CLIP-Driven Open-Vocabulary 3D Scene Graph Generation via Cross-Modality Contrastive Learning

Lianggangxu Chen[1] Xuejiao Wang[1] Jiale Lu[1] Shaohui Lin[1,2]
Changbo Wang[1*] Gaoqi He[1,3*]
[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, China
[3]Chongqing Key Laboratory of Precision Optics, Chongqing Institute of East China Normal University, Chongqing, China
{lgxchen,52275901001,jllu}@stu.ecnu.edu.cn, {shlin, cbwang, gqhe}@cs.ecnu.edu.cn

Abstract

3D Scene Graph Generation (3DSGG) aims to classify objects and their predicates within 3D point cloud scenes. However, current 3DSGG methods struggle with two main challenges. 1) The dependency on labor-intensive ground-truth annotations. 2) Closed-set classes training hampers the recognition of novel objects and predicates. Addressing these issues, our idea is to extract cross-modality features by CLIP from text and image data naturally related to 3D point clouds. Cross-modality features are used to train a robust 3D scene graph (3DSG) feature extractor. Specifically, we propose a novel Cross-Modality Contrastive Learning 3DSGG (CCL-3DSGG) method. Firstly, to align the text with 3DSG, the text is parsed into word level that are consistent with the 3DSG annotation. To enhance robustness during the alignment, adjectives are exchanged for different objects as negative samples. Then, to align the image with 3DSG, the camera view is treated as a positive sample and other views as negatives. Lastly, the recognition of novel object and predicate classes is achieved by calculating the cosine similarity between prompts and 3DSG features. Our rigorous experiments confirm the superior open-vocabulary capability and applicability of CCL-3DSGG in real-world contexts.

1. Introduction

3D scene graph generation (3DSGG) is a fundamental task in scene understanding, which aims to detect objects and represent their relationships using predicates in 3D scenes [12, 31, 48]. Consequently, 3DSGG can be applied to robot planning [1, 30, 67], autonomous driving [36, 46] and human-computer interactive query-answer [5, 41, 55].

Existing 3DSGG models are mainly working in two directions to improve the accuracy. 1) ***Visual contextual based methods*** successively pass 3D scene graph (3DSG) features through a given network, such as graph neural networks [47, 59] and transformers [33]. They collect spatial and semantic clues of adjacent objects by message passing mechanism for the purpose of enhancing the feature representations of objects and predicates. 2) ***Prior knowledge based methods*** extract class embeddings [6, 48, 61] and probability co-occurrence [11, 12] to refine predicate features, with the aim of enhancing recall for the infrequent predicate classes.

Despite notable advancements in 3DSGG, existing state-of-the-art (SOTA) methods still encounter two obstacles that constrain their practicality in the open-vocabulary (OV) settings. 1) The training of 3DSGG models necessitates extensive ground-truth object and predicate labeling. The annotation of 3DSG is costly and time-consuming to perform manually [45, 49, 62]. 2) All currently available 3DSGG methods are trained on a close-set of classes and struggle to extend their capabilities to novel classes in realistic scenarios [11, 47, 48, 59]. As illustrated in Figure 1(a), previous 3DSGG methods trained on closed-set object and predicate classes with fully supervised present a deficiency in recognizing novel classes. Also as Figure 1(b) highlights, in real-world applications, employing prediction results of 3DSG in closed-set might make an incorrect human-computer interactive query-answer or a terrible vehicle collision [10, 21].

Recently, vision-language pre-trained (VLP) models like CLIP [42] have shown its remarkable performance in 2D OV vision understanding tasks [3, 38, 65]. Some works have extended VLP models to comprehend 3D point clouds by transferring the text and image features from VLP to 3D point clouds [58, 60, 66]. However, directly applying them to unsupervised OV 3DSGG is undesirable due to two rea-

[*] Changbo Wang and Gaoqi He are the corresponding authors.

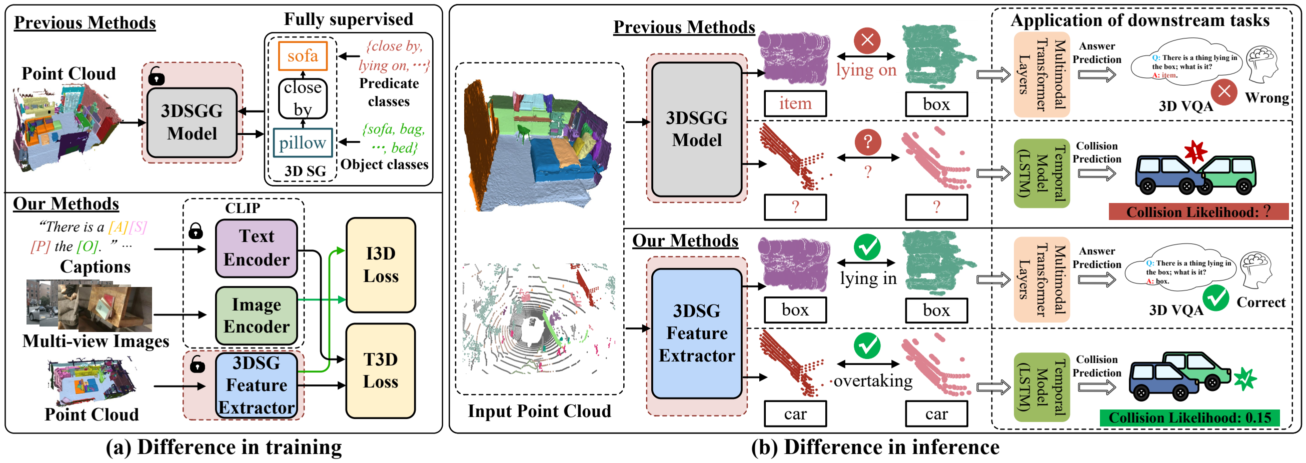

Figure 1. (a) **Difference in training:** Previous 3DSGG models trained on closed-set classes by fully supervised [12, 48, 61]. Our method trains a 3DSG feature extractor on naturally collected image-text pairs without using any 3DSG labels. (b) **Difference in inference:** Previous 3DSGG models fail to recognize novel class, limiting its practicality in real-world scenarios. Our model can recognize novel object and predicate classes.

sons. 1) The sentence-level text only describes the entire scene, which presents disparity with real word-level annotations of 3DSGG. 2) The fused multi-view image data impairs the understanding of spatial relationships. It motivates us to decompose sentences into word-level components and enhance the focus on camera views by using contrastive learning.

