}'_n, \hat{\boldsymbol{T}}_c))}, \tag{4}$$

where $\hat{l}_m$ represents the refined pseudo label for point p_m.

Denote $h_{m,c} = \psi(\boldsymbol{f}'_m, \hat{\boldsymbol{T}}_c)$ and $e_{n,c} = \psi(\boldsymbol{v}'_n, \hat{\boldsymbol{T}}_c)$. We get $y_{m,l_m} = \frac{\exp(h_{m,l_m} + \sum_n w_{mn} e_{n,l_m}))}{\sum_c \exp(h_{m,c} + \sum_n w_{mn} e_{n,c}))}$. The gradient of m-th item of the loss function with respect to $h_{m,c}$ is the same as Eq. 2. The gradient for $e_{n,c}$ can be formulated as

$$\frac{\partial \hat{\mathcal{L}}_m}{\partial e_{n,c}} = \begin{cases} w_{mn} \cdot (y_{m,c} - 1), & c = \hat{l}_m, \\ w_{mn} \cdot y_{m,c}, & c \neq \hat{l}_m. \end{cases} \tag{5}$$

The gradient for w_{mn} can be formulated as

$$\frac{\partial \hat{\mathcal{L}}_m}{\partial w_{mn}} = \begin{cases} e_{n,c} \cdot (y_{m,c} - 1), & c = \hat{l}_m, \\ e_{n,c} \cdot y_{m,c}, & c \neq \hat{l}_m. \end{cases} \tag{6}$$

During the optimization process, the gradient in Eq. 5 forces $e_{n,\hat{l}_m}$ to increase and $e_{n,c}$ $(c \neq \hat{l}_m)$ to decrease if $w_{mn} > 0$, denoting that points with higher correlations should move toward the same text embedding. The degree of this constraint is determined by the value of w_{mn}. The gradient in Eq. 6 forces w_{mn} to increase if $e_{n,\hat{l}_m} > 0$ and decrease if $e_{n,c} > 0$ $(c \neq \hat{l}_m)$. This ensures that point features are positioned closer to each other if they surround the same text embedding and farther from each other if they surround different text embeddings.

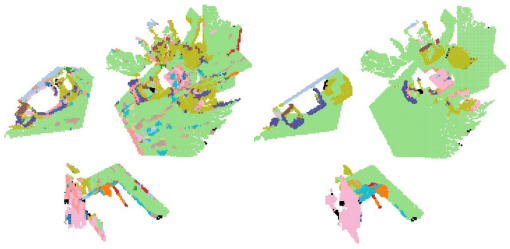
The loss function in Eq. 1 pulls each point feature towards its corresponding text embedding based on pseudo-labels. This can be misleading when pseudo-labels are inconsistent, resulting in a scattered feature distribution. Through our intra-scene correlation learning, the loss function in Eq. 4 further constrains that, (a) points with stronger correlations should be close to the same text embedding, and (b) point features within the same category should be closer together, assisting in the construction of a more focused and concise feature space.

Inter-scene correlation learning. To solve the issue of *inconsistent pseudo labels across different scenes*, we propose an inter-scene attention mechanism to learn feature correlations among objects in different scenes. Specifically, we first include point features from multiple scenes in a training batch as $\{\boldsymbol{f}_m\}_{p_m \in \cup P_i^S}$. Then we process the batched features through a transformer block as

$$\{\hat{\boldsymbol{f}}_m\}_{p_m \in \cup P_i^S} = \text{Transformer}(\{\boldsymbol{f}_m\}_{p_m \in \cup P_i^S}), \tag{7}$$

where the $\text{Transformer}(\cdot)$ function works the same as Eq. 3. By integrating inter-scene attention weights $\{w_{mn}\}$ between $\{\boldsymbol{f}_m\}_{p_m \in P_i^S}$ and $\{\boldsymbol{f}_n\}_{p_n \in P_j^S}$ into $\{\hat{\boldsymbol{f}}_m\}_{p_m \in \cup P_i^S}$, the gradient formulated in Eq. 5 and 6 promotes the compactness of visual features in different scenes during training, leading to a stable and consistent feature distribution. The visualization of intra- and inter-scene attention weights can be found in Section 4 of the supplementary material.

