Multi-view Invariance Learning for 3D Scene Graph Pre-training via Collaborative Cross-Modal Regularization

Yucheng Huang, Luping Ji*, Ruijie Xiao, Jiayuan Sun

School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
jiluping@uestc.edu.cn, {hyc, 202411081612,sunjiayuanro}@uestc.std.edu.cn

Abstract

3D scene graph generation is a pivotal task in scene understanding. Its performance is easy to be constrained by the limited availability of annotated data. Currently, the existing solutions on point cloud pre-training usually emphasize on object-centric representations while neglecting the predicate feature learning. This limitation significantly hinders their relational reasoning capabilities, as inter-object relationships are fundamentally governed by predicate features. To enhance 3D Scene Graph Pre-training, this paper proposes a task-specific Multi-view Invariance Learning framework with Collaborative Cross-modal Regularization. In detail, the inherent horizontal-rotation invariance of 3D objects and their semantic relationships are leveraged to construct a self-supervised paradigm for triplet feature learning. Moreover, our framework harnesses the cross-modal prior knowledge from the vision-language model to regularize model optimization. It could further achieve the semantic discrimination via unsupervised deep clustering. To resolve the knowledge discrepancies arising from the pre-trained model in fine-tuning, a predicate adapter equipped with knowledge filtering gate is devised to selectively aggregate the predicate features of pre-trained model. Extensive experiments demonstrate that our framework is effective in boosting 3D scene graph generation performance, surpassing state-of-the-art ones.

Code — https://github.com/UESTC-nnLab/MVIL-SG

Introduction

3D scene graph generation (SGG) aims to predict the semantics of objects and their corresponding predicate relationships from point clouds for scene understanding. To learn relational triplets, SGPN (Wald et al. 2020) pioneers the creation of a 3D scene graph dataset 3DSSG and proposes an end-to-end 3D scene graph network.

Current 3D scene graph primarily adopts three main schemes. The first is the prior knowledge injection using meta-embeddings (Zhang et al. 2021b) or CLIP-based cross-modal alignment (Chen et al. 2024; Wang et al. 2023) to enhance the recall for infrequent predicate classes. The second is the predicate modeling, like SGFN (Wu et al. 2023; Feng et al. 2025), enhancing predicate classification by designing

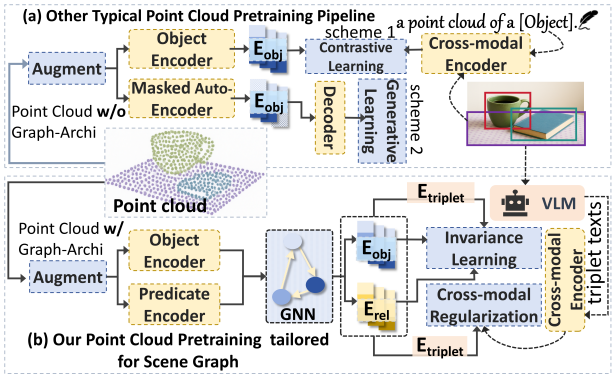


Figure 1: A comparative architecture analysis between typical point cloud pre-training schemes and our scene graph-tailored approach.

robust relationship descriptor. However, these two schemes rely heavily on labeled data, while the scarcity of annotated 3D scene graph data make them fail to learn underlying feature representations for better generalization.

Besides, the typical point cloud pre-training (Zhang et al. 2021c; Pang et al. 2023), offers a compelling solution for the scarcity of annotated datasets, as shown in Figure 1(a). Among them, (Koch et al. 2024b) pioneers a self-supervised pre-training via reconstruction for scene graph networks. Nevertheless, previous approaches lack explicit objectives for optimizing predicate feature learning, thus limiting their ability to model relationship between objects. Moreover, the reliable relational modeling is critical for triplet classification in scene graph generation.

Hence, this paper aims at proposing a pre-training scheme tailored for scene graph through the invariance learning of comprehensive triplet features, **including objects and predicates**, as shown in Figure 1(b). The scheme relies on reliable data augmentations for increasing the distributional divergence between two views while preserving their semantic consistency. Although rotation could alter the distribution of point cloud, it may change the semantics of the predicate. As shown in Figure 2, tests on VL-SAT demonstrate that the horizontal rotation about z-axis significantly degrades the performance for horizontal geometric predicates, while

*Corresponding author

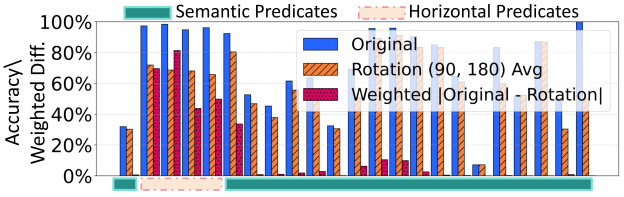


Figure 2: The impact of rotation on predicate-relation classification. The red bar is computed through multiplying the predicate accuracy difference (between the original and rotated models) by the frequency of each category.

having minimal impact on most semantic relationships.

As a compromise, this paper adopts augmentations, such as horizontal rotation, scale variations, elastic distortion, jitter, and patch-wise dropout, to generate multi-view positive samples for relational triplets. And a non-contrastive learning approach (Zbontar et al. 2021) is adopted to learn invariant features for triplet without negative sample requirement. To enhance the semantic discriminability of pretrained features, we employ a cross-modal regularization to inject the prior predicate knowledge from the vision-language model (Wang et al. 2024), and it enriches the semantic representation of triplet features. Upon feature learning stabilization, a unsupervised clustering is designed to improve their discriminability, benefiting the predicate classification in downstream scene graph generation.