In this paper, a novel Cross-Modality Contrastive Learning 3DSGG (CCL-3DSGG) method is proposed for OV 3DSGG. Our goal is to develop a robust 3DSG feature extractor without any object and predicate annotations. To implement it, we require the fixed CLIP-based text and image features. Firstly, to align the text with 3DSG, we parse the captions of 3D scenes into words with different parts of speech. The text-3DSG contrastive loss is designed to align parsed words with 3DSG. To enhance robustness during the alignment, adjectives in parsed words are exchanged for different objects as negative samples. Then, the multi-view image-3DSG contrastive loss is used to align images with 3DSG. To enhance the ability of model to understand spatial features, the current camera view is considered as positives and those from other views as negatives. Finally, novel class recognition is performed during testing by calculating the cosine similarity between designed prompts for each class and 3DSG features. The primary contributions are summarized as:

- We propose the new and practical tasks of OV 3DSGG. A versatile CCL-3DSGG framework is designed to recognize novel classes by leveraging the CLIP. CCL-3DSGG can train without ground-truth 3DSG, minimizing annotation efforts and boosting real-world applicability.
- To align text and 3DSG, a grammar parsing module is introduced to dissect text into words with different parts of speech and enhance negative samples by adjective exchange. To align image and 3DSG, the camera view is treated as a positive sample and other views as negatives, aiding in the differentiation of 3D spatial predicates.
- Two contrastive losses are used to facilitate the training of robust 3DSG feature extractor. Extensive experiments are designed and conducted on multiple datasets. CCL-3DSGG significantly outperforms VL-SAT by an average of 9.8% and 104.4% for supervised and unsupervised. It also sets new SOTA benchmarks for zero-shot and open-vocabulary settings.

2. Related Works

2.1. 3D Scene Graph Generation

3D scene graphs [2, 15, 23, 44] depict objects and their relations within a 3D context. Prior research has predominantly focused on two directions. 1) ***Visual contextual based methods:*** Wald *et al.* [47] and Zhang *et al.* [59] introduced 3DSGG methods utilizing 3D point clouds, employing message passing graphs. After that, various works [20, 27, 39, 51] have explored online instance-incremental 3DSSG based on Wald *et al.* 2) ***Prior knowledge based methods:*** The incorporation of prior knowledge, like class embeddings [31, 40, 61] and statistical priors [6, 11, 12, 17], enhances the accuracy of 3DSGG. Wang *et al.* [48] integrated visual-linguistic cues in 3DSG predictions, leveraging CLIP for auxiliary training. Concurrent works [4, 16, 25, 26] have harnessed 3DSG for robotics, yet they are constrained by large language models (LLM) and lack the capacity for scene understanding. Our paper is the first work that explores unsupervised 3DSGG techniques to predict novel classes.

2.2. Predicting Novel Classes in SGG

Predicting novel classes in SGG involves identifying categories not present in the training set. Novel class prediction within SGG originated from approaches such as few-shot [9, 28], zero-shot, and weakly supervised SGG [57]. Tao *et al.* [19] and Gao *et al.* [13] first explored OV tasks in 2D SGG and video SGG, respectively. Subsequently, Zhang *et al.* [62] and Yu *et al.* [56] explored a pre-trained visual semantic space and entangled cross-modal prompt for effective 2D OV SGG. Contrary to the aforementioned supervised methods, this paper pioneers the unsupervised prediction of novel classes in OV settings within 3D scene graphs.

2.3. Contrastive Learning by CLIP

CLIP [42] has introduced large-scale visual-language contrastive pre-training, which simultaneously pre-trains on extensive text and image datasets, achieving joint representation learning between images and text. To enhance performance in downstream tasks, two strategies are employed. 1) ***Prompt learning based methods*** adjust to downstream tasks using minimal learnable prompts, circumventing extensive model fine-tuning [22, 34, 63]. 2) ***Contrastive loss optimization based methods*** refine representations by augmenting similarity for positive samples and diminishing it for negative ones [7, 18, 35]. CLIP exhibits potent cross-modality learning, excelling in various scene understanding tasks like object detection [32, 49], semantic segmentation [3] and crowd counting [29]. In 3D point cloud understanding, Pointclip [60], Pointclip v2 [66] and CLIP^2 [58] harness CLIP for 3D object and text alignment. ULIP [53] pretrains a unified representation of image, text, and 3D point cloud. Based on ULIP, ULIP-2 [54] leverages large multimodal models to generate more detailed descriptive texts for rendered 2D images. We deploy CLIP innovatively for unsupervised 3DSGG by parsing text grammarly and discriminating multi-view images.

3. Methods

Our framework is depicted in Figure 2. Our approach begins with the extraction of cross-modality features from text $\mathcal{T}$, image $\mathcal{I}$, and 3D point clouds $\mathcal{P}$ (Section 3.1). Notably, we segment text into words, each associated with multiple parts of speech using Grammar Parse [43, 50]. Subsequently, our model augments negative samples by exchanging adjectives. In the final step, we employ cross-modality contrastive losses to align text with 3DSG features and images with 3DSG features independently (Section 3.2). During the inference stage, we calculate similarity scores between prompts and 3DSG features to facilitate open-vocabulary 3DSGG tasks (Section 3.3).

3.1. Cross-modality Features Extraction

3.1.1 Text Features

Grammar Parse: In contrastive learning with textual data $\mathcal{T}$, CLIP models predominantly utilize sentence-level embeddings [48, 58, 60]. However, these embeddings might introduce ambiguity due to the amalgamation of diverse information from distinct words with different parts of speech, such as objects, predicates, and adjectives. Consider the caption sentence, “There is a wooden rectangular door the same shape as the tiny brown rectangular cabinet.” Words such as “door” (object type) or “rectangular” (shape) may dominate, leading to the potential dilution of attributes like “wooden” (material). The presence of adjectives such as “tiny” can further complicate the representation.

To enhance the discriminative power of text features and ensure precise cross-modality feature alignment, we propose segmenting text based on grammatical analysis [43, 50]. Each segmented word is then aligned with 3DSG features, reducing potential ambiguities.

We semantically decompose text into five components based on part-of-speech and dependencies: Subject (S) represents the primary subject; Object (O) identifies the main object; Predicate (P) describes the relationship between subjects and objects; Adjective (A) details the appearance and form attributes. Adjective Subject (AS) and Adjective Object (AO) further differentiate adjectives related to the subject or object. Lastly, Other (OT) covers uncategorized words. The final parsed word set is denoted as $\mathcal{T}_{pw} = \{S, O, P, A, AS, AO, OT\}$.

