GranSSG: Correlating Volumetric Granularities for 3D Semantic Scene Graph Prediction

Kaixiang Huang, Qifeng Zhang, Jin Wang, Jingru Yang, Yang Zhou, Jiao Yi, Guodong Lu, Shengfeng He, *Senior Member, IEEE*

***Abstract*—Predicting 3D Semantic Scene Graphs (3DSSG) is vital for understanding complex scenes by constructing structured representations. Current methods struggle with significant granularity discrepancies among instances, often relying on features at a single scale, which hampers their ability to perceive and interact with differently sized instances. To tackle this challenge, we introduce GranSSG, a novel approach that integrates volumetric granular awareness into 3DSSG prediction. Central to GranSSG is the Volumetric Pooling block, which aggregates features from multiple instance volumes, enhancing the representation of instance patterns across different granularities. Complementing this, the Granularity Transformer block dynamically directs attention to instance features across various network layers, ensuring precise perception of instances regardless of their granularity. Furthermore, the Cross-Granularity Correlation Transformer block mitigates performance degradation in instance pair relationship prediction by adaptively fusing hybrid features from different granularities, providing a comprehensive representation of instance pairs. Extensive evaluations on the challenging 3DSSG benchmark demonstrate that GranSSG significantly enhances prediction performance, setting a new state-of-the-art in 3DSSG prediction.**

***Index Terms*—Point Cloud, 3D Semantic Scene Graph, Granularity-awareness**

I. Introduction

UNDERSTANDING complex 3D real-world environments is fundamental for advancing fields such as AR/VR, autonomous driving, and robotic intelligence [1]. In this context, the prediction of 3D Semantic Scene Graphs (3DSSG) has gained significant attention. Given a 3D point cloud environment with class-agnostic instance masks, the 3DSSG prediction task aims to identify the category of each instance and determine the potential relationships between instance pairs. By representing labeled instances as nodes and their relationships as edges, the constructed graph provides a structural description of the 3D geometric scene.

However, the real-world environment presents substantial challenges for 3DSSG prediction. As illustrated in Figure 1, instances in a 3D scene exhibit significant granularity discrepancies, more pronounced than in 2D due to the additional dimension. Despite these cross-order magnitude size differences, current state-of-the-art approaches [3]–[5] rely indiscriminately on single-granularity feature analysis, degrading 3DSSG prediction performance. Specifically, identifying fine-grained smaller instances (e.g., *lamp*) requires high-resolution, content-descriptive features [6], while coarse-grained items (e.g., *bed*) require features with larger receptive fields for comprehensive perception. Additionally, for instance pairs (e.g., $<lamp, bed>$), it is essential to extend granularity-awareness to process multi-granularity feature combinations, selecting comprehensive hybrid-granularity features to determine their potential relationships. However, the lack of multi-granularity feature analysis in 3DSSG has led to ignored differences in instance information at each granularity, causing attention bias and diminishing discrimination [7].

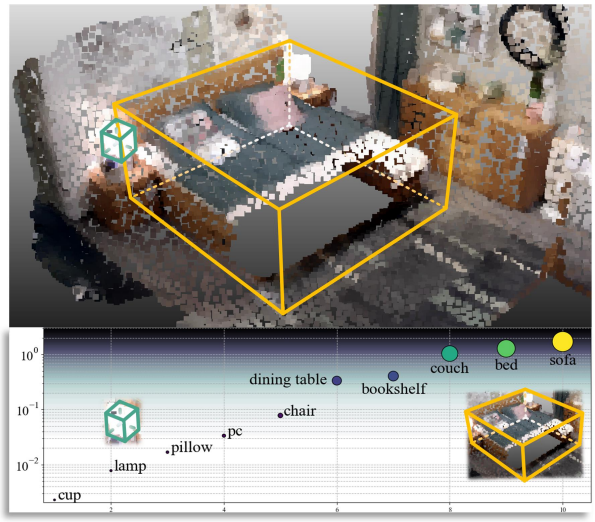


Fig. 1. Illustration of the significant instance granularity discrepancy in a complex indoor scene. In real-world environments, small items and large furniture often coexist, leading to size discrepancies spanning several orders of magnitude (e.g., the lamp and bed). We present typical instances from the 3DSSG dataset [2] and compare their average volumes to highlight these discrepancies. This comparison underscores the importance of granularity-awareness, achieved through adaptive multi-granularity analysis.

Current 3DSSG prediction methods, such as SGG_{point} [8] and SGFN [9], building on SGPN [2], rely heavily on PointNet [10] and Graph Convolutional Networks (GCNs) [11], overlooking the importance of multi-granularity awareness. This approach is vulnerable to scenarios with varying instance sizes, as the backbone serves only one granularity. Recent

This work is supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01020, 2025C01091), National Natural Science Foundation of China (52175032), Key Technology Breakthrough Plan Project of "Science and Technology Innovation Yongjiang 2035" (2024Z295, 2023Z218), Yuyao Science and Technology project (No.2023JH03010019), and Robotics Institute of Zhejiang University under Grant K12107

Kaixiang Huang, Qifeng Zhang, Jin Wang, Yang Zhou, Jiao Yi, and Guodong Lu are with the State Key Laboratory of Fluid Power & Mechatronic Systems, Zhejiang University, Hangzhou, China. (email: dwjcom@zju.edu.cn)

Jingru Yang is with the School of Computer Science, Carnegie Mellon University, Pennsylvania, USA. (email: jingruyang1617@gmail.com)

Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore. (email: shengfenghe@smu.edu.sg)

works [3], [5], [12] leveraging multi-modal features and the CLIP model [13] to assist 3DSSG prediction, like CCL-3DSGG [3] aligning image and language information with the 3D point cloud, still neglect the size discrepancy attribute. The deficiency in granularity-awareness undermines overall performance. Although Granular3D [4] employs a hierarchical network to separately address instances and instance pairs, it still adheres to single-granularity analysis. Multi-granularity analysis for 3D scenes with significant size discrepancies, and adaptive fusion of instance features and cross-granularity instance pair features, remains an open problem requiring urgent attention.

In contrast, with the emergence of hierarchical feature extraction methods such as PointNet++ [14], multi-granularity 3D point cloud analysis has seen considerable progress [15]–[19]. For example, PPT-Net [15] uses a pyramid structure in point cloud retrieval to capture discrimination at different scales, boosting local area perception. Similarly, Retro-FPN [17] summarizes semantic contexts across granularities, refining point features in each stage, enhancing segmentation performance in complex small regions. However, these advancements overlook the weaker information-carrying capacity of point clouds at lower-scale layers due to limited feature dimensions, potentially diminishing network attention to features at lower-scale layers. Current 3DSSG prediction paradigms typically apply pooling layers to directly aggregate instance features [4], [5], [9]. This strategy often leads to increased feature loss and hinders balanced instance pattern perception across granularities, ultimately reducing the effectiveness of multi-granularity analysis.

To address the limitations posed by single-granularity feature extraction in environments with varying instance sizes, we introduce a novel granularity-awareness scheme for 3DSSG prediction, termed GranSSG. Diverging from traditional methods, GranSSG's adaptive fusion of features across volumetric granularities provides a comprehensive pattern representation, effectively modeling instances of distinct sizes simultaneously. Unlike current multi-granularity analysis strategies, which lack balanced information at different granularities, GranSSG features the Volumetric Pooling block (VP). This block aggregates features from multiple instance volumes, with lower-scale layers containing more patches, expanding the representation of instance patterns and alleviating the information imbalance in multi-granularity features. Furthermore, the Granularity Transformer block (GrT) and Cross-Granularity Correlation Transformer block (CGCT) enable multi-granularity feature fusion for instances and subject-object instance pairs (i.e., nodes and edges in the semantic scene graph). By shunting attention to instance features at various granularities, the GrT adaptively fuses fine-grained and coarse-grained features for targeted instance identification. The CGCT module identifies suitable granularity combinations for each instance pair to represent their potential relationships. It explicitly assesses the interplay among cross-granularity feature combinations and enables adaptive fusion of hybrid-granularity features. This leads to more comprehensive instance pair representations and improves relationship prediction.

Extensive evaluations demonstrate that GranSSG significantly outperforms existing 3DSSG prediction methods, establishing a new state-of-the-art performance.

In summary, our main contributions are fourfold:

1) We introduce GranSSG, the first approach to emphasize granularity-awareness in the 3DSSG prediction task. This method adapts to the significant size discrepancies among coexisting instances in the environment, achieving superior performance.
2) To address the imbalance of information across different granularities, we propose the Volumetric Pooling block (VP) in GranSSG. This straightforward yet effective strategy expands the representation of instance patterns from split volumes, directly alleviating the information imbalance problem of multi-granularity features.
3) We design the Granularity Transformer block (GrT) and the Cross-Granularity Correlation Transformer block (CGCT) to enable adaptive multi-granularity fusion of instances and subject-object instance pairs in the scene, significantly enhancing perception through the constructed hybrid cross-granularity feature representation.
4) Extensive experimental evaluations demonstrate the significant advancements brought by GranSSG. Notably, our approach achieves a new state-of-the-art in 3DSSG prediction, highlighting the critical importance of granularity-awareness in this field.

II. Related Work

Scene Graph Prediction in Point Clouds. Significant progress has been made in 3D point cloud semantic scene graph generation, which aims to identify the categories of instances while defining their potential relationships, thereby organizing structured representations of the environment. As a pioneer in this field, [2] introduced the 3DSSG benchmark along with the SGPN model, establishing the paradigm of PointNet [10] combined with GNN [11]. Building upon this framework, subsequent works have continued to make advancements. Specifically, $\text{SGG}_{\text{point}}$ [8] employs an edge-oriented GNN for explicit relationship modeling, enhancing the interaction between instances and instance pairs. SGFN [9] explores incremental 3DSSG prediction through RGB-D sequences. Additionally, VL-SAT [5] incorporates 2D images and natural language descriptions to assist in scene representation. CCL-3DSGG [3] further investigates cross-modality feature alignment via CLIP [13], achieving open-vocabulary prediction. Recognizing the limitations of single-granularity point cloud feature analysis, the recent work Granular3D [4] diverges from the foundational SGPN architecture by employing a hierarchical network that processes the features of instances and their associated pairs at corresponding granularities, achieving superior performance. Yet, the Granular3D merely selects two targeted single-granularity representations for instances and instance pairs, rather than introducing adaptive multi-granularity analysis. Despite these advancements, few works have addressed the significant size discrepancies among instances in the scene, which impacts overall performance and hinders the extraction and fusion of multi-granularity features for adaptive perception.