Our theoretical analysis suggests that integrating correlations within the optimization process can effectively impose further constraints on feature compactness among points of high relevance, leading to a more compact feature space with less ambiguity. This enables the 3D model to learn robust features under noisy pseudo-labels.



Pseudo labels Refined labels

Figure 3. Visualization of intra-set label refinement. By replacing the original pseudo-labels with the most frequent labels in each geometric consistency set, we can effectively reduce the noise.

3.3. Feedback Distillation

To distill the correlation learning capability of our hierarchical framework to the 3D model, we align the output features of the 3D model with the final aggregated point features through a widely used Kullback-Leibler (KL) distance loss, which can be formulated as

$$\mathcal{L}_{kl} = \sum_{m=1}^{M} \text{KL}(\boldsymbol{f}_m, \hat{\boldsymbol{f}}_m), \tag{8}$$

where $\hat{\boldsymbol{f}}_m$ is m-th aggregated feature and $\text{KL}(,)$ is the KL distance that measures the distribution difference between two input vectors. This feedback strategy enables the 3D model to integrate the ability to capture visual correlations, thereby generating more robust semantic predictions.

3.4. Overall Objective

Our training phase is segmented into three parts: the first 10 epochs involve training the 3D model with $\mathcal{L}_{ce}$. Starting at epoch 11, we concurrently train the visual transformer by introducing $\hat{\mathcal{L}}_{ce}$, and post-epoch 20, we incorporate $\mathcal{L}_{kl}$ for feedback distillation. We use the average of the losses at each stage of the training and the entire training process is seamless. During the inference stage, we only retain the 3D model for semantic segmentation.

4. Experiment

4.1. Experiment Setup

Datasets and metrics. We conduct experiments on both an indoor dataset ScanNet [7] and an outdoor dataset nuScenes [2]. ScanNet includes 1,603 scanned indoor scenes, where 1,201 scans are for training, 312 scans are for validation, and 100 scans are for testing. The nuScenes dataset contains 1,000 scenes, where 700 scenes are for training, 150 scenes are for evaluation, and 150 scenes are

Methods	Year	ScanNet mIoU (%)	nuScenes mIoU (%)	Time ($GPU \cdot h$)	Parameter (M)
MaskCLIP [39]	2022	14.2	12.8	-	-
MaskCLIP+ [39]	2022	21.6	15.3	-	-
OpenScene [21]	2023	16.8	14.6	24	15.6
CLIP2Scene [5]	2023	25.6	20.8	**4**	**8.2**
Chen et al. [4]†	2023	<u>33.5</u>	**26.8**	20	31.7
Ours	2023	**36.6**	<u>23.0</u>	<u>14</u>	**8.2**

Table 1. Comparison with previous SOTA label-free 3D semantic segmentation methods on ScanNet and nuScenes datasets. The training time is measured on the ScanNet dataset. † indicates concurrent work.

for testing. We use the mean Intersection of Union (mIoU) metric to evaluate the semantic segmentation performance, the training GPU hour to measure the training cost, and the number of model parameters to evaluate the model size.

Implementation details. We follow CLIP2Scene [5] to use MinkowskiNet14 [6] and SPVCNN [24] as the 3D models for ScanNet and nuScenes, respectively. The image features and text embeddings are generated with MaskCLIP [39], which is fixed during training. To extract intra-scene and inter-scene correlations, we utilize two three-layer vision transformers [9] in our experiments: one for intra-scene correlation learning and the other for inter-scene correlation learning. For the 3D model, we use SGD as the optimizer, while AdamW is employed for the vision transformer. We adopt the cosine learning rate decay strategy, setting the initial learning rate at 0.2 for the 3D model and 2e-4 for the vision transformer. Our framework is developed on PyTorch and trained on four NVIDIA Tesla V100 GPUs. For ScanNet, we train our model for 120 epochs over 3.5 hours. The batch size used is 8. For nuScenes, we train our model for 30 epochs over 12.5 hours, with a batch size of 8.