Negative Text Feature Augmentation: After grammar parsing, the quality of negative samples assumes a critical role in the training process for 3DSG feature extractor. For $Pair_i(A_iS_i, P_i, A_iO_i)$ and $Pair_j(A_jS_j, P_j, A_jO_j)$ in the text, negative text features $\mathcal{T}_{pw}^{-}$ are generated via

$$\mathcal{T}_{pw}^{-} = \text{Swap}\left(Pair_i, Pair_j\right) = \{(A_jX_i, A_iX_j)\}, \quad (1)$$

where $X \in \{S, O\}, X_i \neq X_j$. Adjective exchange maintains sentence structure while altering specific semantics, enabling our model to capture structured representations of detailed semantics effectively. After generating these negative samples, we use $\mathcal{T}_\theta$ to denote the text feature extractor from CLIP and extract the text feature $\mathcal{F}_\mathcal{T} \in \mathbb{R}^{w \times 512}$. Here, w represents the total count of samples, including both positive and negative words.

3.1.2 Image Features

In point cloud data, each 3D scan is complemented by RGB sequences with associated camera poses, enabling us

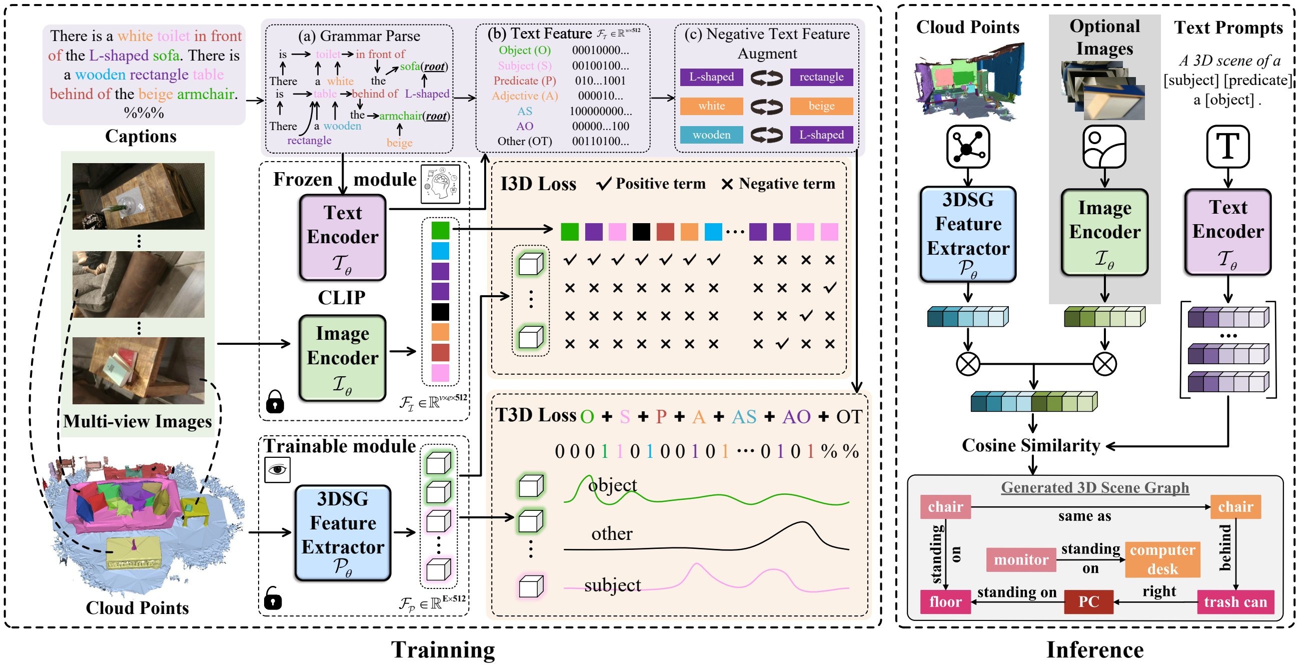

Figure 2. Architecture of the CCL-3DSGG. The CCL-3DSGG architecture begins with inputting image-text pairs and unlabeled 3D point clouds, aiming to train the 3DSG feature extractor $\mathcal{P}_\theta$. We employ part-of-speech analysis and enhance negative samples in the text. Subsequently, the text is processed by text encoder $\mathcal{T}_\theta$ of CLIP to obtain text feature $\mathcal{F}_\mathcal{T}$. The T3D loss aligns text with 3DSG features. Multi-view images are fed into image encoder $\mathcal{I}_\theta$ of CLIP to extract image features $\mathcal{F}_\mathcal{I}$, and the I3D loss aligns images with 3DSG features. In the inference stage, well-aligned 3DSG features $\mathcal{F}_\mathcal{P}$ are facilitated to predict novel 3DSG classes in real-world scenarios.

to derive 2D image patches for each point cloud instance. Drawing from the VL-SAT method described in [48], we use a pretrained CLIP vision encoder $\mathcal{I}_\theta$ to produce features for multi-view images. Importantly, we preserve features across various views for multi-view image-3DSG contrastive loss, represented as $\mathcal{F}_\mathcal{I} \in \mathbb{R}^{v\times e\times 512}$, where v and e denote the number of views and instances, respectively.

3.1.3 3DSG Features

Given the point set $\mathcal{P}$ of a scene s and the class-agnostic instance segmentation $\mathcal{M}$, the task of 3DSGG first parses the input $\mathcal{P}$ into an unclassified scene graph $\mathcal{G}_s = \{\mathcal{V}_s = (\mathcal{V}_o \cup \mathcal{V}_p), \mathcal{E}_s\}$, where $\mathcal{V}_o = \{\mathbf{v}_o^i\}_{i=1}^N$, $\mathbf{v}_o^i$ denotes the feature of object node and $\mathcal{V}_p = \{\mathbf{v}_p^i\}_{i=1}^M$, $\mathbf{v}_p^i$ denotes the feature of predicate node. Here N is the number of objects and M represents the number of predicates between objects. There are two types of undirected edges in $\mathcal{E}_s$, where each edge connects a predicate node to its corresponding subject node or object node. To establish such a graph, the pre-trained PointNet $\mathcal{P}_\theta$ [48] is adopted as our backbone to extract a set of objects feature vectors $\mathcal{V}_o$. Thereafter, the visual features $\mathcal{V}_p$ for predicates are generated by concatenating the object features with the center coordinates of any two object point sets [52]. We simply the 3DSG feature notation to $\mathcal{F}_\mathcal{P} \in \mathbb{R}^{E\times 512}$, where $E = N + M$.