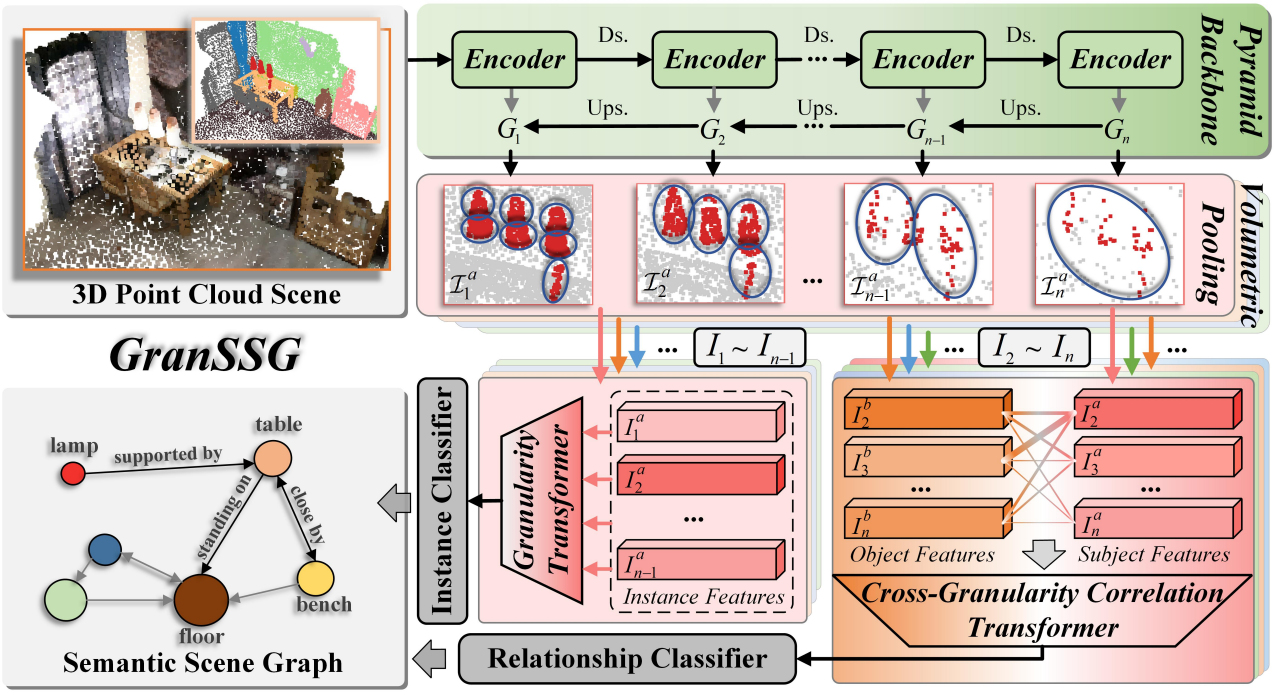


Fig. 2. Overview of the proposed GranSSG, where the stacked colored blocks represent the parallel processing of different instances and instance pairs within a scene. Initially, a pyramid backbone extracts multi-granularity point cloud features. The Volumetric Pooling block (VP) is then utilized on feature maps at all granularities, balancing the information contained across various scales. For single instance identification and instance pair relationship classification required for 3DSSG prediction, the Granularity Transformer block (GrT) and Cross-Granularity Correlation Transformer block (CGCT) are applied respectively. Specifically, the proposed GrT shunts the attention of features at different granularities, leading to a targeted representation. For subject-object instance pairs, the CGCT effectively assesses and fuses all potential cross-granularity feature combination pairs. Altogether, GranSSG adapts to the significant size discrepancies between coexisting instances in the environment, significantly elevating 3DSSG prediction performance.

Multi-granularity Feature Analysis. As a crucial component in addressing the coexistence of instances with diverse sizes in the scene, multi-granularity feature analysis has been widely studied and applied [20]–[22]. FPN [23] is a pioneering work that leverages multi-granularity features in 2D object detection tasks. Seeking to replicate this success, multi-granularity feature analysis has been explored in various 3D tasks [15], [17]–[19]. Following the paradigm of PointNet++ [14], various research with hierarchical networks have emerged [24]–[26]. Building upon this foundation, PatchAugNet [19] employs a pyramid structure to simultaneously capture local and global discriminators, enhancing perspective ability. In the 3D semantic segmentation field, Retro-FPN [17] summarizes semantic contexts across multiple granularities, retrospectively refining point features at each stage, achieving competitive performance in complex small regions. For 3D object detection, M3DETR [27] unifies multiple point cloud representations under multiple feature granularities, particularly strengthening the perception of distant objects. However, despite its effectiveness, multi-granularity feature analysis has not been employed in the 3DSSG prediction field, which is both unreasonable and illogical. Understanding complex 3D scenes with instances exhibiting cross-order magnitude size discrepancies urgently requires adaptive multi-granularity feature representation for both instances and instance pairs.

Feature Imbalance Problem. The success of multi-granularity feature analysis underscores the importance of leveraging both low-level and high-level information. While FPN [23] enriches low-granularity information via a top-down semantic feature pathway, and PANet [21] complements this with a bottom-up fusion of high-resolution features, these techniques primarily address multi-granularity feature transfer without resolving the imbalance of semantic richness between granularities. Separately, pyramid-based pooling methods like PPM [28]–[31] aim to capture contextual information at multiple receptive fields for a more comprehensive perception. Yet, these strategies operate on a single feature map and do not reconcile the semantic disparities. To mitigate this, Libra R-CNN [7] and its successors [32] propose to rescale multi-granularity features to a unified intermediate size. However, such approaches merely rely on uniform compression or expansion, failing to achieve a fundamental balance across granularities. In contrast, the introduced Volumetric Pooling mechanism explicitly enriches lower-granularity features with 3D volumes. This allows each granularity level in our architecture to carry balanced information, providing a solid foundation for effective multi-granularity scene understanding.

III. Methodology

3D Semantic Scene Graph (3DSSG) prediction aims to transform complex point cloud scenes into structured graph representations, which is crucial for understanding 3D environments. Given the significant size variations among instances in real-world scenes, we propose GranSSG, a method that focuses on multi-granularity feature analysis. To elucidate our proposed GranSSG, we first formulate the 3DSSG prediction task and provide an overview of our method. We then offer a detailed and progressive introduction to GranSSG, including

the backbone structure, Volumetric Pooling block (VP), Granularity Transformer block (GrT), Cross-Granularity Correlation Transformer block (CGCT), and the training objective.

A. Problem Formulation and Method Overview

Consider a 3D point cloud scene $\mathcal{X}$ that contains M instances, where each point x is associated with a class-agnostic instance mask $x_{mask} \in \{1, ..., M\}$ (provided by an instance segmentation method such as Mask3D [33] or a dataset), distinguishing individual instances $\{\mathcal{I}^1, ..., \mathcal{I}^M\}$ within $\mathcal{X}$. As illustrated by the instance mask in Figure 2, there is a noticeable size discrepancy between instances. Given the 3D point cloud scene as input, the ultimate goal of 3DSSG prediction is to construct a scene graph $\mathcal{G} = (\mathcal{C}, \mathcal{R})$ [2]. Specifically, the category cat_a of each instance $\mathcal{I}^a$ in $\mathcal{X}$ represents a node in $\mathcal{G}$, while the relationship rel_{ab} between a pair of subject-object instances $\mathcal{I}^a$ and $\mathcal{I}^b$ represents a directional edge in $\mathcal{G}$, starting from the subject instance (head node) to the object instance (tail node).

To address the environment where instances with significant size discrepancy coexist, the proposed GranSSG pivots towards the balance and adaptive multi-granularity analysis, thereby targeting predicates of every element of the 3D scene graph. As shown in Figure 2, GranSSG comprises four main components. First, a pyramid backbone is employed to encode the point cloud at different resolutions, extracting feature maps $\{\mathsf{G}_1, ..., \mathsf{G}_n\}$, which serve as the foundation for subsequent multi-granularity analysis. Next, to balance the information carried by individual instances in multiple feature maps $\{I_1^a, ..., I_n^a\}$, VP aggregates features from split instance volumes. The volume number controls the expansion rate of instance information at different granularities, providing a solid foundation for multi-granularity feature fusion. Subsequently, we introduce GrT to dynamically shunt the attention of multi-granularity features for each instance, enabling targeted instance identification. Parallelly, we present CGCT to process subject-object instance pairs for relationship prediction. CGCT innovatively applies feature correlation assessment to extract valuable hybrid cross-granularity features of instance pairs from the manifold possible cross-granularity feature combinations, resulting in a more comprehensive representation.

B. Backbone Structure

Recognizing the significant size discrepancies of instances in 3D point cloud scenes, it is crucial to extract both descriptive fine-grained features and coarse-grained features with larger receptive fields. As demonstrated in the green part of Figure 3, we employ a pyramid feature extraction backbone organized across five granularities, utilizing the Point Transformer block [34] as an encoder for point feature extraction at varying resolutions. The Farthest Point Sampling (FPS) method is applied during the down-sampling stage, with sampling rates of $[1, 2, 2, 2, 2]$. Consequently, consider the input scene $\mathcal{X}$ contains N points, the subsequent points produced by each granularity are $[N, N/2\ , N/4\ , N/8\ , N/16\ ]$. As the network deepens, the dimension of a single point expands accordingly, reaching $[32, 64, 128, 256, 512]$ that contains gradually richer semantic information. Subsequently, the features at deeper granularities are individually up-sampled. We employ k-NN inverse distance weighting [14] to aggregate semantic features closest to the corresponding points in higher-resolution granularities. Through a Multi-Layer Perceptron (MLP) and a direct addition process, high-level semantic features are propagated downward. Finally, feature maps at different resolutions are extracted, facilitating subsequent multi-granularity analysis.

C. Volumetric Pooling

Building upon the extracted feature maps across various granularities, the multi-granularity representations of each instance can be easily utilized as fundamental elements for scene graph prediction. However, this raises an unavoidable question: Do the various instance representations aggregated from feature maps of different dimensions (e.g., 32-dim in G_1 and 512-dim in G_5) carry balanced information?

Revisiting of current pooling strategy. To date, 3DSSG methods predominantly apply pooling of all point cloud features for an individual instance to aggregate its representation [2], [5], [9]. As a data compression approach, this pooling method does not alter the original dimension of the instance feature map. Therefore, under multi-granularity conditions, an additional MLP is required to resize the aggregated multi-level features to a uniform size, such as 512-dim. Mathematically, this method can be expressed as follows:

$$\begin{aligned} \mathcal{I}_k^a \in \mathbb{R}^{n_k^a \times d_k} &= \{x \in \mathcal{X}_k | x_{mask} = a\} \\ I_k^a &= \mathrm{MLP}_k\left(\mathrm{Pool}\left(\mathcal{I}_k^a\right)\right) \end{aligned} \tag{1}$$

where $\mathcal{I}_k^a$ denotes the split feature set of the instance a under granularity k, identified via the class-agnostic mask x_{mask}, while the n_k^a and d_k respectively represent the number of points and feature dimension. The $\mathrm{Pool}\left(\cdot\right)$ function represents max pooling, and $\mathrm{MLP}_k\left(\cdot\right)$ represents the mapping function used at granularity k.