While our pretraining method provides the guidance for learning semantic predicate relationships, it inevitably conflicts with horizontal geometric relationships (e.g., "left-/right/front/behind"). To address this discrepancy, we implement a adapter-based fine-tuning strategy. By designing a knowledge filtering gate, the prior knowledge is selectively extracted from the pre-trained predicate encoder and aggregates features from the fine-tuning adapters to learn effective horizontal geometric predicates.

The main contributions are summarized as follows: 1) a multi-view self-supervised pre-training scheme for 3D scene graphs with collaborative cross-modal regularization is first proposed to learn generalizable triplet feature representations; 2) an adapter with knowledge filtering gate is devised to learn horizontal geometric relationships during the fine-tuning stage; 3) comprehensive experiments and qualitative analyses are conducted to validate the superiority.

Related Works

Prior Knowledge Injection for Scene Graph

Prior knowledge is injected into downstream scene graph generation by utilizing a teacher model or embeddings pre-trained on proxy tasks. (Zhang et al. 2021b) leverages the learned meta-embeddings as the prior knowledge for SGG tasks. (Koch et al. 2024a; Chen et al. 2024) utilizes CLIP-based (Radford et al. 2021) cross-modal alignment for robustness triplet feature learning. (Koch et al. 2024c) achieves open-vocabulary 3D SGG by leveraging CLIP and BLIP (Li et al. 2022) to incorporate image-text prior knowledge.

Predicate Modeling for Scene Graph

Predicate relation modeling underpins the classification accuracy of predicates in scene graph generation. The pioneered SGPN (Wald et al. 2020) formulates the feature descriptor for predicates as the union of object pairs. Subsequent predicate descriptors like statistical metrics (Wu et al. 2023) and hyper-rectangle embeddings (Feng et al. 2025) effectively enhance the predicate feature learning. (Feng et al. 2023) refines predicate classification through spatial hierarchical modeling.

Pre-Training for 3D Scene Graph

The 3D understanding task to address the scarcity of annotated data relies on point cloud pretraining. Current point cloud pretraining schemes adopt contrastive learning (Xie et al. 2020; Zhang et al. 2021c) or mask autoencoder-based generative learning (Pang et al. 2023). While generative methods focus on local point cloud features (Qi et al. 2023), which is misaligned with predicate relation modeling (as they require holistic inter-object distribution modeling). Contrastive learning faces challenges in constructing predicate-negative samples from unlabeled data.

To address this, we propose a non-contrastive framework that learns multi-view invariant features for triplet through cross-modal alignment and deep clustering, and it could achieve discriminative capability of scene graph features without negatives.

Methods

The overall scene graph pre-training pipeline as shown in Figure 3. Multi-view invariance learning serves as the primary representation learning approach, assisted by cross-modal regularization and deep clustering to achieve effective semantic discrimination. Upon completion of pre-training, a knowledge-filtering gate-based adapter is employed to accomplish model fine-tuning.

Overview

The required pre-training data on ScanNet (Dai et al. 2017) includes: scene point clouds $\mathcal{P} \in \mathbb{R}^{N_s \times 3}$, class-agnostic instance segmentation masks $\mathcal{M} \in \mathbb{R}^{N_s}$, and scene image sequences $\mathcal{I}_{img} \in \mathbb{R}^{T}$. $\mathcal{P}$ and $\mathcal{M}$ are organized into a directed scene graph $\mathcal{G} = \{\mathcal{O}, \mathcal{R}\}$, where the vertex set $\mathcal{O}$ represents all instance objects, and the edge set $\mathcal{R}$ denotes potential predicate relationships between any two instance objects.

For the downstream scene graph prediction task, the organized graph data $\mathcal{G}$ is fed into the neural network $\mathscr{F}_\theta$ (including encoders and graph neural network) and classification heads to predict probability distributions for each vertex $\mathcal{O}_i$ and edge $\mathcal{R}_i$. To ensure optimal task adaptation, the whole network $\mathscr{F}_\theta$ during pre-training is employed to learn multi-view invariant features.

Multi-view Invariance Learning

Multi-view Augmentation. Given a point cloud $\mathcal{P}$ in the standard coordinate system, another view could be obtained through the transformation computed by

$$\hat{\mathcal{P}} = R_Y(\gamma) \cdot R_X(\beta) \cdot R_Z(\alpha) \cdot \mathcal{P} + t \tag{1}$$