3.2. Cross-Modality Contrastive Losses

So far, cross-modality features are obtained, that is text features $f_t \in \mathcal{F}_\mathcal{T}$, image features $f_I \in \mathcal{F}_\mathcal{I}$, and 3DSG features $f_o, f_p \in \mathcal{F}_\mathcal{P}$. f_o are the object features and f_p are the predicate features. The purpose of cross-modality contrastive losses is to align image and text to 3DSG, which consists of Multi-view Image-3DSG Contrastive (I3D) Loss and Text-3DSG Contrastive (T3D) Loss.

3.2.1 Multi-view Image-3DSG Contrastive Loss

It is noteworthy that CLIP has associated image features with semantic label embeddings. More specifically, these semantic label embeddings can be propagated to 3DSG features by designed contrastive loss. Meanwhile, image features can enhance the spatial discrimination of the learned 3DSG features when the camera view is treated as a positive sample, with other views considered negative. The I3D loss, denoted as $\mathcal{L}_{I3D}$, is formulated as follows:

$$\mathcal{L}_{I3D} = \sum_{i=1}^{e} -\log\left(\frac{\exp\left((f_o^i)^T f_I^{ij}\right)/\tau}{\sum_{k=1}^{v}\exp\left(((f_o^i)^T f_I^{ik})/\tau\right)}\right), \tag{2}$$

where $\tau = 0.1$ is the temperature coefficient and f_I^{ij} is the camera view image feature corresponding to the 3DSG feature f_o^i.

3.2.2 Text-3DSG Contrastive Loss

The goal of T3D loss is to align the word features of different parts of speech with 3DSG features through contrastive learning [14]. Firstly, for text and object alignment, we decompose $\mathcal{T}_{pw}$ into two subsets $\mathcal{T}_o^+$ and $\mathcal{T}_o^-$, where

$$\mathcal{T}_o^+ = \{S, O\}, \mathcal{T}_o^- = \{P, A, AS, AO, OT\}. \tag{3}$$

For the object feature $f_o^i \in \mathcal{F}_\mathcal{P}$ of the i_{th} object, we first assign the text feature $f_t^i \in \mathcal{T}_o^+$ with the highest similarity by calculating the cosine similarity $\phi\left(f_o^i, f_t^i\right) = \frac{(f_o^i)^T f_t^i}{\|f_o^i\| \cdot \|f_t^i\|}$. Then, we align the object words in the text with 3DSG feature using the following formula:

$$\mathcal{L}_{T3D\_o} = \sum_{i=1}^{N} -\log\left(\frac{\exp\left(\mathbf{w} * \phi\left(f_o^i, f_t^i\right)/\tau\right)}{\sum_{f_t^j \in \mathcal{T}_{pw} \cup \mathcal{T}_{pw}^-}\exp\left(\mathbf{w} * \phi\left(f_o^i, f_t^j\right)/\tau\right)}\right), \tag{4}$$

where $\mathbf{w}$ is the weight of each term. $\mathbf{w}$ for $\mathcal{T}_o^+$ is set to 1, while for $\mathcal{T}_o^-$ and $\mathcal{T}_{pw}^-$ is set to 2. Similarly, for text and predicate alignment, we decompose $\mathcal{T}_{pw}$ into two subsets $\mathcal{T}_p^+$ and $\mathcal{T}_p^-$ by another way, where

$$\mathcal{T}_p^+ = \{P, A, AS, AO\}, \mathcal{T}_p^- = \{S, O, OT\}. \tag{5}$$

Then, we align the predicate words in the text with 3DSG features using the following formula:

$$\mathcal{L}_{T3D\_p} = \sum_{i=1}^{M} -\log\left(\frac{\exp\left(\mathbf{w} * \phi\left(f_p^i, f_t^i\right)/\tau\right)}{\sum_{f_t^j \in \mathcal{T}_{pw} \cup \mathcal{T}_{pw}^-}\exp\left(\mathbf{w} * \phi\left(f_p^i, f_t^j\right)/\tau\right)}\right), \tag{6}$$

where $f_p^i \in \mathcal{T}_p^+$. $\mathbf{w}$ for $\mathcal{T}_p^+$ is set to 1, while for $\mathcal{T}_p^-$ and $\mathcal{T}_{pw}^-$ is set to 2. The final T3D loss is the mean of the two: $\mathcal{L}_{T3D} = \left(\mathcal{L}_{T3D\_o} + \mathcal{L}_{T3D\_p}\right)/2$. Finally, our complete loss function is formed by the following combination of loss functions:

$$\mathcal{L}_{\text{CCL}-\text{3DSGG}} = \lambda_1 \mathcal{L}_{I3D} + \lambda_2 \mathcal{L}_{T3D}, \tag{7}$$

where the hyper-parameters λ_1 and λ_2 are both set to 0.5.

3.3. Open-Vocabulary/Zero-Shot Inference

In Sections 3.1 and 3.2, we have trained a 3DSG feature extractor $\mathcal{P}_\theta$. In the inference stage, $\mathcal{P}_\theta$ is used to derive object features $\mathcal{F}_o^{inf}$ and predicate features $\mathcal{F}_p^{inf}$. Concurrently, $\mathcal{I}_\theta$ is employed to extract the image features $\mathcal{F}_I^{inf}$, which are optional in the inference stage. If images are available in the test dataset, we combine these with the 3DSG features, resulting in $\mathcal{F}_{oI}^{inf}$ and $\mathcal{F}_{pI}^{inf}$.