4.2. Comparison Methods

We compare our method with state-of-the-art methods, including MaskCLIP [39], MaskCLIP+ [39], OpenScene [21], CLIP2Scene [5], and Chen et al. [4], on the ScanNet and nuScenes datasets.

MaskCLIP & MaskCLIP+ [39] generate segmentation results through projecting 2D pixel-wise predictions to 3D points according to camera parameters and depth maps. The 2D predictions are derived by evaluating the similarities between the CLIP visual features and text embeddings.

OpenScene [21] combines multi-view image features with point features and evaluates their similarity with CLIP text embeddings to make predictions. For a fair comparison, the reported results are produced by using the same image encoder as used in MaskCLIP.

CLIP2Scene [5] achieves 3D segmentation by utilizing semantic consistency regularization and spatial-temporal consistency regularization to align point features with corresponding text embeddings and concentrating point features in each spatial-temporal interval.

Chen et al. [4] adopt SAM's [16] segmentation masks to refine noisy pseudo-labels from MaskCLIP, thereby improving the 3D model's performance.

Our approach hierarchically utilizes intra-modal correlations to promote the compactness between point features and thereby reduces prediction ambiguity. Our principal innovation capitalizes on intra-modal feature alignment and does not conflict with the concurrent work of Chen et al. [4], which leverages cross-modal feature alignment to learn robust point features from refined pseudo-labels. Finally, we achieve state-of-the-art results without additional priors, like the feature fusion strategy in OpenScene or the SAM model used by Chen et al.

4.3. Label-free 3D Semantic Segmentation

We demonstrate the effectiveness of our hierarchical intra-modal correlation learning framework on both indoor dataset ScanNet and outdoor dataset nuScenes. In Table 1, we report the semantic segmentation metric 'mIoU' on ScanNet and nuScenes, as well as training time and model size for models trained on ScanNet. On the ScanNet dataset, our method outperforms all the state-of-the-art methods by achieving 36.6% mIoU, with a significant gain of 11.0% mIoU compared with CLIP2Scene [5] and 3.1% mIoU gain compared with Chen et al. [4]. Furthermore, the training expenses and model size of our methods are notably minimal. On the nuScenes dataset, we achieve 23.0% mIoU with a gain of 2.2% mIoU compared with CLIP2Scene. All these results show the superiority of our hierarchical intra-modal correlation learning framework. In Fig. 4, we present the qualitative results of our method and CLIP2Scene on the validation set of ScanNet. Note that Chen et al. [4] have not released code yet, so their visualization results are not presented. Compared with CLIP2Scene, our method produces more consistent predictions, such as the *wall* in the first row, the *table* in the second row, and the *sofa* in the last row. Besides, our method can make more accurate recognition for categories with limited training samples, like the *sink* in the first row and the *cabinet* in the second row. By employing our hierarchical intra-model correlation learning framework, the 3D model can generate more coherent and accurate semantic segmentation results.