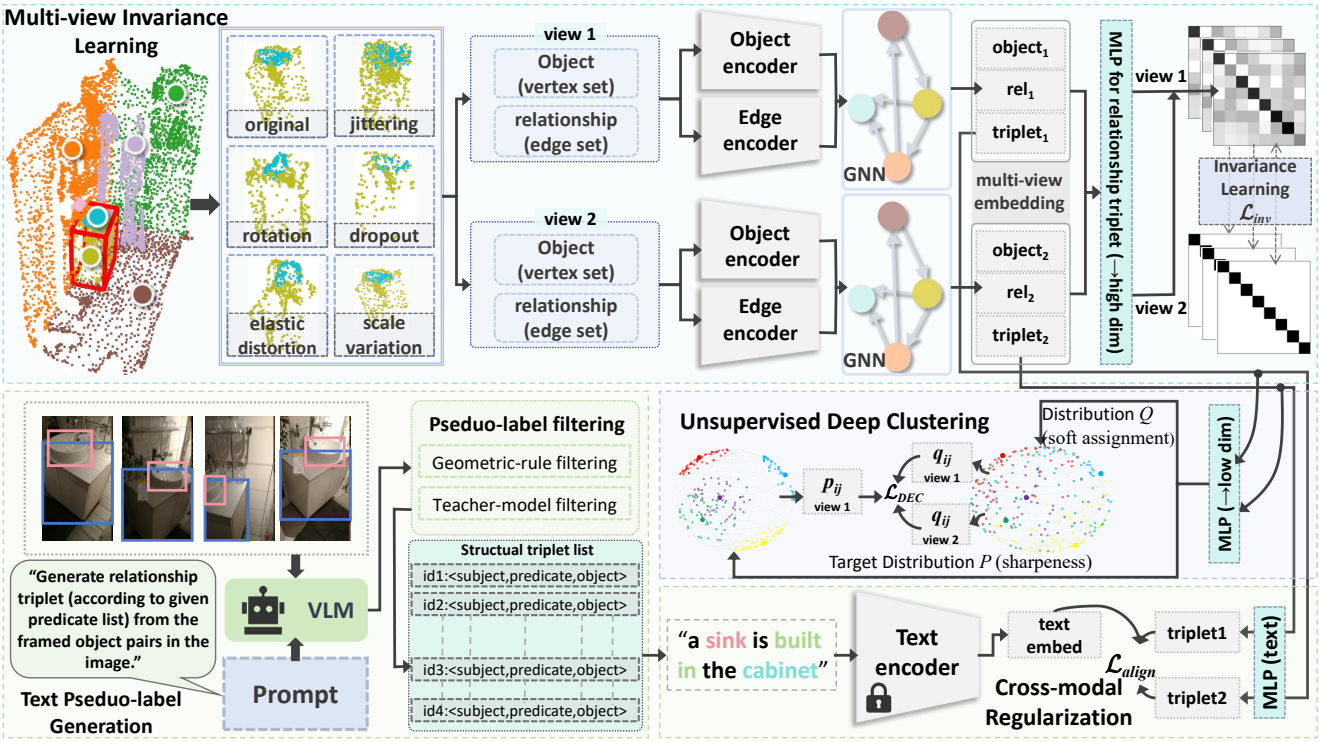


Figure 3: Overview of the proposed scene graph pre-training pipeline.

where R_X, R_Y, and R_Z denote rotation angles about the x-, y-, and z-axes, t is the translation. Notably, rotations about the y- and x-axes could corrupt the semantic information of objects and predicates. For instance, rotating ceiling point clouds about the y-axis could misclassify them as walls. In contrast, rotations about the z-axis (R_Z) preserve the semantic features of both objects and most predicate relationships.

Hence, this study incorporates augmentation strategies including: horizontal rotation about the z-axis, elastic distortion, patch-wise dropout, jittering, and scale variation.

Invariance Learning. Given a graph $\mathcal{G}_1 = \{\mathcal{O}_1, \mathcal{R}_1\}$ from view 1 and a graph $\mathcal{G}_2 = \{\mathcal{O}_2, \mathcal{R}_2\}$ from view 2, the pre-training objective is to achieve multi-view invariance learning of both objects and predicate relationships through network parameter optimization:

$$\begin{cases} \theta^* = \arg\min_{\theta} \mathbb{E}_{(\mathcal{G}_1,\mathcal{G}_2)\sim\mathcal{D}}[\mathcal{L}_{inv}(\theta)] \\ \mathcal{L}_{inv}(\theta) = dist(\mathscr{F}_\theta(\mathcal{G}_1;\theta), \mathscr{F}_\theta(\mathcal{G}_2;\theta)) \end{cases} \tag{2}$$

where $dist(\cdot)$ denotes an abstract distance metric.

To avoid negative sample construction, we chose the non-contrastive approach, Barlow Twins, as the distance metrics $dist(\cdot)$. Then, the learning objective shifts to achieving multi-view feature invariance learning and decorrelation for avoiding the feature representation collapse. Given the vertex features $f_{\mathcal{O}}$ and edge features $f_{\mathcal{R}}$ output by the scene graph network, the triplet features $f_{\mathcal{T}}$ are constructed by concatenating vertex and edge features through graph indexing, i.e., $f_{\mathcal{T}} = concat\langle f^i_{\mathcal{O}}, f^{ij}_{\mathcal{R}}, f^j_{\mathcal{O}}\rangle$. For notational simplicity, we use f_z to represent any of these three feature types, where $z \in \{\mathcal{O}, \mathcal{R}, \mathcal{T}\}$, and f_z is obtained by mapping $f_{\mathcal{O}}$, $f_{\mathcal{R}}$, and $f_{\mathcal{T}}$ to a higher-dimensional space ($dim \rightarrow 4096$) through three projection heads. The total invariance learning objective could be expressed as

$$\begin{cases} \mathcal{L}_{inv} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \underbrace{\lambda_b \sum_{i\neq j} \mathcal{C}_{ij}^2}_{\text{decorrelation term}} \\ \mathcal{C}_{ij} \triangleq \dfrac{\sum_b f^A_{z,b,i} f^B_{z,b,j}}{\sqrt{\sum_b (f^A_{z,b,i})^2}\sqrt{\sum_b (f^B_{z,b,j})^2}} \end{cases} \tag{3}$$

where b indexes batch samples and i, j index the vector dimension of the f_z, $\mathcal{C}$ is a square matrix with size the dimensionality of the f_z, λ_b is the weight of decorrelation term.

Since Barlow Twins lacks the samples interaction within a batch, it cannot explicitly achieve semantic discrimination through the objective function. To address this limitation, we propose two regularization schemes to enhance their semantic representation.

Cross-Modal Regularization

This study utilizes scene image sequences as an intermediary, harnessing the extensive cross-modal priors of the QWEN-VL-MAX (Wang et al. 2024) model to regularize the feature invariance learning.