Technically, as an affine mapping, the key of MLP function is to reconstruct the feature space, it is not capable of enriching the information carried in the aggregated instance feature [35]. Consequently, due to differences in semantic feature dimensions, deeper instance representations inherently possess a stronger capacity to carry information. Moreover, in the pyramid network architecture that applies point down-sampling, instances at lower-scale layers contain more points, leading to exacerbated feature loss after max pooling. Recognizing the inevitable issue of information asymmetry in existing methods, we propose a novel Volumetric Pooling block (VP), which substantially alleviates multi-granularity feature imbalance by expanding the contained information.

Volumetric Searching. Except for the global instance representation, neighboring points among the instance also form a meaningful 3D subset [10]. Therefore, the key idea of our VP is to collect more information from the split 3D volumes of the target instance, thereby fundamentally expanding the representations at lower-granularity layers. However, the 3D point cloud is unordered, while instances are with significant

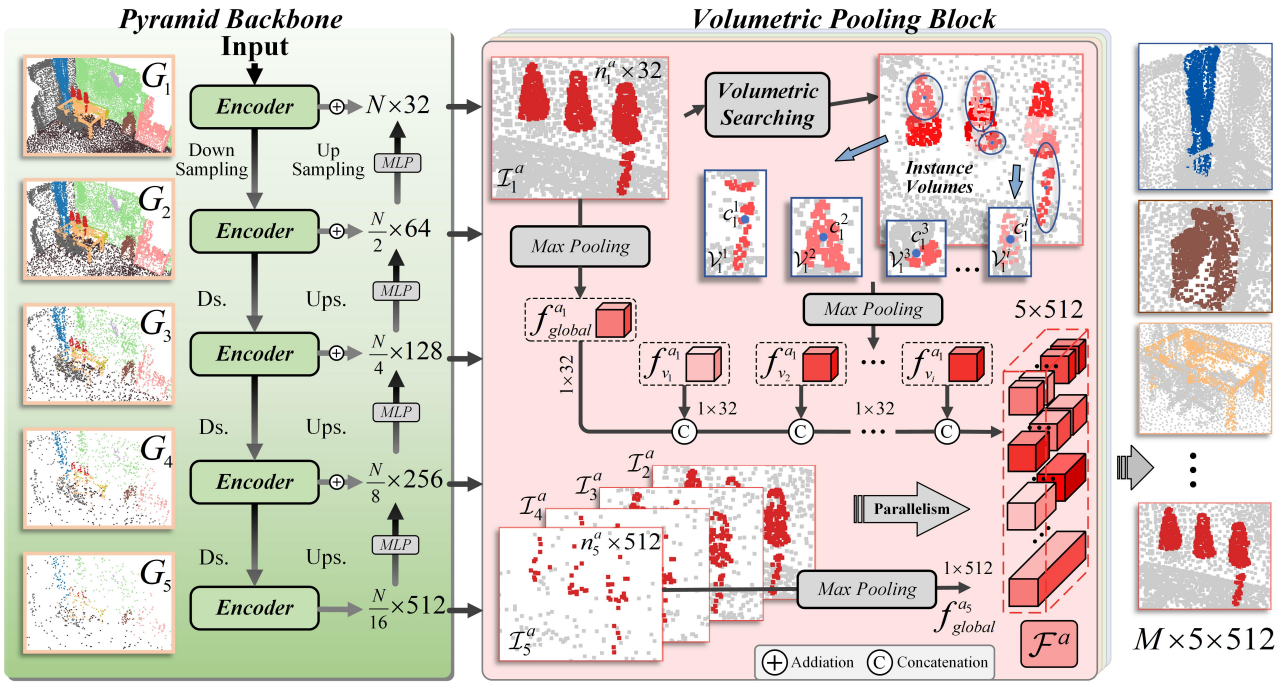


Fig. 3. Illustration of the backbone and our proposed Volumetric Pooling block (VP). The left part shows the pyramid structure applied for multi-granularity point cloud feature extraction, while the sub-images on the left side demonstrate the down-sampled 3D scene under various resolutions. The right part shows the pipeline of the VP. Contrasting with traditional feature pooling strategies, for each instance, VP first splits it into instance volumes that represent local patterns and then reorganizes the volume features after individual pooling. VP operates on every granularity of the instance in parallel. The shallower the feature granularity, the more volumes VP will search and combine. This approach alleviates the feature loss in shallow scales, enriches the contained information, and balances the features across various granularities.

size discrepancies. To adaptively capture the informative local pattern from instances, we state two principles for the volumetric searching process: 1) Instance volumes should be evenly distributed in the 3D space to avoid overlap, and 2) Merging all volumes should approximate the whole instance for comprehensive representation. For this, when processing the target $\mathcal{I}_k^a$, the FPS method is innovatively applied to search the centers of volumes, ensuring that the selected points evenly cover the entire 3D point set [36]. Subsequently, for each center point c_k^n, we perform a k-Nearest Neighbor (kNN) search to find surrounding points, constructing the volumes $\mathcal{V}_k^n$. Compared to simpler ball-query with a fixed radius, kNN can better search the local 3D structure of irregular instances, while ensuring different instance volumes contain the same number of points, balancing the information distribution. Through the cooperation of FPS and kNN, the split volumes adaptively represent the local patterns of instances, maintaining robustness when dealing with irregular 3D instances. As illustrated in Figure 3, the VP captures the *lamp post* and *lampshade* from the complex point cloud of the *lamp*, with each subset representing a valuable local pattern. In summary, the Volumetric Searching in VP is expressed as follows:

$$\begin{aligned} \mathbf{C}_k^a &= \left\{c_k^1, ..., c_k^i\right\} = \mathrm{FPS}\left(\mathcal{I}_k^a, i\right) \\ \mathbf{V}_k^a &= \left\{\mathcal{V}_k^1, ..., \mathcal{V}_k^i | \mathcal{V}_k^t = \mathrm{kNN}\left(c_k^t, \mathcal{I}_k^a, n_k^a/i\right)\right\} \end{aligned} \tag{2}$$

where $\mathbf{C}_k^a$ denotes the set of volume centers, and $\mathbf{V}_k^a$ represents the instance volumes set identified by VP. The i denotes the number of selected volume centers, while the n_k^a/i represents the sampled points when searching each instance volume.

Pooling and Reorganization. Unlike previous methods that rely on additional MLP layers, our VP method aligns the dimensions of instance semantic features at various granularities by pooling and reorganizing the identified instance volumes. Taking the $\mathcal{I}_k^a$ as an example, VP applies max pooling uniformly across all instance volumes in $\mathbf{V}_k^a$. Consequently, each aggregated volume feature $f_{v_n}^{a_k}$ will describe a targeted instance local feature in detail. Furthermore, compared to pooling the entire instance, since fewer points are distributed in each volume, the information loss from the pooling process is correspondingly reduced. Additionally, we maintain global instance pooling to capture the overall pattern, as represented in Eq.1, constructing the global feature $f_{global}^{a_k}$.

The instance volume features and the global feature obtained are then reorganized through a straightforward feature concatenation process. This scheme preserves the independence of information in each part, effectively enhancing the instance features across different granularities. The mathematical expression can be described as follows:

$$\begin{aligned} f_{v_i}^{a_n} &= \mathrm{Pool}\left(\mathcal{V}_i\right), \ f_{global}^{a_k} = \mathrm{Pool}\left(\mathcal{I}_k^a\right) \\ I_k^a &= \mathrm{Concat}\left(f_{global}^{a_k}, \underbrace{f_{v_1}^{a_k}, ..., f_{v_i}^{a_k}}_{i \text{ volume features}}\right) \end{aligned} \tag{3}$$

Meanwhile, to ensure the dimensional uniformity of integrated semantic features across various granularities $\mathcal{F}_a = \{I_1^a, ..., I_5^a\}$, which are aggregated from feature maps with different dimensions, the number of searched volumes is carefully controlled. For clarity, Table I outlines the VP strategy at each

TABLE I
VOLUMETRIC POOLING STRATEGY UNDER DIFFERENT GRANULARITIES.

Granularity	Global	Volume	Dimension
1	1	15	[32 ; 32 × 15]
2	1	7	[64 ; 64 × 7]
3	1	3	[128 ; 128 × 3]
4	0	2	[0 ; 256 × 2]
5	1	0	[512 ; 0]

granularity. Intuitively, at layers with lower down-sampling rates, VP adaptively introduces more instance volumes (e.g., the parameter in Eq.2) to enrich semantic features and mitigate the information loss caused by the compression of massive point features. Specifically, the hyperparameter configuration strictly corresponds to the feature dimensions of each granularity level, where the combined features composed of the global feature and volume features are unified to 512-dim across all granularities. This also eliminates the need for additional MLP layers for dimension alignment. Notably, the feature reorganization strategy of VP slightly differs in the last two granularities. In the fourth granularity, since the point features are relatively descriptive and using an isolated instance volume would violate principle 2 of Volumetric Searching, we apply two instance volumes to perceive the instance. In the final stage, a single global feature is naturally chosen. Overall, by inversely balancing volume number and feature dimension, the reorganized multi-granularity instance features achieve dimensional uniformity.

In conclusion, we introduce a novel approach that uses 3D instance volumes to complement local instance features via Volumetric Pooling (VP). Combined with an adaptive volumetric searching strategy, this design effectively balances feature information across granularities without introducing additional parameters. By uniformly applying VP to all instances in the scene, a robust foundation for multi-granularity scene graph analysis is established.

D. Granularity Transformer

In complex 3D real-world scenes, where instances with cross-order of magnitude size differences coexist, the ability of multi-granularity awareness becomes critical for predicting the two fundamental elements of 3DSSG: instance categories and the relationships between instance pairs (i.e., the *nodes* and *edges*). Therefore, we first focus on instance identification and propose the Granularity Transformer block (GrT), which dynamically shunts the attention of instance features across multiple granularities to better model instances with significant size discrepancies.