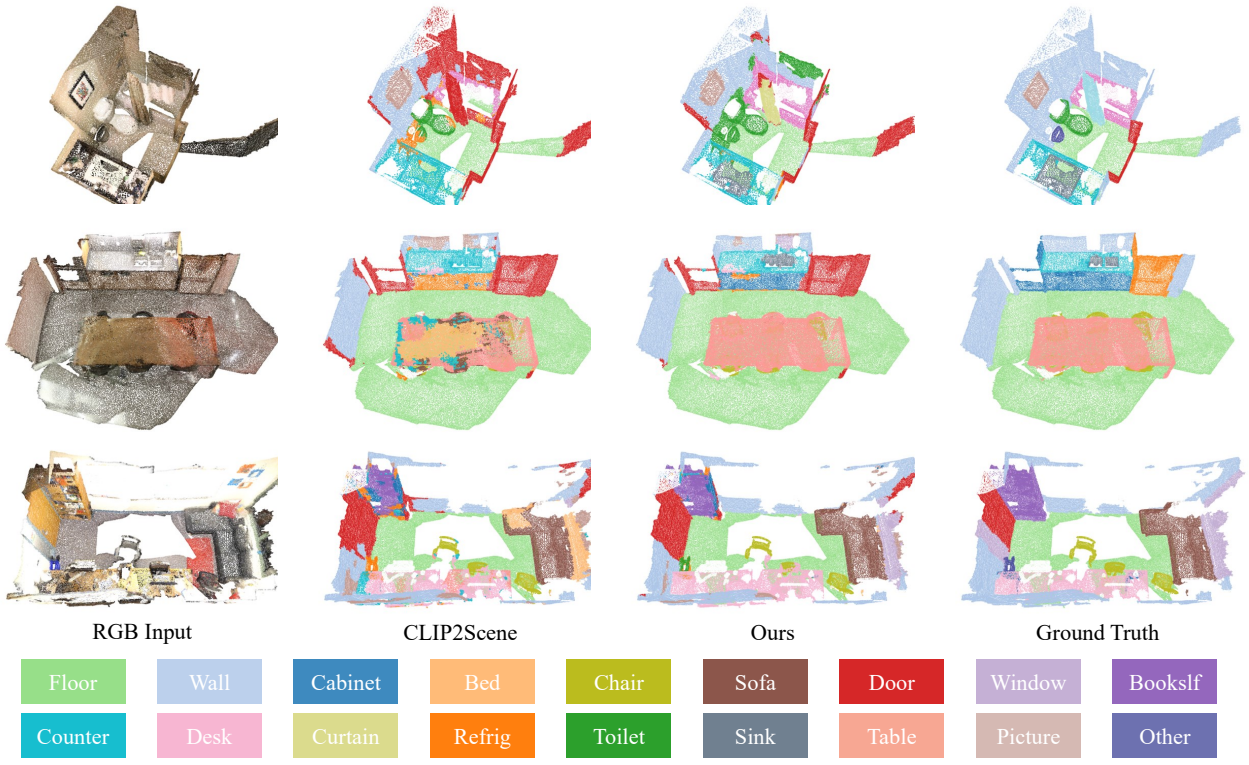


Figure 4. Qualitative comparison for semantic segmentation of our method and CLIP2Scene [5] on the ScanNet dataset. More visualization results are reported in the supplementary material.

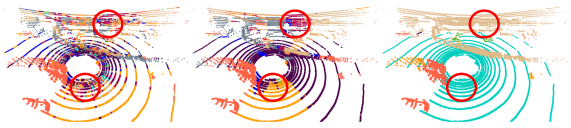


Figure 5. Visualization of pseudo-labels on nuScenes dataset.

We also observe that the improvement of our method on the nuScenes dataset is not as much as that on the ScanNet dataset. This is because the pseudo-labels generated from CLIP [22, 39] in the nuScenes dataset are notably deficient in quality. As shown in Fig. 5, despite the fact that intra-set refinement can improve the local consistency, the overall label quality from CLIP is so poor that erroneous labels constitute the majority. Consequently, the refined labels may not accurately depict the true categories of the points. In the meanwhile, Chen et al. [4] implement an additional step of training a 2D semantic segmentation model. Although this step is costly, it effectively corrects these erroneous labels, contributing to a more substantial improvement.

4.4. Ablation Study

To evaluate the effectiveness of different components of our hierarchical intra-modal correlation learning framework, we conduct a series of ablation experiments on the ScanNet dataset and results are shown in Table 2. All experiments are implemented on four NVIDIA V100 GPUs with a batch size of 4 over 120 training epochs. In EXP I, we reproduce CLIP2Scene [5] as our baseline model, achieving a mIoU of 28.8% on the ScanNet validation set.