Pesduo-Label Generation. Given a set of object instances $\mathcal{O} = \{o_1, o_2, \ldots, o_N\}$ with a IDs-shared instance masks in both point clouds and a sequence of images $\mathcal{I}_{img} =$

$\{I_1, I_2, \ldots, I_T\}$, the process begins by establishing two mappings.

Let $\Phi_I : \mathcal{O} \rightarrow 2^{\{1,\ldots,T\}}$ be mapping where $\Phi_I(o_i)$ return the set of image indices in which object o_i is visible. Let $\Phi_B : \{1, \ldots, T\} \rightarrow 2^{\mathbb{R}^4 \times \mathcal{O}}$ be a mapping where $\Phi_B(T)$ returns the set of bounding boxes for all objects present in image I_T. Bounding boxes with dimensions below a specified threshold ($height < 30px$ or $width < 30px$) are filtered out to discard instances lacking salient semantic information.

For each edge $(o_s, o_o) \in \mathcal{R}$ in the scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{R})$, the set of co-occurring images is determined by the intersection $\mathcal{I}^{co}_{s,o} = \Phi_I(o_s) \cap \Phi_I(o_o)$. A subset of images $\mathcal{I}^{sel}_{s,o} \subseteq \mathcal{I}^{co}_{s,o}$ is randomly sampled, where $|\mathcal{I}^{sel}_{s,o}| = min(4, |\mathcal{I}^{co}_{s,o}|)$.

For each selected edge $I_k \in \mathcal{I}^{sel}_{s,o}$, the corresponding bounding boxes for the object pair (o_s, o_o) are retrieved via $\Phi_B(k)$ and rendered onto the image. These annotated images, along with structured prompts, are input to the QWEN-VL model. The model is constrained to generate a set of the top-two most probable textual triplets, $\mathcal{T}_{s,o,k} = \{\langle subject, predicate_j, object\rangle\}^2_{j=1}$, where each predicate is drawn from the predefined predicate categories of the 3DSSG. The final set of pseduo-labels for the edge (o_s, o_o) is the union of outputs from all sampled images:

$$\xi_{s,o} = \bigcup_{I_k \in \mathcal{I}^{sel}_{s,o}} \mathcal{T}_{s,o,k} \tag{4}$$

After obtaining pseudo-labels, we first filter the duplicate labels among multi-view images, and further refine subject-object order using geometric rules. Then, leveraging a teacher model trained on 3DSSG, we compute each predicate's expected confidence $E_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{R}| \times cls}$ and uncertainty via Monte Carlo dropout. If uncertainty is below threshold $\mathcal{U} \in \mathbb{R}^{cls}$, we retain the pseudo-label only if it ranks among the top 5 in expected confidence $E_{\mathcal{S}}$, otherwise, we replace it with the top-2 prediction[1].

Cross-Modal Triplet Features Alignment. To guide cross-modal alignment, we employ frozen text encoder of CLIP (Radford et al. 2021) to generate the corresponding embeddings f_{text} of triplet pseudo-labels as supervisory signals. We adopt the L_1 distance as the optimization objective to align point cloud triplet features with their textual triplet counterparts, as the given text pseudo-labels are mapped to a deterministic feature space. Furthermore, through data augmentation, we obtain multi-view point cloud features of the same triplet that semantically correspond to identical textual descriptions. The final cross-modal regularization are computed as follows

$$\begin{aligned}\mathcal{L}_{align} = \mathbb{E}_{\mathcal{T}}[&||\mathscr{F}_t(f^{view_1}_{\mathcal{T}}) - \mathcal{C}_{lip}(\hat{\xi})||_1 \\ &+||\mathscr{F}_t(f^{view_2}_{\mathcal{T}}) - \mathcal{C}_{lip}(\hat{\xi})||_1]\end{aligned} \tag{5}$$

where $\mathscr{F}_t(\cdot)$ is is a multilayer perceptron (MLP) employed for dimension alignment. For each predicate sample potentially linked to multiple pseudo-labels, we randomly sample a single one for alignment calculation.

[1] cls is the total number of predicate categories. $\mathcal{U}$ is the 35th percentile of uncertainties across the entire dataset for each class.

Unsupervised Deep Clustering for Triplets

Cross-modal alignment provides preliminary clustering centroids. A unsupervised deep clustering is employed to further enforce class-aware feature separation.

During the mid-training phase when representations stabilize, the Deep Embedding Clustering (DEC) strategy (Xie, Girshick, and Farhadi 2016) is utilized to generate cluster centroids $\{u_1, u_2, \ldots, u_k\}$ for the triplet features across the complete dataset, where k denotes the number of clusters. To prioritize relational semantics, we set k equals to the number of predefined predicate categories. Formally, the soft assignment probability q_{ij} for the j-th cluster is computed as

$$q_{ij} = \frac{(1 + \|\mathscr{F}_c(f_{\mathcal{T},i}) - \mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1 + \|\mathscr{F}_c(f_{\mathcal{T},i}) - \mu_{j'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{6}$$

here, $\mathscr{F}_c(\cdot)$ is the MLP for dimensionality reduction ($dim \rightarrow 64$), α denotes the degrees of freedom (set to 1 in our experiments), where the probability q_{ij} is derived from Student's t-distribution.