During inference, we input the prompt "a point cloud of a {object class}" into $\mathcal{T}_\theta$ to obtain features $\mathcal{F}_T^{inf}$ for all object classes. The final object label for each 3D instance is then determined by the maximum cosine similarity:

$$\text{argmax}\left\{\phi\left(\mathcal{F}_o^{inf}, \mathcal{F}_T^{inf}\right)\right\} \textbf{ or } \text{argmax}\left\{\phi\left(\mathcal{F}_{oI}^{inf}, \mathcal{F}_T^{inf}\right)\right\}. \tag{8}$$

Similarly, the predicate classes are predicted to be the cosine distance closest to the prompt "a point cloud of a {subject class} {predicate class} a {object class}":

$$\text{argmax}\left\{\phi\left(\mathcal{F}_p^{inf}, \mathcal{F}_T^{inf}\right)\right\} \textbf{ or } \text{argmax}\left\{\phi\left(\mathcal{F}_{pI}^{inf}, \mathcal{F}_T^{inf}\right)\right\}. \tag{9}$$

4. Experiments

In this section, we evaluate the performance of CCL-3DSGG on two datasets: 3DSSG [47] and ScanNet [8]. We provide a detailed account of the task description and experimental settings, compare our model to SOTA methods, and conduct ablation studies to emphasize the efficacy of CCL-3DSGG.

4.1. Task Description

The training set of 3DSSG [47] contains 3582 scenes, while the testing set comprises 548 scenes. The dataset includes 160 object classes and 27 predicate classes. We adopt two standard tasks from [61] for evaluation: (1) Predicate Classification (PREDCLS) which, given ground-truth object labels and bounding boxes, predicts the predicate labels of object pairs; and (2) Scene Graph Classification (SGCLS) which classifies the ground-truth bounding boxes and predicts predicate labels. To demonstrate the unsupervised and OV capability of our approach, we performed visualization experiments on the unlabeled indoor ScanNet dataset [8].

4.2. Implementation Details

To ensure a fair comparison, we employ the same pretrained PointNet with a GNN-based structure as our backbone of 3DSG feature extractor [48]. We set the dimension of cross-modality feature to 512. Training is conducted using the Adam optimizer [24], with a batch size of 8, over 100 epochs. The initial learning rate for the backbone is 0.001. We conduct all experiments on an Nvidia RTX 2080Ti GPU and implement our methodology using PyTorch [37]. The training of our full method takes approximately 48-50 hours. We follow the sub-scene split presented in [47]. Moreover, we reproduce the VL-SAT [48], KISGP [61], and SGPN [47] for comparison in this study. The captions utilized are sourced from the works of Yan *et al.* [55] and ScanRefer [5].

Table 1. Comparisons with state-of-the-arts on the 3DSSG dataset. Because the 3DSGG task inputs the instance segmentation, we only compute the mean of the two tasks of SGCLS and PREDCLS. The best and second best results are marked according to formats.

Method	SGCLS		PREDCLS			Testing Time
	R@{20/50/100}	mR@{20/50/100}	R@{20/50/100}	mR@{20/50/100}	Mean{R/mR}	(Second/Scene)
SGPN [47] (CVPR2020)	27.0/28.8/29.0	19.5/22.6/23.1	51.9/58.0/58.5	32.1/38.4/38.9	42.2/29.1	9
SGFN [51] (CVPR2021)	27.5/29.2/29.2	24.2/28.1/28.2	52.6/58.9/59.4	45.3/53.1/53.2	42.8/38.7	11
EdgeGCN [59] (CVPR2021)	28.0/29.8/29.8	24.5/29.1/29.2	54.7/60.9/61.5	54.3/62.1/62.2	44.1/43.6	15
KISGP [61] (NIPS2022)	28.5/30.0/30.1	24.4/28.6/28.8	59.3/65.0/65.3	56.6/63.5/63.8	46.4/44.3	12
Liu *et al.* [31] (TVCG2022)	-/-/-	-/-/-	-/-/-	-/53.2/61.5	-/-	14
Chen *et al.* [6] (CGF2022)	33.5/36.0/36.2	25.4/29.7/29.8	60.1/66.2/72.8	57.7/64.0/64.3	50.8/45.2	19
Feng *et al.* [12] (CVPR2023)	-/31.5/31.6	-/30.3/30.5	-/68.3/69.5	-/66.5/66.9	-/-	11
VL-SAT [48] (CVPR2023)	32.0/33.5/33.7	31.0/32.6/32.7	67.8/79.9/80.8	57.8/64.2/64.3	54.4/47.1	24
Our method	37.6/40.3/45.7	35.0/37.3/40.6	73.6/80.5/82.9	59.1/66.7/69.1	60.1/51.3	30

Table 2. Following the VL-SAT, the 26 predicate classes in the 3DSSG dataset are categorized into head, body, and tail parts based on the predicate distribution. The mean top-k accuracy (mA@k) metric is calculated for each part. Meanwhile, both unseen and seen triplets from the validation set are used to evaluate the robustness of our trained 3DSG feature extractor.

Method	Predicate						Triplet			
	Head		Body		Tail		Unseen		Seen	
	mA@3	mA@5	mA@3	mA@5	mA@3	mA@5	A@50	A@100	A@50	A@100
SGPN [47]	96.66	99.17	66.19	85.73	10.18	28.41	15.78	29.60	66.60	77.03
SGFN [51]	95.08	99.38	70.02	87.81	38.67	58.21	22.59	35.68	71.44	80.11
non-VL-SAT[48]	95.32	99.01	71.88	88.64	40.01	58.33	21.99	35.44	71.52	80.34
VL-SAT [48]	96.31	99.21	80.03	93.64	52.38	66.13	31.28	47.26	75.09	82.25
Our method	98.54	99.78	84.72	96.03	61.24	75.91	36.72	52.47	80.58	88.92

4.3. Comparisons with SOTA Methods on Close-Set

Quantitative Results with Supervised: Our method can be extended to supervised by modifying Eqn. (7) to:

$$\mathcal{L}_{\mathrm{CCL-3DSGG}} = \lambda_1 \mathcal{L}_{I3D} + \lambda_2 \mathcal{L}_{T3D} + \lambda_{obj}\mathcal{L}^{3d}_{obj} + \lambda_{pred}\mathcal{L}^{3d}_{pred}. \tag{10}$$

This modification entails the addition of object and predicate classification losses. Notably, the terms $\lambda_{obj}\mathcal{L}^{3d}_{obj}$ and $\lambda_{pred}\mathcal{L}^{3d}_{pred}$ are consistent with those in VL-SAT [48].

Table 1 presents a comparison of our approach with other 3DSGG methods based on Recall (R) and Mean Recall (mR) metrics. Our method surpasses the competing models in both tasks, registering an average Recall of 60.1 and a Mean Recall of 51.3. Despite introducing additional information, our model achieves a significant performance boost without a substantial increase in time (24 to 30).