Granularity Embedding. Although the multi-granularity instance features organized via VP can capture complex geometric representations, the fundamental granularity attributes of the instance remain unignorable [9]. Therefore, we introduce granularity embedding in GrT to provide a macro perspective, increasing sensitivity in distinguishing instance granularity differences, thereby guiding the subsequent attention shunt. As demonstrated in Figure 4, the 3D bounding box (3D BBox) of an instance offers intuitive granularity attributes, including the geometry center $\mathbf{c} = (x, y, z)$, instance extents $\mathbf{e} = (e_x, e_y, e_z)$, and BBox volume $v = e_x e_y e_z$. However, the volume calculated by BBox, while effectively representing size discrepancies between instances, is coarsely defined by the range of instances, lacking precision and robustness. For example, in the case of a *wall* (e.g., the point cloud masked in green in Figure 2), the 3D BBox encompasses a significant amount of empty space. Therefore, we further introduce the convex hull volume H_v and the convex hull surface H_s to depict granularity attributes more precisely, reducing the deviation of granularity judgment caused by loose envelope. Altogether, the embedding is organized as follows:

$$Gran = \mathrm{MLP}_{ins}\left(\mathbf{c}, \mathbf{e}, log\left(v\right), log\left(H_v\right), log\left(H_s\right)\right) \tag{4}$$

where, both volume v, H_v and surface attribute H_s are log-transformed to mitigate training instability caused by cross-order magnitude differences between instances.

Attention Shunt. Since the reliance on higher resolution feature maps for distinguishing instances, which require relatively fine-grained details [4], the proposed GrT employs the first four granularities of features output by the VP. Taking the instance a as an example, the initial step involves utilizing linear layers to project the input multi-granularity feature sequence $\mathcal{F}^a_{ins} = \{I^a_1, ..., I^a_4\}$ into set Ins^a_i and Ins^a_j. Simultaneously, the instance granularity embedding $Gran^a$ is introduced into both tensors to supplement the basic volumetric attribute, enhancing the perception of GrT to instance size. The next step is to guide the attention shunt based on Ins^a_i. In particular, considering every feature in the set Ins^a_i as a pattern representation of instance a, possessing equivalent value in the multi-granularity analysis, we then diverge from the current vector attention [37] approach, which requires relation calculation with a specific center. Therefore, the mapping function (e.g., a linear layer) is directly performed in each feature $Ins^{a_k}_i$ of Ins^a_i. Subsequently, a normalization operation, such as *softmax*, is applied along the feature channel. With this approach, the attention weights are vectors that can modulate individual feature channels, providing delicate feature shunting. Finally, the value set Ins^a_j is multiplied by the corresponding attention weights while aggregating the ultimate instance feature $\mathbf{y}^a$. The above process can be formulated as:

$$\begin{aligned} &Ins^a_i = \varphi\left(\mathcal{F}^a_{ins}\right) + Gran^a,\ \ Ins^a_j = \psi\left(\mathcal{F}^a_{ins}\right) + Gran^a \\ &\mathbf{y}^a = \sum_{k=1}^{4} \left(\alpha\left(\gamma\left(Ins^{a_k}_i\right)\right) \odot Ins^{a_k}_j\right) \end{aligned} \tag{5}$$

where, the $\varphi(\cdot)$, $\psi(\cdot)$, and $\gamma(\cdot)$ represent the individual mapping functions used in GrT, and the $\alpha(\cdot)$ represents the *softmax* operation.

In summary, by focusing on multi-granularity instance feature analysis with the aid of basic granularity attributes, the GrT dynamically directs fine-grained and coarse-grained features for precise instance identification. Consequently, by applying GrT in parallel to the numerous instances in the 3D scene, our GranSSG demonstrates enhanced capability in understanding the coexistence of multi-size instances.

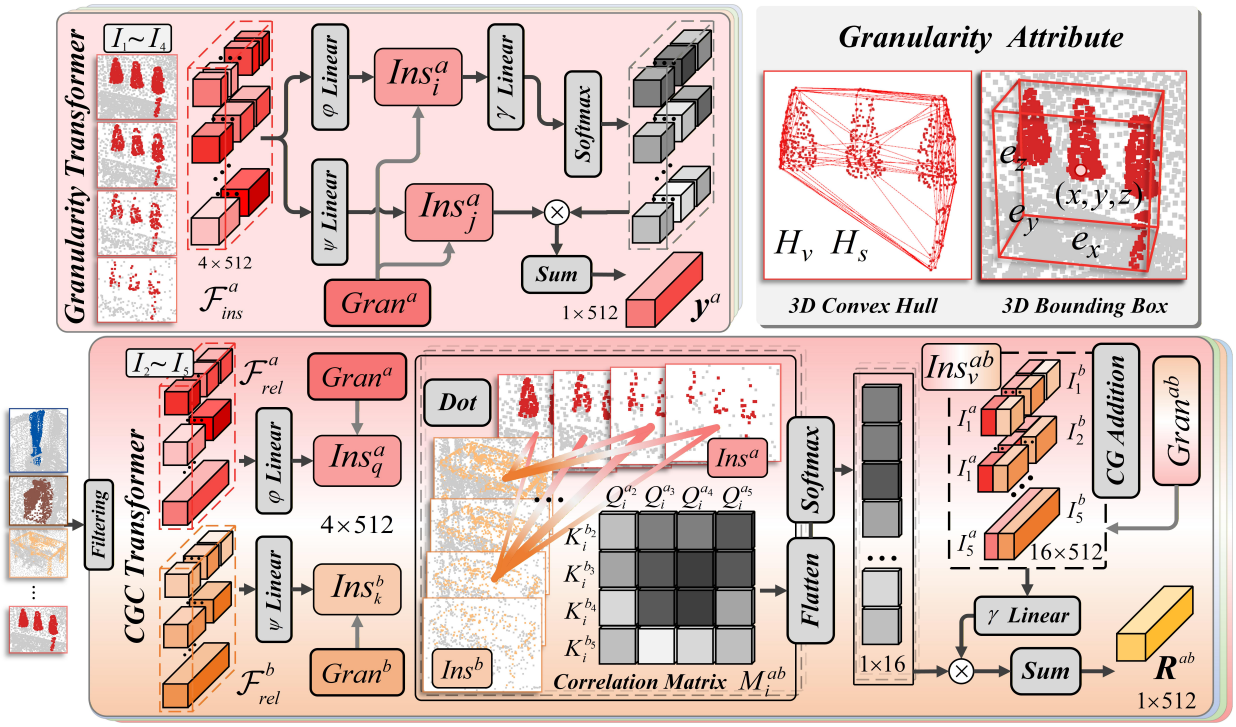


Fig. 4. Illustration of the proposed Granularity Transformer block (GrT, upper left), the Cross-Granularity Correlation Transformer block (CGCT, lower part), and the granularity attribute (upper right). GrT equally considers features at various granularities for a single instance, performing fine-grained attention shunting across feature channels to achieve adaptive weight distribution. The basic granularity attribute message is introduced via embedding, enhancing the sensitivity of GrT to instances of different sizes. CGCT evaluates all possible manifold cross-granularity feature combinations for subject-object instance pairs, assessing and fusing valuable cross-granularity features, thereby achieving a targeted and comprehensive perception of relationships among instance pairs.

E. Cross-Granularity Correlation Transformer

Different from instance identification that considers merely its intrinsic multi-granularity feature, relationship reasoning involves a subject–object pair where the two instances may differ significantly in size. This demands an analysis of cross-granularity feature combinations to accurately perceive their interaction. To address the manifold interplay possibilities of multi-granularity features for both sides while adaptively selecting more valuable granularity combinations, we introduce the Cross-Granularity Correlation Transformer block (CGCT).

Candidate Pair Filtering. Since a scene typically contains numerous instances, performing computations on all potential instance pairs results in an exhaustive pairwise computation. Consequently, we first perform candidate subject-object instance pair filtering. In this process, we flatten the features across all granularities to obtain a complementary representation. Each pair is evaluated using cosine similarity, and a filtering threshold τ is applied based on a Gaussian distribution. Only pairs that exceed the threshold are retained and passed to the CGCT module. Although straightforward, this approach effectively captures the spatial and semantic commonalities between subjects and objects, serving as a highly efficient filter. This process can be described as follows:

$$\begin{aligned} &Rel\left(i,j\right) = cos\left(\mathrm{Flt}\left(\mathcal{F}_{rel}^{i}\right), \mathrm{Flt}\left(\mathcal{F}_{rel}^{j}\right)\right) \\ &\tau = \mu_{rel} + 0.5 \times \sigma_{rel} \\ &\mathcal{P}_{rel} = \left\{(i,j) \,|Rel\left(i,j\right) > \tau, i \neq j\right\} \end{aligned} \tag{6}$$

where $Rel\left(\cdot\right)$ represents instance pair relevance, $\mathrm{Flt}\left(\cdot\right)$ represent the flatten operation, μ_{rel} and σ_{rel} are the mean and standard deviation of relevance scores calculated within each batch, respectively. This allows for an adaptive filtering threshold τ that responds to the specific distribution of instance pairs. And $\mathcal{P}_{rel}$ is the set of kept candidate instance pairs.

Correlation Granularity Embedding. Despite the introduction of instance granularity embedding in GrT, we also employ a specific correlation granularity embedding, which emphasizes the relative bias among volumes of the instance pair, targeted serves the relationship perception. Specifically, the embedding is organized as follows:

$$\begin{aligned} Gran^{ab} = \mathrm{MLP}_{rel}\,(&\mathbf{c}_a - \mathbf{c}_b, \mathbf{e}_a - \mathbf{e}_b, \\ &\log\left(\frac{v_a}{v_b}\right), \log\left(\frac{H_v^a}{H_v^b}\right), \log\left(\frac{H_s^a}{H_s^b}\right)\Big) \end{aligned} \tag{7}$$

Cross-Granularity Correlation Assessment. Considering the necessity for a larger receptive field to accurately judge the relationship of instance pairs [4], we utilize features from the last four granularities in our processing. For a linked instance pair $\langle a, b\rangle$, designated as *subject* and *object*, correlation assessment is the initial step. Specifically, the multi-granularity feature sets $\mathcal{F}_{rel}^{a} = \{I_2^a, ..., I_5^a\}$ and $\mathcal{F}_{rel}^{b} = \{I_2^b, ..., I_5^b\}$ are projected into query $Ins_q^a = \{Q^{a_2}, ..., Q^{a_5}\}$ and key $Ins_k^b = \{K^{b_2}, ..., K^{b_5}\}$ via independent linear layers. Simultaneously, similar to GrT, the relevant instance granularity embeddings $Gran^a$ and $Gran^b$ are injected to represent the relative relationship from a rough yet global perspective.