Since DEC does not rely on ground-truth labels, it requires the construction of a target distribution P to guide the optimization of the soft assignment probabilities Q. The design principle of P is to amplify high-confidence predictions, computed as follows:

$$p_{ij} = \frac{q^2_{ij}/f_j}{\sum_{j'} q^2_{ij'}/f_{j'}}, \quad \text{where} \quad f_j = \sum_i q_{ij} \tag{7}$$

the squaring operation on q_{ij} accentuates the disparity between high- and low-confidence assignments.

During actual training, we observed that dynamically computing the target distribution P per batch leads to non-convergent loss. To address this, we instead compute P of "view 1" every 10 epochs using the entire dataset, with gradient backpropagation disabled. Meanwhile, the predicted soft assignment distributions Q_1, Q_2 of both "view 1" and "view 2" are updated batch-wise. The final loss is computed by indexing the precomputed P with current batch IDs as follows:

$$\mathcal{L}_{DEC} = KL(P||Q_1, Q_2) = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \sum_{v=1}^{2} \sum_{j=1}^{k} p_{ij} \log \frac{p^v_{ij}}{q_{ij}} \tag{8}$$

here, $\mathcal{B}_t$ denotes the samples in the t-th batch. Notably, when introducing the DEC loss, since the target distribution P requires global computation, we must disable the shuffle operation in the dataloader to prevent batch index misalignment.

Pre-Training and Fine-Tuning

Pre-Training Loss. The complete loss function is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{inv} + \mathcal{L}_{align} + \lambda_d * \mathcal{L}_{DEC} \tag{9}$$

During the pre-training phase, only $\mathcal{L}_{inv}$ and $\mathcal{L}_{align}$ losses are utilized in the first 300 epochs to construct robust representations, while $\mathcal{L}_{DEC}$ is introduced in the remaining 50 epochs to enhance feature clustering effects.

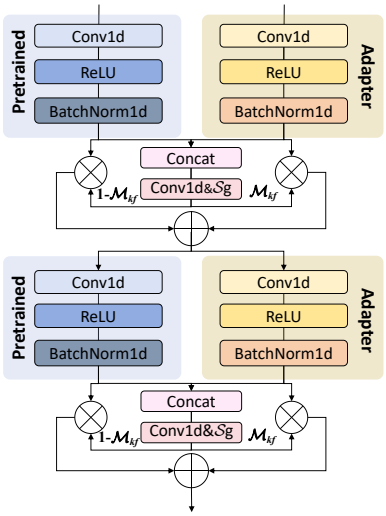


Figure 4: Overview of the proposed Adapter.

Predicate Adapter with Knowledge Filtering Gate. To address the learning conflict of predicate during the fine-tuning stage, a knowledge-filtering gate-equipped adapter, as shown in Figure 4, is devised for the edge encoder.

First, a three-layer MLP identical to the relation encoder is constructed as the fine-tuning adapter, and it connect with the pretrained edge encoder in parallel. Each layer produces two outputs: one from the pre-trained edge encoder and another from the adapter.

Inspired by (Jacobs et al. 1991; Hochreiter and Schmidhuber 1997), we employ a gating mechanism for information filtering. Given the features $f_{\mathcal{A}}$ and $f_{\mathcal{P}}$ from the adapter branch and pre-trained edge encoder respectively, the knowledge-filtered feature fusion is implemented by the following formula:

$$\begin{cases} \mathcal{M}_{kf} = Sigmoid\Big(Conv\big(Concat(f_{\mathcal{A}}, f_{\mathcal{P}})\big)\Big) \\ f_{\mathcal{E}} = \mathcal{M}_{kf} * f_{\mathcal{A}} + (1 - \mathcal{M}_{kf}) * f_{\mathcal{P}} \end{cases} \tag{10}$$

where, $f_{\mathcal{E}}$ is the fusion features, $\mathcal{M}_{kf}$ is the filtering gate.

Experiments

Experiments Setup

Pretraining Setup. This paper selects 900 scenes from the ScanNet train-set to construct the pre-training dataset. The dataset is divided into subsets, yielding a total of 9,208 scene subset segments, each containing no fewer than 6 predicate-relationship triplets. AdamW is employed as the optimizer. CosineAnnealingLR is our learning rate scheduler with a Linear Warm-up, where the initial learning rate is set to 2e-3. The batch size is 128. λ_d is set to 5. λ_b is set to 0.005 according to (Zbontar et al. 2021).

Fine-Tuning Setup. For scene graph fine-tuning, we adopt 3DSSG dataset. In fine-tuning phase, we employ the same optimizer and scheduler as in pre-training, the learning rate of our pretrained network $\mathscr{F}_\theta$ is set to 1e-4, and the learning rate of adapters and predicted heads is set to 4e-4. The fine-tuning process runs for 100 epochs with a batch size of 16. Each object contains 128 points. The experimental device utilizes an NVIDIA RTX 3090 GPU.

Evaluation Metrics

The top-k accuracy "A@k" is employed to evaluate object and predicate classification. The mean top-k accuracy "mA@k" was used to assess the impact of long-tail categories in predicates. Following 2D SGG (Xu et al. 2017; Yang et al. 2018), Scene Graph Classification (SGCLs, with graph-constrained) and Predicate Classification (PredCls) are introduced to evaluate predicate recall capability - both employing the Top-k recall "R@k".

Comparison with State-of-the-Art Methods

Based on the currently open-source SGFN (Wu et al. 2023) and VL-SAT (Wang et al. 2023) frameworks, we construct scene graph networks incorporating point cloud encoders such as PointNet (Qi et al. 2017a), PointNet++ (Qi et al. 2017b), and PointTransformer (Zhao et al. 2021). For each encoder variant, we systematically validated the effectiveness of our proposed pre-training approach denoted as "w/ PR", with DepthContrast's pre-training scheme being abbreviated as "w/ DC" for comparison.