EXP	IntraSet	IntraScene	InterScene	FB	mIoU
I	-	-	-	-	28.8
II	✓	-	-	-	31.1 (+2.3)
III	✓	✓	-	-	32.1 (+3.3)
IV	✓	✓	✓	-	36.1 (+7.3)
V	✓	✓	✓	✓	36.6 (+7.8)

Table 2. Ablation study of our hierarchical intra-modal correlation learning framework on the ScanNet validation set. 'FB' denotes the feedback distillation mechanism.

Ablation of intra-set label refinement. In EXP II, we cluster the input point cloud into geometric consistent sets and perform label refinement in each set, leading to a mIoU of 31.1% and a gain of 2.3% mIoU compared to EXP I. In Fig. 3, we demonstrate that intra-set label refinement can produce cleaner pseudo-labels, thus improving the coherence of the 3D model's predictions.

Ablation of intra-scene correlation learning. In EXP III, we incorporate the intra-scene correlation learning module based on EXP II, resulting in a higher gain of 3.3% mIoU. As shown in the second column of Fig. 7, the intra-scene correlation learning enables the model to learn a more con-

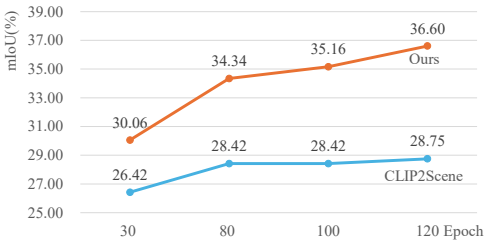


Figure 6. Semantic segmentation results of our method and CLIP2Scene [5] on ScanNet across different training epochs.

centrated and concise feature space, thereby reducing confusion between different categories.

Ablation of inter-scene correlation learning. Building on EXP III, we additionally employ the inter-scene correlation learning module in EXP IV, yielding a gain of 7.3% mIoU. As shown in the third row of Fig. 7, this strategy further constrains that the feature spaces of multiple scenes should be consistent, assisting the model in generating stable feature distributions in different scenes.

Ablation of feedback distillation. In EXP V, we implement the feedback distillation mechanism and achieve a mIoU of 36.6%, with a total gain of 7.8% mIoU compared with CLIP2Scene. By aligning the feature space, we can transfer the correlation learning capacity into the 3D model, further enhancing its performance.

Ablation of training epochs. We compare the segmentation performance of our method with CLIP2Scene [5] as the number of training epochs increases from 30 to 120. The results are presented in Fig. 6. Our method consistently outperforms CLIP2Scene across various training epochs, achieving a maximum mIoU gain of 7.8% at the 120th epoch. Additionally, our method shows a better upward trend with a 1.4% mIoU increase from 100 to 120 epochs, compared to only a 0.3% mIoU increase exhibited by CLIP2Scene across the same epoch range.

Additional ablation studies of geometric consistency sets are detailed in Section 3 of the supplementary material.

4.5. Feature Space Visualization

Our hierarchical intra-modal correlation learning framework aims to create a concise feature space. To validate this, we compare the t-SNE visualizations of the feature space from our method and CLIP2Scene [5] in Fig. 7. The first two rows show the t-SNE map of two unique scenes. We present the feature spaces separately for CLIP2Scene, our method that incorporates intra-scene correlation learning, and our method that integrates inter-scene correlation learning. Compared to CLIP2Scene, both the intra- and inter-scene strategies assist the 3D model in learning a more focused feature space with fewer conflicts, as indicated by the dotted box. This verifies the effectiveness of introducing point correlations to help concentrate point features. In the last row of Fig. 7, we concatenate point features from