Following this preparation, CGCT aims to identify suitable cross-granularity combinations that represent the instance pair

well. Unlike the single-instance attention weight in GrT, we posit that the granularity emphasis for the same instance can vary in different instance pairs. For example, deeper scales with a broader receptive field might better determine the relationship between *table* and *walls*, whereas shallower scales may better capture the interaction details when *table* involves a small *lamp*. Moreover, CGCT does not restrict the features to be at the same granularity. Instead, we encourage cross-granularity feature interaction. As shown in Figure 4, the features of subject-object instances interact freely, creating 4×4 potential combinations. This endows CGCT with a comprehensive cross-granularity perceptive capability.

In practice, CGCT splits Ins_q^a and Ins_k^b into h independent heads to assess the correlation of cross-scale combinations in parallel. For each head, we perform a dot-product operation on Q_i^a and K_i^b. While straightforward, each element $Q_i^{a_m} \cdot \left(K_i^{b_n}\right)^T$ effectively represents the similarity of a feature pair (i.e., $\langle I_m^a, I_n^b \rangle$) in semantic space. Further, the constructed correlation matrix comprehensively covers the assessment of all potential combinations, adaptively guiding the manifold cross-scale feature fusion. Subsequently, we apply the *softmax* function to normalize the correlation matrix, obtaining the attention weight value. Altogether, the cross-granularity feature assessment in CGCT is calculated by:

$$\begin{aligned} &Ins_q^a = \varphi\left(\mathcal{F}_{rel}^a\right) + Gran^a, \;\; Ins_k^b = \psi\left(\mathcal{F}_{rel}^b\right) + Gran^b \\ &M_i^{ab} = \text{Softmax}\left(\frac{Q_i^a \cdot \left(K_i^b\right)^T}{\sqrt{d_h}}\right) \end{aligned} \tag{8}$$

where, the $\varphi(\cdot)$ and $\psi(\cdot)$ represent the mapping function used in CGCT, M_i^{ab} represents the correlation matrix of the instance pair $\langle a, b \rangle$ in the i-th head, and $\sqrt{d_h}$ is a scaling factor.

Cross-Granularity Feature Fusion. We then aim to construct the value Ins_v^{ab} that describes the relationship of instance pairs under hybrid granularities. First, permutation is applied to the multi-granularity features of $\mathcal{F}_{rel}^a$ and $\mathcal{F}_{rel}^b$ while keeping the combination scheme consistent with the correlation matrix for uniformity. Next, each cross-granularity feature pair $\langle I_m^a, I_n^b \rangle$ is combined element-wise, denoted as I_{m-n}^{ab}. In this manner, the cross-granularity feature set $\mathcal{F}_{rel}^{ab} = \left\{I_{2-2}^{ab}, ..., I_{5-5}^{ab}\right\}$ is organized and can be projected into Ins_v^{ab} through a linear layer. Notably, instead of employing single-instance granularity embedding, the correlation granularity embedding $Gran^{ab}$ is then injected into Ins_v^{ab}, which focuses more on the discrepancy between subject and object, better reflecting their relationship. Since multi-head attention is applied, the feature channel of V^{ab} is then h-fold split, enabling more detailed cross-granularity fusion. Finally, for each head, the flattened correlation matrix controls the attention distribution of cross-granularity feature fusion, leading to a comprehensive perception of instance pair relationships. The mathematical expression can be represented as follows:

$$\begin{aligned} &\mathcal{F}_{rel}^{ab} = \left\{\text{Add}\left(I_m^a, I_n^b\right) | \forall I_m^a \in \mathcal{F}_{rel}^a, \forall I_n^b \in \mathcal{F}_{rel}^b\right\} \\ &Ins_v^{ab} = \gamma\left(\mathcal{F}_{rel}^{ab}\right) + Gran^{ab} \\ &\mathbf{R}_i^{ab} = \sum_{k=1}^{4\times4}\left(\text{Flt}\left(M_i^{ab_k}\right) \cdot V_i^{ab_k}\right) \end{aligned} \tag{9}$$

where, $\gamma(\cdot)$ represents a mapping function, $\text{Add}(\cdot)$ and $\text{Flt}(\cdot)$ represent the element-wise addition and the flatten operation. $\mathbf{R}_i^{ab}$ denotes the i-th head of the ultimate relation feature $\mathbf{R}^{ab}$, while $M_i^{ab_k}$ and $V_i^{ab_k}$ represent the attention weight and the value for the k-th cross-granularity combination in head i.

Overall, the proposed CGCT facilitates complex cross-granularity analysis. Moreover, by extending CGCT to every potential instance pair within the scene, all edges are comprehensively modeled. In summary, the introduction of CGCT, together with GrT, enables our GranSSG to perform the essential multi-granularity analysis for both fundamental elements of 3DSSG: single instances (*nodes*) and relationships between instance pairs (*edges*), particularly in environments with significant discrepancies in instance sizes.

F. The Training Objective

As the features of *nodes* and *edges* are extracted separately, we utilize two MLP layers as classifiers, applying a joint loss $\mathcal{L}_{scene}$ to train the entire GranSSG. The training objective of the network is defined as follows:

$$\mathcal{L}_{scene} = \alpha_{ins}\mathcal{L}_{ins} + \alpha_{rel}\mathcal{L}_{rel} \tag{10}$$

where α_{ins} and α_{rel} are the respective weighting factors, set to 0.3 and 1, respectively. As defined in [2], a linked instance pair can exhibit multiple relationships simultaneously, with relationship prediction supervised by per-class binary cross-entropy. The instance identification is supervised by multi-class cross-entropy to ensure the uniqueness of categories.

IV. Experiment

In this section, we perform intensive experiments in the challenge and large-scale 3D point cloud semantic scene graph prediction dataset, 3DSSG [2], to validate the competitive performance of our proposed GranSSG. Furthermore, we perform extensive ablation experiments to evaluate the benefits of each component within our methods in detail.

A. Setups and Implementation Details

Dataset: We conduct our experiment on the large-scale 3DSSG dataset [2] to validate the effectiveness of the proposed method. Building upon the 3RScan dataset [38], the 3DSSG dataset encompasses 1553 real-world indoor point cloud scenes, where each point contains spatial coordinates, RGB information, and a class-agnostic instance mask to identify instances. Moreover, the 3DSSG dataset provides detailed semantic scene graph annotations for the understanding of complex everyday environments, including instance labels from 160 semantic categories and 26 types of relationships to represent the potential connections of linked instance pairs. We follow the same training/validation dataset splitting setting applied in [2] when evaluating our proposed methods.

Evaluation Metrics: Following the experiment setting in [4], [5], we choose the top-k accuracy as the main metric. Specifically, while evaluating the two basic elements: category of instances (*nodes*) and relationship of instance pairs (*edges*), the Object A@k and Predicate A@k are respectively performed, as defined in [5]. This allows us to finely evaluate

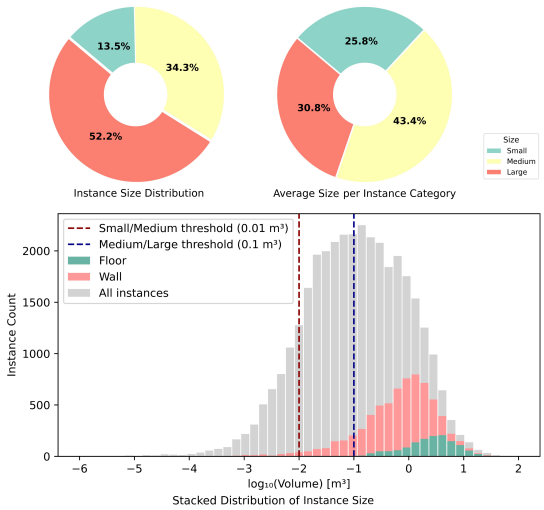


Fig. 5. Statistical analysis of instance size distribution in 3DSSG [2] dataset. Left: Proportion of small, medium, and large instances. Right: Average object size per category. Bottom: Log-scale histogram showing the overall volume distribution (gray), with stacked overlays for frequent large-size classes.

the performance of two multi-granularity analyses in our GranSSG. Meanwhile, the Triplet A@k is further applied to comprehensively represent the prediction effect. Conducting the methodology from [5], for a relation triplet <*subject*, *predicate*, *object*>, the predicting scores of *subject*, *predicate*, and *object* are multiplied first, subsequently ranked in order. As a strict standard, the relation triplet is ultimately considered correct only if all components are accurately predicted.

Additionally, as intuitively demonstrated in Figure 5, the instance volumes in the 3DSSG dataset span multiple orders of magnitude. To more rigorously evaluate model performance across varying object sizes, we introduce size-related metrics, denoted as A@*k-s* (*small*), A@*k-m* (*medium*), and A@*k-l* (*large*). Taking the convex hull volume of an instance as standard, instances are classified as small, medium, and large. Specifically, as shown in the stacked histogram of Figure 5, to maintain representative separation while avoiding overrepresentation of *wall* and *floor* in *large* categories, we choose 0.01 m^3 and 0.1 m^3 as thresholds. The right-side pie chart of Figure 5 further confirms that each volume range covers a well-balanced variety of instance categories, achieving a semantically meaningful distribution.

Notably, when calculating A@k accuracy for Predicate and Triplet in each size range, a subject-object pair is included if either instance meets the size threshold, thereby ensuring comprehensive statistics. Different from the traditional metrics, the introduced size-related metrics then present the effectiveness of methods in a clear and rigorous way, highlighting the vital importance of granularity-awareness in the 3DSSG prediction.

Furthermore, we report the performance of methods in two tasks that focus on the scene graph as defined by Zhang et al. [39]: (1) Scene Graph Classification (SGCls) and (2) Predicate Classification (PredCls). The SGCls task comprehensively considers all relation triplets in the scene, while the PredCls task assumes the correction of instance identification and solely assesses the prediction of instance pair relationships. Following the definition in Zhang et al. [39], the recall at the top-k (R@k) triplets is considered as the metric for these tasks.

Implementation details: The model is optimized using AdamW for 35 epochs with a batch size of 8. The initial learning rate is 0.001 with a weight decay of 1×10^{-4}, adjusted by a StepLR scheduler (step size 10, $\gamma = 0.1$). We apply standard point cloud augmentations, including random scaling, coordinate jittering, and chromatic jittering. As pointed out in [5], since the view-dependent spatial relationships strongly correlate with the coordinate system, 3D scenes are standardized in the same coordinate, leading to unambiguous spatial relation predicates. Our experiments are conducted on a computational platform equipped with Intel(R) Xeon(R) CPU E5-2680v3 CPU x2, and RTX 3090 GPU x8. All models are trained using a single RTX 3090 GPU.

B. Evaluation on 3DSSG dataset

To verify the performance of our proposed GranSSG, we first compare our method with other leading approaches in the 3DSSG validation set. Through the evaluation conducted on standard object/predicate/triplet and size-related metrics, the effectiveness of GranSSG is comprehensively exhibited.