Quantitative Results. As shown in Table 1. It is obvious that integrating our proposed pre-training scheme and adapter leads to consistent performance improvements across all classifications and different point cloud encoders. Particularly in "VL-SAT$_{\text{pointTrans w/ PR}}$", it achieves the improvements of 2.96 on "A@1" for object classification, 5.57 on "mA@1" for predicate classification, and 5.6 on "mA@50" for triplet classification. While "VL-SAT $_{\text{PointNet++ w/ DC}}$" shows positive performance in object classification, it results in a 2.64 decrease on predicate "mA@1" metrics compared to "VL-SAT $_{\text{PointNet++ w/ PR}}$".

Furthermore, our pre-training scheme demonstrates the superiority in handling class imbalance. Specifically, "SGFN$_{\text{PoinNet w/ PR}}$" shows significant gains of 14.21 on "mA@1" for predicate classification. Finally, under the "mR@20", "VL-SAT$_{\text{pointTrans w/ PR}}$" outperforms "VL-SAT$_{\text{pointTrans}}$" by 6.5 on SGCls and by 4.5 on PredCls.

Quantitative Results on Predicate. On 3DSSG dataset, predicate relationships can be categorized into Head, Body, and Tail, according to occurrence frequency. As shown in Table 2, the results demonstrate that "SGFN $_{\text{pointTrans w/ PR}}$" and "VL-SAT $_{\text{pointTrans w/ PR}}$" achieve comprehensive performance improvements. Particularly in Tail category, "VL-SAT $_{\text{pointTrans w/ PR}}$" shows the obvious gains of 11.04 and 10.52 on "mA@3" and "mA@5" compared to VL-SAT, respectively.

As shown in Figure 5, "VL-SAT $_{\text{PointNet w/ PR}}$" shows better performance on long-tailed predicates. While it still slightly under-performs the original VL-SAT on horizontal predicate (the 2nd, 3rd, 6th, and 7th bars from the left). One possible reason is that our method emphasizes semantic predicate learning, neglecting the learning balance of semantic and geometric predicates.

Quantitative Results on Unseen Triplets. Unseen categories are define as those triplets in the 3DSSG test set that

Model	Object		Predicate				Triplet		SGCLs	PredCLs
	A@1	A@5	A@1	A@3	mA@1	mA@3	mA@50	mA@100	mR@20/50/100	mR@20/50/100
SGPN (Wald et al. 2020)	50.32	74.56	89.89	98.15	40.63	63.41	52.74	65.58	19.7/22.6/23.1	32.1/38.4/38.9
SGG_{point} (Zhang et al. 2021a)	51.42	74.56	92.4	97.78	27.95	49.98	45.02	56.03	21.5/22.8/23.4	29.3/36.7/38.4
$SGFN_{PointNet}$	53.67	77.18	90.19	98.17	41.89	70.82	58.37	67.61	20.5/23.1/23.1	46.1/54.8/55.1
$SGFN_{PointNet++}$	54.43	78.42	89.85	98.08	45.68	73.26	59.32	70.33	24.2/24.8/26.4	49.6/56.7/56.5
$SGFN_{pointTrans}$	55.07	78.26	89.54	97.55	42.86	71.64	60.57	69.73	23.6/25.5/25.8	47.8/56.3/57.4
$VL\text{-}SAT_{PointNet}$	55.66	78.66	89.81	98.45	54.03	77.67	65.09	73.59	31.0/32.6/32.7	57.8/64.2/64.3
$VL\text{-}SAT_{PointNet++}$	55.27	78.84	89.43	98.21	52.69	77.34	63.61	72.47	31.8/32.4/33.7	56.6/63.7/64.9
$VL\text{-}SAT_{pointTrans}$	55.38	79.02	89.76	97.92	52.86	74.59	62.97	72.56	30.2/32.6/32.9	55.3/63.8/64.1
SGFN $_{PointNet++\ w/\ DC}$	58.31	79.94	89.20	97.56	50.72	73.44	61.79	72.30	30.7/32.4/33.2	53.3/58.2/63.6
VL-SAT $_{PointNet++\ w/\ DC}$	58.49	80.13	89.75	98.13	52.68	74.24	63.24	71.87	31.0/32.8/33.4	55.3/63.6/64.1
SGFN $_{PointNet\ w/\ PR}$	56.38	79.06	90.42	98.47	56.10	77.05	65.82	73.94	32.6/33.5/33.8	56.5/64.2/64.6
VL-SAT $_{PointNet\ w/\ PR}$	56.87	79.77	90.86	98.65	56.84	76.21	66.67	74.93	33.6/34.2/36.6	57.4/65.6/66.3
SGFN $_{PointNet++\ w/\ PR}$	57.61	79.72	90.13	98.44	55.67	74.35	65.71	73.49	32.6/33.4/33.7	58.4/65.4/65.6
VL-SAT $_{PointNet++\ w/\ PR}$	57.90	79.18	90.26	98.67	55.32	76.13	65.42	73.27	33.8/35.8/37.4	56.6/65.3/66.5
SGFN $_{pointTrans\ w/\ PR}$	**58.76**	**80.53**	90.46	98.64	56.51	77.49	66.63	74.18	35.2/38.1/**39.4**	58.4/66.5/68.9
VL-SAT $_{pointTrans\ w/\ PR}$	58.34	80.26	**91.03**	**98.96**	**58.43**	**79.63**	**68.57**	**76.89**	**36.7/38.2**/38.8	**59.8/68.3/69.6**

Table 1: Comparisons with state-of-the-arts on the 3DSSG dataset.