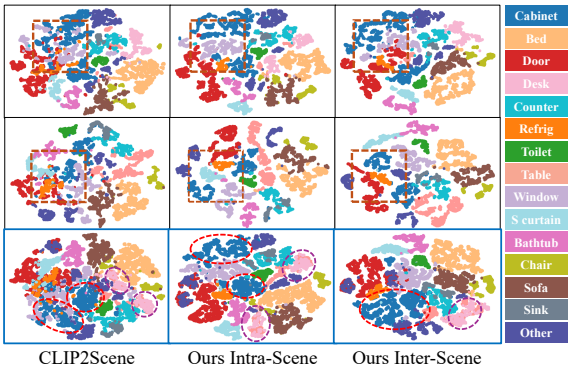


Figure 7. Feature space visualization. We visualize the learned point features of our method and CLIP2Scene [5] using t-SNE maps. The first two rows are t-SNE maps of two distinct scenes and the last row shows the t-SNE map of the concatenated features from these two scenes. *Refrig* and *S curtain* are the abbreviations for *refrigerator* and *shower curtain*, respectively.

the two aforementioned scenes to verify the consistency of feature distributions across different scenes. Compared to CLIP2Scene and the intra-scene correlation learning strategy, our inter-scene setting results in a more stable feature space across diverse scenes. This indicates that incorporating global correlations across multiple scenes can enhance the alignment of feature distributions in different scenes and help the 3D model produce more robust predictions.

5. Limitations and Future Work

Our work demonstrates that exploring intra-modal correlations can help the 3D model training for label-free 3D semantic segmentation. However, we have not yet explored intra-modal correlations within images and texts. By using correlations in images, we can create a more consistent 2D feature space, leading to more accurate and coherent pseudo-labels. Similarly, investigating correlations within texts enables us to use more precise and detailed descriptions, offering rich guidance for cross-modal alignment. In the future, we will extend our hierarchical intra-modal correlation learning framework to images and texts to achieve better label-free 3D semantic segmentation performance.

6. Conclusion

In this paper, we introduce a hierarchical intra-modal correlation learning framework for label-free 3D semantic segmentation. Our method hierarchically utilizes visual and geometric correlations at three scales including intra-set, intra-scene, and inter-scene, to help mitigate noise in pseudo-labels and concentrate point features, resulting in a focused and concise 3D feature space with fewer conflicts. Experiments on both indoor and outdoor datasets demonstrate the superiority of our method. Comprehensive theoretical analysis and extensive ablation studies further support the effectiveness of our framework.

References

[1] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5664–5673, 2019. 3

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[3] Nenglun Chen, Lei Chu, Hao Pan, Yan Lu, and Wenping Wang. Self-supervised image representation learning with geometric set consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19292–19302, 2022. 4

[4] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, ZHU Xinge, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 6, 7

[5] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 1, 2, 3, 6, 7, 8

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 3, 6

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5

[8] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019, 2023. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[10] Nico Engel, Vasileios Belagiannis, and Klaus Dietmayer. Point transformer. *IEEE access*, 9:134826–134840, 2021. 3

[11] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 4

[12] Ben Graham. Sparse 3d convolutional neural networks. *arXiv preprint arXiv:1505.02890*, 2015. 3

[13] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 3

[14] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 3

[15] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023. 1, 2

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4, 6

[17] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8500–8509, 2022. 3

[18] Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*, 2021. 2

[19] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1190–1199, 2023. 1, 2

[20] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2027–2034, 2013. 4

[21] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 1, 2, 6

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 7

[23] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 2

[24] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 6

[25] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 3

[26] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *arXiv preprint arXiv:2305.03045*, 2023. 3

[27] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017. 3

[28] Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21560–21569, 2023. 3

[29] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 3

[30] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 3

[31] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023. 1, 2

[32] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, pages 677–695. Springer, 2022. 2

[33] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 3

[34] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 3

[35] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 3

[36] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022. 1, 2

[37] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 1, 2

[38] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 3

[39] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2, 3, 6, 7

[40] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 1, 2