As shown in Table II, GranSSG exhibits superiority over existing methods by a large margin. Considering the two types of 3DSSG prediction: 1) Predicting the whole scene graph simultaneously, which is the original state of our GranSSG, and 2) Predicting a single relation triplet at a time, which explicitly isolates individual subject-object pairs and their local vicinities to suppress background interference from irrelevant objects, marked with *. Notably, in the common condition, our GranSSG outperforms the current state-of-the-art method VL-SAT across all dimensions, achieving significant improvements of +7.48, +2.3, and +2.29 in Object A@1, Predicate A@1, and Triplet A@50, respectively. Although VL-SAT achieves additional gains via incorporating RGB information, natural language representation, and the CLIP model [13], the superior performance of GranSSG directly illustrates the significance of granularity-awareness ability. Specifically, GranSSG exhibits dominant improvement in identifying instance categories, breaking through 60% and 80% in Object A@1/5 for the first time. Meanwhile, in terms of predicate prediction, our method shows remarkable performance. Even though GranSSG slightly lags in the Predicate A@1 metric compared to SGGpoint by -0.29, it still achieves +0.73 and +0.55 at A@3/5 respectively. When distributed prediction relation triples, the competitive results of +1.07 and +1.38 in Object A@1 and Predicate A@1 over the current Granular3D consistently demonstrate the effectiveness of our GranSSG. Overall, these comparisons highlight the advantage of providing multi-granularity analysis in real-world environments where instances with significant size discrepancies coexist.

On the other hand, our proposed size-related metrics evaluate the algorithm in a more rigorous and detailed manner, as shown in Table II. Notably, in instance recognition, all evaluated methods significantly underperform in identifying smaller instances compared to medium or large instances. This difficulty underscores the importance of granularity-awareness in understanding complex environments. Specifically, compared to VL-SAT, GranSSG boosts the overall performance

TABLE II
QUANTITATIVE RESULTS (%) ON 3DSSG [2] VALIDATION SET. THE **BOLD** DENOTES THE BEST PERFORMANCE. THE MODEL WITH ∗ PREDICTS A SINGLE RELATION TRIPLET AT A TIME.

Model	Type	Object						Predicate						Triplet				
		A@1/5/10			A@1-s/m/l			A@1/3/5			A@1-s/m/l			A@50	A@100	A@50-s/m/l		
SGPN [2]	Point cloud	48.28	72.94	82.74	-	-	-	91.32	98.09	99.15	-	-	-	87.55	90.66	-	-	-
SGG_{point} [8]	Point cloud	51.42	74.56	84.15	-	-	-	**92.40**	97.78	98.92	-	-	-	87.89	90.16	-	-	-
SGFN [9]	RGB-D sequence	53.67	77.18	85.14	25.25	43.44	67.13	90.19	98.17	99.33	89.28	87.99	91.18	89.02	91.71	87.55	87.60	91.11
SGRec3D [40]	Point cloud	-	80	87	-	-	-	-	97	99	-	-	-	89	91	-	-	-
VL-SAT [5]	Point cloud	55.66	78.66	85.91	26.12	48.02	72.02	89.81	98.45	**99.53**	90.44	89.34	91.69	90.35	92.89	88.30	88.42	92.01
GranSSG	Point cloud	**63.14**	**83.89**	**89.45**	**34.00**	**55.34**	**76.45**	92.11	**98.51**	99.47	**92.48**	**91.26**	**93.24**	**92.64**	**94.64**	**90.74**	**91.99**	**94.35**
Open3DSG* [12]	RGB sequence	-	57	68	-	-	-	-	63	70	-	-	-	64	66	-	-	-
Granular3D* [4]	Point cloud	66.02	86.13	91.60	34.61	56.85	79.62	91.29	98.35	99.45	91.94	90.54	92.85	93.15	95.13	91.33	92.97	94.75
GranSSG*	Point cloud	**67.09**	**87.35**	**91.99**	**35.17**	**58.48**	**80.82**	**92.67**	**98.56**	**99.50**	**93.09**	**91.81**	**93.86**	**93.49**	**95.40**	**91.44**	**93.20**	**95.04**

TABLE III
QUANTITATIVE RESULTS (%) OF THE SGCLS AND PREDCLS TASKS, WITH AND WITHOUT GRAPH CONSTRAINTS. THE **BOLD** DENOTES THE BEST PERFORMANCE. THE MODEL WITH ∗ PREDICTS A SINGLE RELATION TRIPLET AT A TIME.

Model	Type	with Graph Constraints						without Graph Constraints					
		SGCls			PredCls			SGCls			PredCls		
		R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
Co-Occurrence [39]	Point cloud	14.8	19.7	19.9	34.7	47.4	47.9	14.1	20.2	25.8	35.1	55.6	70.6
KERN [41]	RGB sequence	20.3	22.4	22.7	46.8	55.7	56.5	20.8	24.7	27.6	48.3	64.8	77.2
SGPN [2]	Point cloud	27.0	28.8	29.0	51.9	58.0	58.5	28.2	32.6	35.3	54.5	70.1	82.4
Schemata [42]	RGB sequence	27.4	29.2	29.4	48.7	58.2	59.1	28.8	33.5	36.3	49.6	67.1	80.2
Zhang et al. [39]	Point cloud	28.5	30.0	30.1	59.3	65.0	65.3	29.8	34.3	37.0	62.2	78.4	88.3
3D-HetSGP [43]	Point cloud	28.5	29.8	29.9	61.8	65.9	65.9	-	-	-	-	-	-
SMKA [44]	Point cloud	-	31.5	31.6	-	68.3	69.5	-	-	-	-	-	-
SGFN [9]	Point cloud	29.5	31.2	31.2	65.9	78.8	79.6	31.9	39.3	45.0	68.9	82.8	91.2
VL-SAT [5]	Point cloud	32.0	33.5	33.7	67.8	79.9	80.8	33.8	41.3	47.0	**70.5**	85.0	92.5
MonoSSG [45]	RGB sequence	-	35.2	36.1	-	80.8	81.0	-	-	-	-	-	-
HE-3DGSR [46]	RGB-D sequence	-	35.1	35.2	-	81.3	81.6	-	-	-	-	-	-
HEDSGP [47]	RGB sequence	-	37.5	38.2	-	**83.3**	**83.6**	-	-	-	-	-	-
USG-Par [48]	Point cloud	36.6	41.4	46.2	71.9	81.0	83.4	-	-	-	-	-	-
CCL-3DSGG [3]	Point cloud + Images	37.6	40.3	45.7	73.6	80.5	82.9	-	-	-	-	-	-
GranSSG	Point cloud	**41.5**	**51.4**	**57.3**	**73.7**	80.2	80.3	**39.9**	**49.5**	**57.0**	70.4	**86.8**	**92.5**
Granular3D* [4]	Point cloud	43.4	53.0	59.2	73.2	79.4	79.5	41.7	51.8	59.2	69.5	86.1	91.3
GranSSG*	Point cloud	**45.9**	**56.0**	**61.9**	**75.2**	**81.3**	**81.5**	**43.6**	**54.6**	**61.6**	**80.0**	**88.2**	**92.7**

by +2.44/+3.57/+2.34 in Triplet A@1-s/m/l, while achieving a significant improvement of +7.88/+7.32/+4.43 in Object A@1-s/m/l. The greater improvement in identifying smaller objects further demonstrates the effectiveness of our method. Meanwhile, compared to Granular3D, our GranSSG also leads to notable performance gains in instance and instance pair perception, with +0.56/+1.63/+1.2 in Object A@1-s/m/l, and +1.15/+1.27/+1.01 in Predicate A@1-s/m/l. These results further showcase that granularity-awareness ability is a prerequisite in the 3DSSG prediction task.

For a more detailed and comprehensive evaluation of our proposed method, we additionally report results on the Scene Graph Classification (SGCls) and Predicate Classification (PredCls) tasks, following the protocol in [39]. As shown in Table III, GranSSG significantly outperforms existing point cloud-based and multi-modal approaches on the SGCls task, which strictly measures overall scene graph generation performance [5]. Specifically, GranSSG achieves improvements of +4.9/+10.0/+11.1 in R@20/50/100 in compare with USG-Par. On the PredCls task, which focuses solely on the predicate, our method also achieves comparable performance. It is worth noting that some recent methods benefit from pretrained priors, such as CLIP [13] in CCL-3DSGG, and Point-BERT [49] in USG-Par, which provide additional performance gains. Meanwhile, methods based on RGB image sequences, such as HEDSGP, perform well in PredCls. This may be due to their ability to make temporally consistent filtering and prediction of instance-pair relationships across frames. Overall, the results subsequently confirm the effectiveness of our method, which benefits from adaptive multi-granularity feature analysis.

C. Qualitative Results

In this part, we present qualitative results of the GranSSG, demonstrating its 3DSSG prediction capabilities and the performance of the proposed GrT and CGCT methods on instances of various sizes.

To comprehensively demonstrate the performance of GranSSG in complex real-life scenes, we selected several typical indoor environments as demonstrated in Figure 6, including a kitchen (left part in 1st row), a laundry room (right part in 1st row), a bathroom (2nd row), and a living room (3rd row). From each scene, the significant size discrepancies between instances are clearly observable, and we use circles of different sizes to mark and distinguish them. Overall, the results depicted in Figure 6 exhibit that our GranSSG maintains excellent scene graph prediction capabilities across various scenarios, highlighting compatibility and stability. Simultaneously, the accurate predictions for objects with substantial size discrepancies, such as the *item* and *cabinet* in the kitchen, and the *toilet paper* and *counter* in the bathroom, reflect the effectiveness of our method in granularity adaptation. However, in the living room, the confusion between *suitcase*, *bag*, and *box* suggests a limitation in distinguishing objects with similar shapes. This highlights the limitation of multi-granularity point cloud analysis when texture cues are essential.