Model	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 (Wald et al. 2020)	96.66	99.17	66.19	85.73	10.18	28.41	15.78	29.60	66.60	77.03
SGFN (Wu et al. 2023)	95.08	99.38	70.02	87.81	38.67	58.21	22.59	35.68	71.44	80.11
VL-SAT (Wang et al. 2023)	96.31	99.21	80.03	93.64	52.38	66.13	31.28	47.26	75.09	82.25
Chen *et al.* (Chen et al. 2024)	98.54	**99.78**	84.72	**96.03**	61.24	75.91	36.72	52.47	80.58	**88.92**
SGFN $_{pointTrans\ w/\ PR}$	97.89	99.63	82.47	94.59	58.96	73.22	37.64	51.25	78.91	86.44
VL-SAT $_{pointTrans\ w/\ PR}$	**98.67**	99.53	**86.25**	95.36	**63.42**	**76.65**	**39.75**	**55.83**	**83.26**	88.37

Table 2: Quantitative results on predicate and unseen triplets.

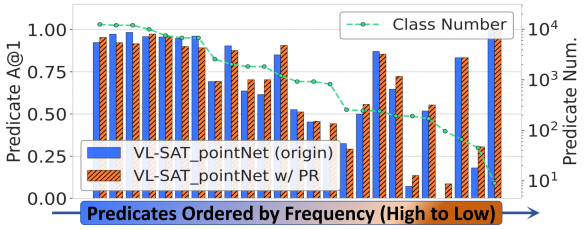


Figure 5: Predicate A@1 for all predicate categories.

BT	CM	DEC	AD_+	AD_{kf}	Object	Predicate		Triplet	
					A@1	mA@1	mA@3	mA@50	mA@100
					53.67	41.89	70.82	58.37	67.61
✓				✓	55.35	48.63	72.28	59.84	68.70
✓	✓			✓	55.24	53.69	74.46	60.15	69.84
✓		✓		✓	55.69	52.92	75.73	59.62	70.42
✓	✓	✓		✓	**56.38**	**56.10**	**77.05**	**65.82**	**73.94**
✓	✓	✓			55.27	45.43	70.66	58.74	69.82
✓	✓	✓	✓		55.86	52.34	73.28	60.41	70.53

Table 3: Ablation study on the proposed pretrained methods and the predicate adapter.

do not appear in the training set. As demonstrated in Table 2, "VL-SAT $_{pointTrans\ w/\ PR}$" achieves significant improvements over the (Chen et al. 2024) - increases 3.03 on "A@50" for "Unseen" and increases 2.68 on "A@50" metric for "Seen".

Ablation Study and Analysis

Validation of Pre-Training Strategy. As shown in Table 3, the invariance learning and two regularization terms are evaluated on the 3DSSG dataset. We generate four pretrained weights by combining different regularization components. These weights are used to initialize $SGFN_{PointNet}$ with our proposed adapter "AD_{kf}" for evaluation.

The results demonstrate that using only Barlow Twins for invariance learning yields improvements in object classification - increasing 1.68 on "Object A@1" compared to $SGFN_{PointNet}$. After incorporating cross-modal regularization,"SGFN w/BT+CM" achieves 11.8 improvement on predicate "mA@1".

The complete solution ("SGFN w/BT+CM+DEC") integrates cross-modal alignment and DEC to enhance semantic discrimination. It delivers 14.21 increase on predicate "mA@1" and 7.45 increase on triplet "mA@50".

Validation of Adapter Fine-Tuning. As shown in Table 3, After removing the adapter, the predicate "mA@1" decreased by 10.67. Switching to element-wise addition "AD_+" for feature fusion results in a decline of 3.76 on

Dataset size	(Batch Size)		
(Scene splits)	64	128	256
4672	59.93	**62.87**	60.05
9208	60.48	**65.82**	65.16

Table 4: Impact of batch size and data size on triplet accuracy (mA@50, %).

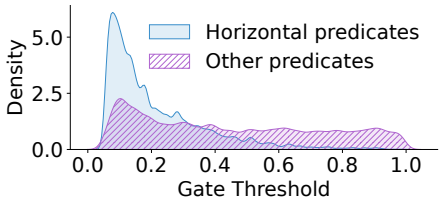


Figure 6: Distribution of knowledge filtering gate threshold.

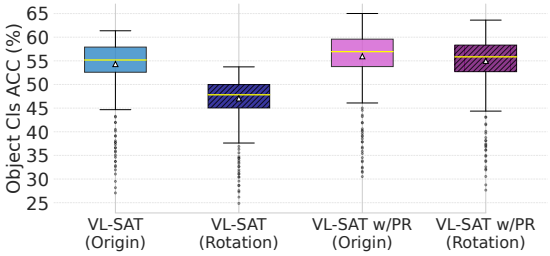


Figure 7: Analysis of rotation invariance.

predicate "mA@1". Only the knowledge-filtered gate-based adapter can effectively reuse and learn knowledge.

In the validation set, we collect the knowledge filter gate thresholds across all categories. Figure 6 demonstrates that the gate actively inhibits knowledge incorporation from the pre-trained model upon receiving horizontal orientation predicates, whereas other relational predicates yield uniformly distributed gate thresholds.

Impact of Batch Size and Data. As shown in Table 4, the triplet classification performance shows a positive correlation with the increasing volume of pre-training data across different batch sizes, and it achieves optimal performance at a batch size of 128 (tests on $\text{SGFN}_{\text{PointNet}}$).