Head-tail and Unseen Triple with Supervised: As evidenced in Table 2, our approach achieves SOTA performance when benchmarked against SGFN and VL-SAT for the infrequent predicate classes and unseen triplets. These results substantiate that our model furnishes more robust 3DSG feature representations, enhancing its generalization capability for unseen triplets.

Table 3. Unsupervised experimental results of mR on the 3DSSG dataset. w/o CL means without classification losses.

Method	SGCLS	PREDCLS
	mR@{20/50/100}	mR@{20/50/100}
SGPN [47] w/o CL	-/-/-	0.7/2.5/11.8
KISGP [61] w/o CL	2.5/5.4/8.9	5.6/9.2/23.5
VL-SAT [48] w/o CL	8.2/10.1/13.3	9.5/13.9/27.4
Our method	13.4/19.6/23.7	29.4/33.2/49.1

Quantitative Results with Unsupervised: Given that existing methods are designed for supervised, we modified them by removing object classification loss and predicate classification loss. Specifically, for SGPN, only the object loss was retained. Table 3 shows that SGPN is terrible for unsupervised tasks, largely because it is totally designed for supervised and lacks auxiliary loss functions. In contrast, KISGP and VL-SAT show higher mR results with SGPN, mainly due to the utilization of object and predicate labels as auxiliary training resources. Our method achieves SOTA

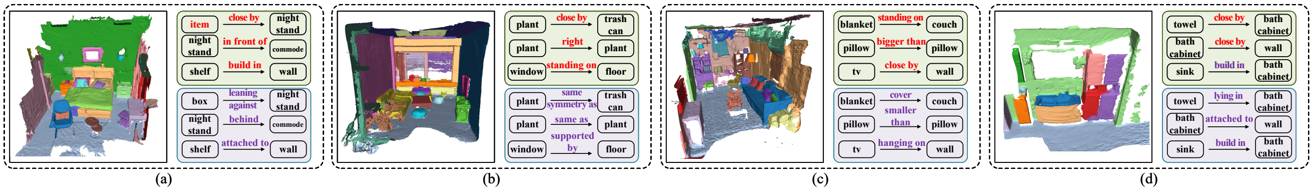

Figure 3. Qualitative examples of the improvement in supervised 3DSGG. On the right side of each scene, the result of the VL-SAT [48] is at the top, and our result is at the bottom. The purple predicates are those correctly classified relationships (in ground truth), and the red predicates are those incorrectly classified relationships. For better viewing, we only show failure cases.

Table 4. Open-vocabulary 3DSGG and Zero-shot 3DSGG of R@{50/100} on the 3DSSG dataset.

Method	Open-vocabulary 3DSGG		Zero-shot 3DSGG	
	SGCLS	PREDCLS	SGCLS	PREDCLS
KISGP [61]	19.3/24.8	38.1/44.6	15.7/20.1	32.0/38.9
Chen *et al.* [6]	20.5/25.8	46.2/52.8	15.8/18.7	41.2/47.6
VL-SAT [48]	23.1/29.4	60.3/66.9	21.6/28.1	43.5/59.4
Our method	37.1/42.3	64.8/71.2	35.5/40.6	49.1/65.7

Table 5. Ablation studies on CCL-3DSGG with unsupervised.

Exp	Module	PREDCLS			Object	
		mR@20	mR@50	mR@100	A@1	A@5
1	**our full method**	**29.4**	**33.2**	**49.1**	**49.2**	**73.1**
2	w/o Grammar Parse	6.8	10.7	26.6	28.0	52.5
3	w/o adjective exchange	25.8	29.9	45.6	46.5	70.0
4	w/o I3D loss	13.8	17.4	32.1	37.0	61.5
5	w/o T3D loss	5.3	10.1	18.6	20.5	41.0
6	w/o image information in testing	24.3	27.9	43.1	47.5	71.8
7	Learnable prompt [64] in testing	25.3	28.9	44.6	47.2	71.2
8	only add object label	31.4	35.7	53.5	60.4	88.6
9	only add predicate label	50.5	53.5	57.4	55.0	77.7
10	$\mathcal{P}_\theta$ + prediction head in VL-SAT	28.6	31.1	48.3	47.6	71.1
11	$\mathcal{P}_\theta$ + prediction head with fine-tune	40.7	48.1	53.7	52.6	74.8

with an average mR SGCLS of 18.9 and PREDCLS of 37.2 by carefully designed unsupervised contrastive losses.

Qualitative Results with Supervised: Figure 3 depicts four challenging scenes from diverse indoor rooms, including bedrooms, living rooms, and toilets. Specifically, Figure 3(a) illustrates proficiency of our model in distinguishing easily confounded objects and predicates, such as **box** and **leaning against**. The application of the I3D loss across multiple views enhances the capability of model to differentiate spatial predicates like **in front of** and **behind**. Notably, our model addresses the language biases observed in VL-SAT, converting expressions **build in** to more accurate predicate **attached to**. In Figure 3(b) and Figure 3(c), our model parses the textual components based on their parts of speech, emphasizing attributes between objects, such as **same symmetry as**, **same as**, **bigger than**, and **smaller than**. Furthermore, our approach excels in identifying infrequent predicates such as **cover** and **lying in**.

Qualitative Results with Unsupervised: Figure 4 presents the visualization results on ScanNet in the absence of ground truth. Our method is capable of predicting precise spatial predicates like **above** and **beside**, whereas VL-SAT typically predicts the more generic **close by**. Furthermore, our approach can predict predicates not present in the training set, such as **mounted to** and **hanging in**, underscoring the strength of our model in unsupervised learning and predicting novel classes.

4.4. Predicting Novel Classes

Building upon the methodologies in [19, 62], we train our proposed CCL-3DSGG using 70% of the object and predicate classes from the 3DSSG dataset, designated as base classes. We aim for CCL-3DSGG to effectively recognize the remaining 30% of novel objects and predicates during inference. For evaluation, we employ metrics across two object category sets: a combination of 70% base and 30% novel classes (Open-vocabulary 3DSGG), and solely the 30% novel classes (zero-shot 3DSGG).

Open-vocabulary 3DSGG: In Table 4, we replicated three methods in the supervised setting, as their unsupervised performance proved to be subpar, leading to further deterioration in the OV task. Our approach demonstrates SOTA even in unsupervised, achieving an average Recall of 53.85 when compared to supervised methods on OV settings. The results illustrated in Figure 5 show the performance of individual predicates in terms of mA@1 scores. Our evaluations reveal that, contrary to the challenges faced in VL-SAT when predicting novel classes, our method excels not only in proficiently predicting these novel classes but also consistently performs well with base classes.