For the GrT block, designed to integrate multi-granularity features to achieve adaptive instance category identification, we comprehensively report the attention distribution across

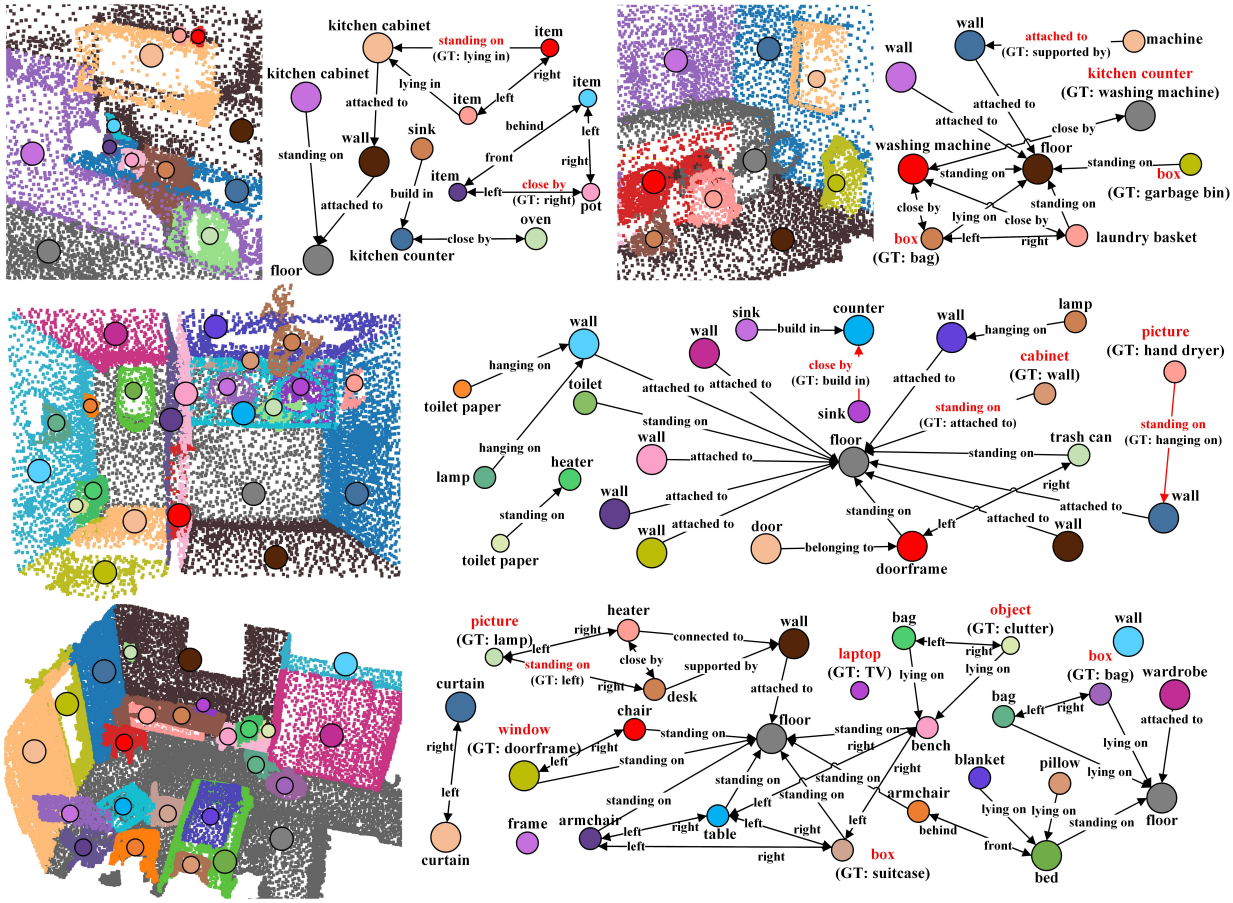


Fig. 6. Visualization of some typical 3D semantic scene graph prediction results on the 3DSSG validation set. The small colored circles with various sizes represent the instances with significant size discrepancies. And the misclassified instances or relationships in the scene graph are marked with red.

all instance categories in the 3DSSG validation dataset. As depicted in Figure 7, after sorting, the categories on the left exhibit a focus on shallow-granularity point features, while those on the right predominantly absorb deep-granularity features. In this context, the average convex hull volume of each category is further overlaid. The line representing the size of various instances illustrates a clear upward trend. This supports the underlying theory that larger instances benefit from deep-granularity features with larger receptive fields, while smaller instances rely more on shallow-granularity features that preserve local detail. This trend is both logical and well-validated in the context of 3DSSG prediction for complex scenes. However, we also observe a few exceptions. For example, attention distributions for large instances like *blinds* and small ones like *sockets* appear relatively uniform. This may be due to insufficient training samples, leading to underfitting in these categories. Meanwhile, although the general trend aligns with our expectations, the relationship between instance volume and attention distribution is inherently non-linear. This complexity may be attributed to the influence of structural properties and the varying degrees of contextual support provided by the surrounding environment during the recognition of different instances. In conclusion, by leveraging attention shunts through the GrT module, our method dynamically routes the most appropriate multi-granularity features to each instance, contributing to enhanced recognition performance.

Moreover, we selected representative instance pairs to cover interactions between subject-object instances of varying sizes, including the *book-folder* pair for small instances, the *desk-chair* pair for mid-sized instances, and the *bed-object* pair representing instances with significant size discrepancies. The results demonstrate the assessment of cross-granularity feature correlations by the proposed CGCT method. As illustrated in Figure 8, different instances exhibit varying but traceable tendencies when dealing with combinations of subject-object instance scale features. For example, small instances (e.g., *book*, *lamp*) typically choose shallow-granularity features when interacting with other objects. Meanwhile, if their interacting instance is large, such as in the *book-sofa* pair, deep-granularity features are also utilized. When interactions involve medium and large instances, such as the *desk-wall* pair, deeper-granularity features dominate. However, we also observe that the attention to the deepest granularity combinations (i.e., the lower-right corner of the attention map) is not particularly strong. This may be due to overly large receptive fields introducing irrelevant environmental information. Despite these exceptions, the overall trend in the experimental results clearly demonstrates that our proposed CGCT block effectively adapts to the diverse interactions across granularities. By fusing the most informative feature pairs, it significantly improves relationship perception.

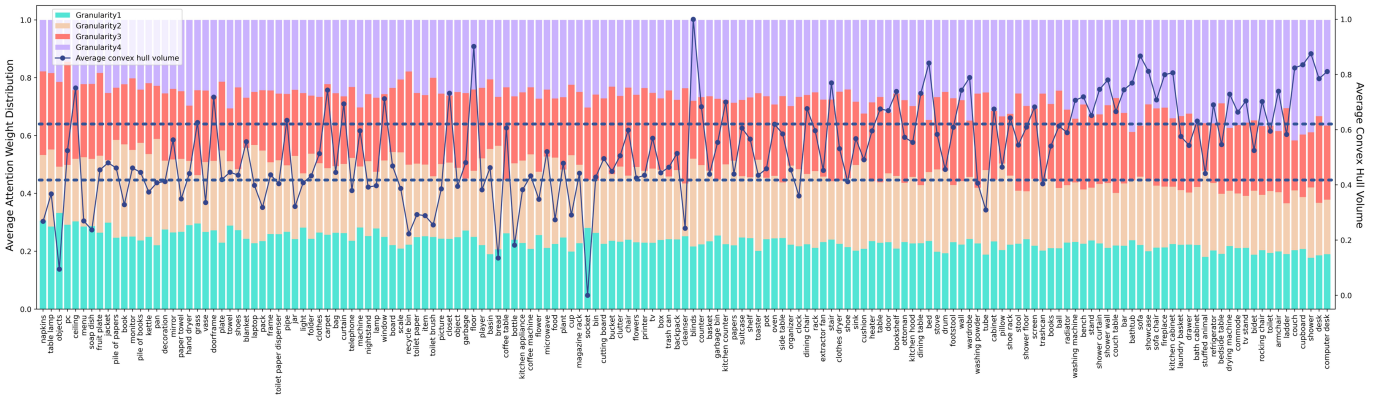


Fig. 7. Visualization of the correlation between instance size and attention shunt in the GrT block. Along the x-axis, we analyze all 160 categories of instances in the 3DSSG dataset. Each color in the stacked bar chart indicates the average attention weight distributed on a granularity. The line chart shows the normalized average hull volumes across all instances, while the two dashed lines represent the volume thresholds for small, medium, and large instances.

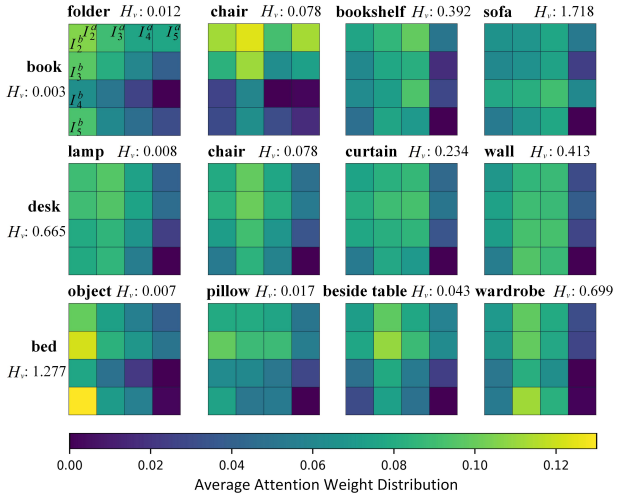


Fig. 8. Visualization of the average cross-granularity correlations of typical instance pairs in the CGCT block. From the top left to the bottom right of each matrix, the granularity features of the subject-object instance interactions deepen. The brighter areas indicate more valuable cross-granularity feature interplay assessed by the CGCT block, with the reported H_v representing the average hull volume of the instance category.

D. Ablation study

For a more detailed assessment of the proposed GranSSG, we conducted extensive ablation studies on the 3DSSG validation set. These experiments comprehensively cover the various components within GranSSG and demonstrate the effectiveness of our methods.

Volumetric Pooling: We initially focus on the ablation study of our proposed Volumetric Pooling block (VP), which aims to balance the information carried by instance features across different granularities and is a crucial component of the subsequent multi-granularity analysis. In this ablation study, we compare our VP with some traditional strategies that extend features at various granularities to an identical size: 1) 'L+P': Extend point features with a linear layer, then apply pooling. 2) 'P+L': Pool instance point features, then apply a linear layer. 3) 'P+C': Pool features, then replicate and concatenate them to match the target size (a non-learning approach). As shown in Table IV, our VP significantly outperforms all baselines, with +0.66/+0.61/+0.76 gains in Triplet A@50 respectively, which illustrates the effectiveness of through information balancing. Moreover, the 'P+C' strategy, a simple non-parametric approach, does not perform significantly worse than methods that use a linear layer for feature expansion (only a -0.1 drop in Triplet A@50 compared to the 'L+P' method). This further demonstrates the difficulty of altering the information carried in the aggregated instance feature. Additionally, the proposed VP demonstrates stronger advantages in recognizing small instances (+2.81 in Object A@1-s compared to the 'L+P' method), confirming its ability to retain fine-grained details from shallower representations. These results substantiate the value of VP in enabling robust, granularity-aware feature representation, especially benefiting small and challenging objects.