Analysis of Rotation Invariance. The invariance to horizontal rotation about the z-axis is crucial for scene graph tasks. As demonstrated in Figure 7, $\text{VL-SAT}_{\text{PointNet w/PR}}$ maintains robust object classification performance under Z-axis rotations of the scene. In contrast, the original VL-SAT exhibits substantial performance degradation due to its limited rotation equivariance.

Analysis of Predicate Clustering Effects. As illustrated in Figure 8. The latent features from 3DSSG are extracted by the pre-trained model without fine-tuning. Notably, the pretrained model without regularization exhibits chaotic effect, confirming that pure invariance learning struggles to acquire meaningful semantic clustering. The pretrained model with cross-modal regularization achieves preliminary clustering effects. After applying the deep clustering, the enhanced semantic discriminative is achieved.

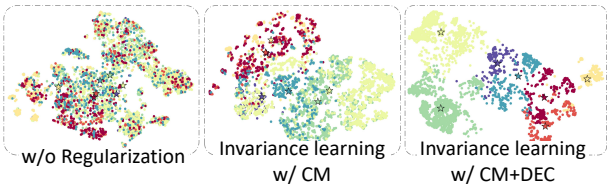
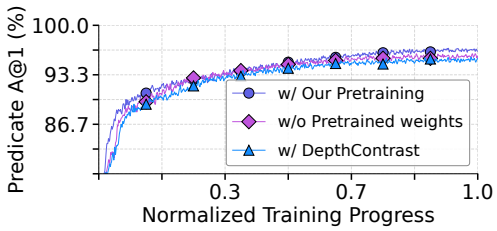
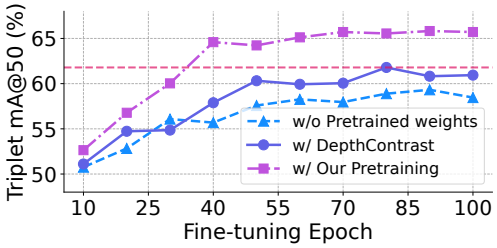


Figure 8: Predicate Clustering Effect Analysis.



(a) Predicate A@1 (Train)



(b) Triplet mA@50 (Val)

Figure 9: Accuracy curve Comparison.

Analysis of Accuracy Curves. We compare classification accuracy curves for both predicates and triplets across three conditions: 1) weights from the depthContrast, 2) our pre-trained weights, and 3) a baseline ($\text{SGFN}_{\text{PointNet++}}$) without pre-training. As shown in Figure 9, Our method achieves faster predicate accuracy convergence during training and superior triplet classification performance versus baseline and DepthContrast on validation.

Conclusions

To solve the absence of predicate representation learning in the point cloud pre-training of 3D scene graph generation, this paper proposes a new self-supervised multi-view invariance learning scheme with collaborative cross-modal regularization. It utilizes a non-contrastive learning to capture the semantic consistency under data augmentation, with a complement of cross-modal regularization to enhance semantic discriminability. To resolve the knowledge conflict in fine-tuning, an adapter with knowledge filtering gate is proposed to selectively fuses pre-trained knowledge, avoiding catastrophic forgetting. Comprehensive experiments validate the effectiveness and superiority of our new scheme. In 3D scene graph generation, it could obviously promote the primary performance metrics of existing methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No.62476049.

References

Chen, L.; Wang, X.; Lu, J.; Lin, S.; Wang, C.; and He, G. 2024. Clip-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27863–27873.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Feng, M.; Hou, H.; Zhang, L.; Guo, Y.; Yu, H.; Wang, Y.; and Mian, A. 2023. Exploring hierarchical spatial layout cues for 3D point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 27: 731–743.

Feng, M.; Yan, C.; Wu, Z.; Dong, W.; Wang, Y.; and Mian, A. 2025. Hyperrectangle embedding for debiased 3D scene graph prediction from RGB sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Koch, S.; Hermosilla, P.; Vaskevicius, N.; Colosi, M.; and Ropinski, T. 2024a. Lang3dsg: Language-based contrastive pre-training for 3d scene graph prediction. In *2024 International Conference on 3D Vision (3DV)*, 1037–1047. IEEE.

Koch, S.; Hermosilla, P.; Vaskevicius, N.; Colosi, M.; and Ropinski, T. 2024b. Sgrec3d: Self-supervised 3d scene graph learning via object-level scene reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3404–3414.

Koch, S.; Vaskevicius, N.; Colosi, M.; Hermosilla, P.; and Ropinski, T. 2024c. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14183–14193.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Pang, Y.; Tay, E. H. F.; Yuan, L.; and Chen, Z. 2023. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1: 2440001.

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

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

Qi, Z.; Dong, R.; Fan, G.; Ge, Z.; Zhang, X.; Ma, K.; and Yi, L. 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, 28223–28243. PMLR.

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

Wald, J.; Dhamo, H.; Navab, N.; and Tombari, F. 2020. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Cheng, B.; Zhao, L.; Xu, D.; Tang, Y.; and Sheng, L. 2023. 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*, 21560–21569.

Wu, S.-C.; Tateno, K.; Navab, N.; and Tombari, F. 2023. Incremental 3d semantic scene graph prediction from rgb sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5064–5074.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, 574–591. Springer.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419.

Yang, J.; Lu, J.; Lee, S.; Batra, D.; and Parikh, D. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, 670–685.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, 12310–12320. PMLR.

Zhang, C.; Yu, J.; Song, Y.; and Cai, W. 2021a. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9705–9715.

Zhang, S.; Hao, A.; Qin, H.; et al. 2021b. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances*

in Neural Information Processing Systems, 34: 18620–18632.

Zhang, Z.; Girdhar, R.; Joulin, A.; and Misra, I. 2021c. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10252–10263.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.