Zero-shot 3DSGG: In Table 4, our approach significantly enhances the classification outcomes, surpassing VL-SAT by an average of 25.1% in the zero-shot setting. These findings underscore the efficacy of our pretraining strategy, leveraging naturally occurring free-form captions and images. This approach gleans substantial knowledge from realistic open-world scenarios, culminating in transferable 3DSG representations that outperform supervised methods on the more constrained 3DSSG dataset.

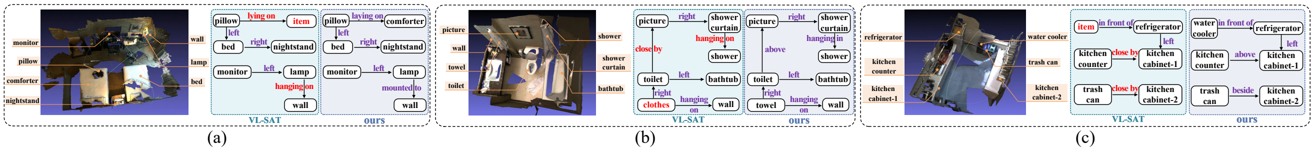

Figure 4. Qualitative results are drawn from both VL-SAT [48] and our method utilizing the ScanNet dataset [8]. Note that ScanNet does not provide annotations specific to 3DSG. As such, we rely on captions from [5] and undertake a manual evaluation process.

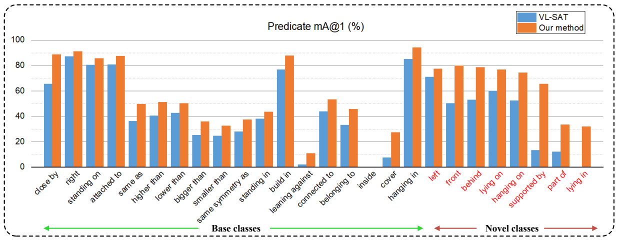

Figure 5. In the open-vocabulary scenario, our holistic model exhibits enhanced performance in mA@1 relative to VL-SAT [48]. Classes highlighted in red signify novel classes.

4.5. Ablation Study

In this section, we show the ablation performance on the 3DSSG dataset in Table 5.

Effectiveness of Cross-Modality Features: EXP 1 denotes our full method. For EXP 2, we replaced the Grammar Parse in Section 3.1.1 with sentence-level features to facilitate similarity computations and contrastive losses training. This modification resulted in a significant performance decline, particularly in managing word-level 3DSGG tasks (29.4 to 6.8 on mR@20). During EXP 3, we removed the negative sample enhancement using adjective exchange, observing a decline in performance (29.4 to 25.8 on mR@20). In EXP 5, after eliminating the T3D loss, the model exhibited difficulties in unsupervised tasks, faring even worse than in EXP 2. EXP 7 replaced our designated prompt with a learnable one as proposed by [64]. However, this did not enhance performance, potentially due to the single test data domain. In EXP 4, when we eliminated the I3D loss, we observed a performance dip. This suggests that images play a supplementary role, with text features being the primary determinant of unsupervised performance. In EXP 6, we further confirmed the optional of image features by removing them during inference.

Influence of Object and Predicate Annotation: In both EXP 8 and EXP 9, we incorporated the object loss $\mathcal{L}^{3d}_{obj}$ and predicate loss $\mathcal{L}^{3d}_{pred}$, respectively. The inclusion of the object loss yielded a marked increase in object classification accuracy (49.2 to 60.4 on A@1), while the predicate loss led to a significant enhancement in predicate classification accuracy (29.4 to 50.5 on mR@20).

Has Our Model Effectively Learned Robust 3DSG Features? In EXP 10, we employed the prediction head from VL-SAT to infer features without prompts during the testing phase. The performance aligns closely with our unsupervised approach with slightly reduced (29.4 vs 28.6 on mR@20). In EXP 11, fine-tuning the prediction head in VL-SAT with a limited dataset enhanced the performance, making them comparable to those achieved with supervised methods. All the results prove the robustness of the 3DSG features learned by our model.

5. Conclusion

In this paper, we introduce a CCL-3DSGG method for 3DSGG without any object and predicate annotations. To achieve efficient open-vocabulary 3DSGG, we decompose captions into words with different parts of speech. Text is aligned with 3DSG features utilizing the T3D loss, enriched with adjective exchange based negative sample augmentation. Concurrently, images are aligned with 3DSG features via the I3D loss, furnishing 2D contextual clues from diverse views. Our CCL-3DSGG framework exhibits robust open-vocabulary and zero-shot capabilities across multiple datasets.

Limitations: There are several limitations of our work and still much to do to realize the full potential of the proposed approach. First, the inference algorithm could probably take better advantage of pixel features when images are present at test time using earlier fusion. Second, we evaluated extensively on open-vocabulary 3D SGG, but provide only qualitative results for other datasets since ground truth are missing. In future work, it will be interesting to design experiments to quantify the success of open vocabulary queries for 3DSGG where ground truth is not available.

Acknowledgments This research is sponsored by Natural Science Foundation of Chongqing, China (CSTB2022NSCQ-MSX0552), National Natural Science Foundation of China (No. 62002121, 62072183, 62102151), Shanghai Science and Technology Commission (No. 21511100700, 22511104600), Shanghai Sailing Program (21YF1411200), CCF-Tencent Rhino-Bird Open Research Fund, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education (KLATASDS2305).

References

[1] Christopher Agia, Krishna Murthy Jatavallabhula, Mohamed Khodeir, Ondrej Miksik, Vibhav Vineet, Mustafa Mukadam, Liam Paull, and Florian Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs. In *Conference on Robot Learning*, pages 46–58. PMLR, 2022. 1

[2] 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. 2

[3] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 1, 3

[4] Haonan Chang, Kowndinya Boyalakuntla, Shiyang Lu, Siwei Cai, Eric Pu Jing, Shreesh Keskar, Shijie Geng, Adeeb Abbas, Lifeng Zhou, Kostas Bekris, et al. Context-aware entity grounding with open-vocabulary 3d scene graphs. In *International Conference on Robot Learning*, 2023. 2

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 1, 5, 8

[6] Lianggangxu Chen, Jiale Lu, Yiqing Cai, Changbo Wang, and Gaoqi He. Exploring contextual relationships in 3d cloud points by semantic knowledge mining. In *Computer Graphics Forum*, pages 75–86. Wiley Online Library, 2022. 1, 2, 6, 7

[7] Yihao Chen, Xianbiao Qi, Jianan Wang, and Lei Zhang. Disco-clip: A distributed contrastive loss for memory efficient clip training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22648–22657, 2023. 3

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5, 8

[9] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael Bernstein, and Fei-Fei Li. Visual relationships as functions: Enabling few-shot scene graph prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 3

[10] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3722–3731, 2021. 1

[11] Mingtao Feng, Haoran Hou, Liang Zhang, Yulan Guo, Hongshan Yu, Yaonan Wang, and Ajmal Mian. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 2023. 1, 2

[12] Mingtao Feng, Haoran Hou, Liang Zhang, Zijie Wu, Yulan Guo, and Ajmal Mian. 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9182–9191, 2023. 1, 2, 6

[13] Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*, 2023. 3

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 5

[15] Paul Gay, James Stuart, and Alessio Del Bue. Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 330–346. Springer, 2019. 2

[16] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023. 2

[17] Chaolin Han, Hongwei Li, Jian Xu, Bing Dong, Yalin Wang, Xiaowen Zhou, and Shan Zhao. Unbiased 3d semantic scene graph prediction in point cloud using deep learning. *Applied Sciences*, 13(9):5657, 2023. 2

[18] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878, 2023. 3

[19] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *European Conference on Computer Vision*, pages 56–73. Springer, 2022. 3, 7

[20] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022. 2

[21] Ziyuan Jiao, Yida Niu, Zeyu Zhang, Song-Chun Zhu, Yixin Zhu, and Hangxin Liu. Sequential manipulation planning on scene graph. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8203–8210. IEEE, 2022. 1

[22] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Domain prompt tuning via meta relabeling for unsupervised adversarial adaptation. *IEEE Transactions on Multimedia*, 2023. 3

[23] Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics*, 50(12):4921–4933, 2019. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for

stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Sebastian Koch, Pedro Hermosilla, Narunas Vaskevicius, Mirco Colosi, and Timo Ropinski. Lang3dsg: Language-based contrastive pre-training for 3d scene graph prediction. *arXiv preprint arXiv:2310.16494*, 2023. 2

[26] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. *arXiv preprint arXiv:2402.12259*, 2024. 2

[27] Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. Embodied semantic scene graph generation. In *Conference on Robot Learning*, pages 1585–1594. PMLR, 2022. 2

[28] Xingchen Li, Long Chen, Guikun Chen, Yinfu Feng, Yi Yang, and Jun Xiao. Decomposed prototype learning for few-shot scene graph generation. *arXiv preprint arXiv:2303.10863*, 2023. 3

[29] Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. Crowdclip: Unsupervised crowd counting via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2893–2903, 2023. 3

[30] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10968–10980, 2023. 1

[31] Yuanyuan Liu, Chengjiang Long, Zhaoxuan Zhang, Bokai Liu, Qiang Zhang, Baocai Yin, and Xin Yang. Explore contextual information for 3d scene graph generation. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1, 2, 6

[32] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. Capdet: Unifying dense captioning and open-world detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15233–15243, 2023. 3

[33] Changsheng Lv, Mengshi Qi, Xia Li, Zhengyuan Yang, and Huadong Ma. Revisiting transformer for point cloud-based 3d scene graph generation. *arXiv preprint arXiv:2303.11048*, 2023. 1

[34] Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, and Changsheng Xu. Understanding and mitigating overfitting in prompt tuning for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[35] Anas Mahmoud, Jordan SK Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, and Steven L Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7102–7110, 2023. 3

[36] Arnav Vaibhav Malawade, Shih-Yuan Yu, Brandon Hsu, Deepan Muthirayan, Pramod P Khargonekar, and Mohammad Abdullah Al Faruque. Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction. *IEEE Internet of Things Journal*, 9(12):9379–9388, 2022. 1

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[38] Yash Patel, Yusheng Xie, Yi Zhu, Srikar Appalaraju, and R Manmatha. Simcon loss with multiple views for text supervised semantic segmentation. *arXiv preprint arXiv:2302.03432*, 2023. 1

[39] Chao Qi, Jianqin Yin, Jinghang Xu, and Pengxiang Ding. Instance-incremental scene graph generation from real-world point clouds via normalizing flows. *arXiv preprint arXiv:2302.10425*, 2023. 2

[40] Chao Qi, Jianqin Yin, Zhicheng Zhang, and Jin Tang. Dynamic scene graph generation of point clouds with structural representation learning. *Tsinghua Science and Technology*, 29(1):232–243, 2023. 2

[41] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023. 1

[42] 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, 3

[43] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 3

[44] Rajat Talak, Siyi Hu, Lisa Peng, and Luca Carlone. Neural trees for learning on graphs. *Advances in Neural Information Processing Systems*, 34:26395–26408, 2021. 2

[45] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 1

[46] Yafu Tian, Alexander Carballo, Ruifeng Li, and Kazuya Takeda. Road scene graph: A semantic graph-based scene representation dataset for intelligent vehicles. *arXiv preprint arXiv:2011.13588*, 2020. 1

[47] 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. 1, 2, 5, 6

[48] 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. 1, 2, 3, 4, 5, 6, 7, 8

[49] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detec-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023. 1, 3

[50] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 3

[51] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 2, 6

[52] 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. 4

[53] 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. 3

[54] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023. 3

[55] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. Comprehensive visual question answering on point clouds through compositional scene manipulation. *arXiv preprint arXiv:2112.11691*, 2021. 1, 5

[56] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. *arXiv preprint arXiv:2303.13233*, 2023. 3

[57] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3736–3745, 2020. 3

[58] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2023. 1, 3

[59] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9705–9715, 2021. 1, 2, 6

[60] 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, 3

[61] Shoulong Zhang, Aimin Hao, Hong Qin, et al. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 5, 6, 7

[62] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2915–2924, 2023. 1, 3, 7

[63] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 7, 8

[65] Peipei Zhu, Xiao Wang, Lin Zhu, Zhenglong Sun, Wei-Shi Zheng, Yaowei Wang, and Changwen Chen. Prompt-based learning for unpaired image captioning. *IEEE Transactions on Multimedia*, 2023. 1

[66] 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, 3

[67] Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, and Yuke Zhu. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6541–6548. IEEE, 2021. 1