TABLE IV
ABLATION STUDY OF THE VOLUMETRIC POOLING BLOCK (%).

Strategy	Object			Predicate			Triplet		
	A@1	A@1-s/l		A@1	A@1-s/l		A@50	A@50-s/l	
L+P	61.95	31.19	75.90	91.38	92.22	92.60	91.98	89.95	93.94
P+L	62.68	32.81	76.26	91.77	92.64	93.17	92.03	90.02	93.77
P+C	62.00	30.57	76.39	91.83	**92.72**	93.20	91.88	90.14	93.75
VP (Ours)	**63.14**	**34.00**	**76.45**	**92.11**	92.48	**93.24**	**92.64**	**90.74**	**94.35**

TABLE V
ABLATION STUDY OF THE HYPERPARAMETER SETTING FOR THE VOLUMETRIC POOLING BLOCK (%).

Strategy	Object		Predicate		Triplet				
	A@1	A@5	A@1	A@3	A@50	A@100	A@50-s/m/l		
1, [7, 3, 2, 0, 0]	62.88	83.67	91.52	98.35	92.41	94.30	90.27	91.85	93.99
1, [15, 7, 3, 2, 0]	**63.14**	83.89	**92.11**	**98.51**	**92.64**	**94.64**	**90.74**	**91.99**	94.35
1, [31, 15, 7, 3, 2]	63.00	83.93	92.00	98.39	92.45	94.50	90.45	91.57	**94.40**
0, [16, 8, 4, 2, 0]	62.70	**84.09**	90.95	98.29	92.36	94.59	90.03	91.24	94.16

Additionally, we also provide an ablation study of the hyperparameter setting of our VP block. To accommodate different instance volume settings, we added corresponding MLP layers to ensure the final output dimension of instance feature is 512. As shown in Table V, our proposed method achieves the optimal performance without requiring additional dimension alignment measures. Furthermore, it can be observed that the performance degradation caused by excluding global features is the most significant, demonstrating the effectiveness of combining local volume features with global instance awareness.

Meanwhile, we further introduce an innovative experiment

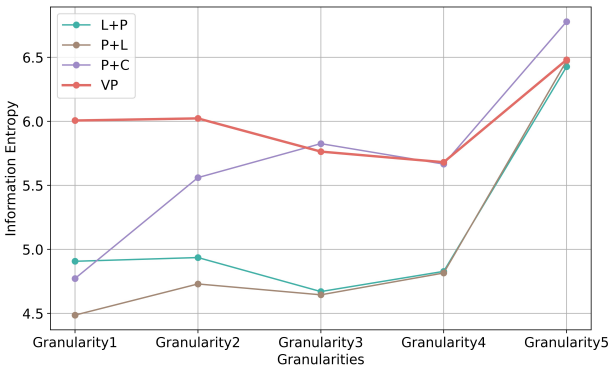


Fig. 9. Illustration of the average information entropy of instance features at various granularities under different pooling strategies.

that measures the average information entropy of instance features at each granularity, providing an angle to observe the information richness. The results are presented in Figure 9. Since the 'P+C' strategy merely replicates the original point features, it can be considered a baseline for information carried at each granularity. It is evident that the average information entropy for shallower-granularity features is less than that for deeper-granularity features. Simultaneously, methods employing linear layers, 'L+P' and 'P+L', do not exhibit high information entropy from Granularity1 to Granularity4, even lower than the 'P+C' strategy. This implies that the network may prefer to filter and retain key semantic information. In contrast, the information entropy of features at all granularities in our proposed VP does not fluctuate significantly. Specifically, at Granularity1, there is a +26% increase in information entropy compared to the 'P+C' strategy. These continuously demonstrates the beneficial effect of our method in balancing features across granularities, thereby laying a solid foundation for subsequent adaptive multi-granularity analysis.

Granularity Transformer: We then study the adaptive attention fusion of multi-granularity instance features facilitated by the proposed Granularity Transformer (GrT) block, which aims to enhance the understanding of the coexistence of multi-size instances. In this part, we compare our GrT with several common strategies. In Table VI, 'Concatenation' refers to a straightforward baseline method that simply concatenates features from all granularities for prediction without any sophisticated fusion mechanism. 'Single granularity' refers to selecting instance features from a specific scale, while 'Pooling granularities' involves fusing multi-granularity features through direct pooling. We also include comparisons with the widely used 'Soft attention' and multi-head self-attention 'MHSA' methods. The results indicate that the 'Concatenation', 'Single granularity' and 'Pooling granularities' methods, which lack the ability to shunt multi-granularity feature attention, exhibit performance declines of -1.21, -0.76 and -0.58 in Triplet A@50, compared to our GrT block. Furthermore, due to the difficulty in finely modeling individual channels of multi-granularity instance features, the 'Soft attention' and 'MHSA' methods also failed to achieve competitive results, particularly in instance identification (-1.42 and -0.83 in Object A@1). Overall, this study highlights the importance of our GrT block in granularity-aware instance identification through dynamic and precise feature shunting.

TABLE VI
ABLATION STUDY OF THE GRANULARITY TRANSFORMER BLOCK (%).

Strategy	Object		Predicate		Triplet				
	A@1	A@5	A@1	A@3	A@50	A@100	A@50-s/m/l		
Concatenation	61.42	81.56	91.26	98.24	91.43	93.79	89.65	90.62	93.37
Single granularity	62.42	83.72	91.99	98.36	91.88	94.01	90.17	91.44	93.84
Pooling granularities	62.27	**84.18**	91.85	98.38	92.06	94.29	90.07	91.67	93.91
Soft attention	61.72	83.12	91.74	98.44	91.82	94.06	90.02	91.35	93.53
MHSA	62.31	83.92	91.92	98.47	92.06	94.26	90.33	91.20	93.76
GrT (Ours)	**63.14**	83.89	**92.11**	**98.51**	**92.64**	**94.64**	**90.74**	**91.99**	**94.35**

TABLE VII
ABLATION STUDY OF THE CROSS-GRANULARITY CORRELATION TRANSFORMER BLOCK (%).

Strategy	Object		Predicate		Triplet				
	A@1	A@5	A@1	A@3	A@50	A@100	A@50-s/m/l		
Concatenation	63.05	83.12	83.15	94.88	90.54	93.03	88.51	89.72	92.73
W-GrT-M	62.52	83.33	91.34	98.32	91.78	93.99	89.81	91.22	93.59
W-GrT-F	62.55	83.67	91.45	98.36	92.16	94.30	89.91	91.68	94.05
CGCT-M	62.50	**83.99**	91.54	98.31	92.10	94.36	90.04	91.73	93.86
CGCT-F (Ours)	**63.14**	83.89	**92.11**	**98.51**	**92.64**	**94.64**	**90.74**	**91.99**	**94.35**

Cross-Granularity Correlation Transformer: We subsequently explore the Cross-Granularity Correlation Transformer (CGCT) block, designed to enhance the perception of subject-object instance pairs. The core of the CGCT block lies in interactively assessing cross-granularity features for each instance pair and adaptively fusing all potential hybrid combinations. Therefore, we first establish a baseline named 'Concatenation,' which aggregates features across all subject and object granularities through simple concatenation followed by element-wise addition. Besides, we devised comparison strategies that inherit the weight attention distribution from the GrT module, thereby excluding feature interaction between instance pair, denoted as 'W-GrT-M' and 'W-GrT-F'. The 'W-GrT-M' strategy considers only the most valuable feature combination, while 'W-GrT-F' considers all combinations. Similarly, we introduced a strategy where the CGCT block considers only the most correlated features, denoted as 'CGCT-M'. As illustrated in Table VII, the performance of 'Concatenation' is significantly inferior, underscoring the critical importance of selective information filtering. Meanwhile, compared to 'W-GrT-M' and 'W-GrT-F', our proposed CGCT block gains an improvement of +0.86 and +0.48 in Triplet A@50, respectively. This highlights the importance of targeted analysis of hybrid cross-granularity features for each instance pair. Meanwhile, after comprehensively fusing the cross-granularity features of instance pairs, the 'CGCT-F' further boosts the Predicate A@1 performance by +0.57 compared to 'CGCT-M'. Similarly, 'W-GrT-F' is better than 'W-GrT-M' as well. In conclusion, these results underscore that the proposed CGCT block excels in managing the complex manifold of cross-granularity awareness for instance pairs by establishing comprehensive feature interactions between subject-object instances in the scene graph.

Granularity Embedding: As a vital component for both GrT and CGCT block, granularity embeddings are crucial in supplementing the basic granularity attributes of instances, while enhancing the spatial relationships between instance pairs. To study the impact of these embeddings within our GranSSG, the ablation study separately removed the 3D

TABLE VIII
ABLATION STUDY OF THE GRANULARITY EMBEDDING (%).

Strategy	Object		Predicate		Triplet				
	A@1	A@5	A@1	A@3	A@50	A@100	A@50-s/m/l		
No embedding	60.87	82.23	81.93	93.71	89.27	91.91	89.09	90.43	92.90
Only embedding	45.82	69.59	90.64	97.85	85.76	89.08	85.11	84.13	88.76
Without hull	61.93	83.46	91.71	98.44	91.98	94.29	90.27	91.68	93.96
Ours	**63.14**	**83.89**	**92.11**	**98.51**	**92.64**	**94.64**	**90.74**	**91.99**	**94.35**

TABLE IX
ABLATION STUDY OF THE GRANULARITY DISTRIBUTION (%). THE MODEL WITH † INTRODUCES ADDITIONAL GRANULARITY-SPECIFIC EMBEDDING.

Strategy	Object		Predicate		Triplet				
	A@1	A@5	A@1	A@3	A@50	A@100	A@50-s/m/l		
[3;5]	61.04	82.74	91.86	98.32	91.68	94.05	90.13	91.41	93.55
[123; 345]	61.83	83.10	91.39	98.39	92.02	94.20	90.35	91.60	93.73
[1234; 2345]	63.14	**83.89**	**92.11**	**98.51**	**92.64**	94.64	**90.74**	**91.99**	**94.35**
[12345; 12345]	62.84	83.35	91.54	98.32	92.00	94.32	90.03	91.60	93.91
[2345; 1234]	62.42	83.54	91.08	98.24	92.08	94.29	89.80	91.46	93.90
[123; 345]†	61.75	83.30	91.56	98.17	92.08	94.12	90.28	91.46	93.80
[1234; 2345]†	**63.18**	83.73	92.08	98.46	92.46	**94.65**	90.19	91.78	94.21
[12345; 12345]†	62.69	83.61	91.63	98.30	92.20	94.30	90.20	91.47	93.94

convex hull attributes and the entire embedding component. As displayed in Table VIII, in general, the application of granularity embedding significantly raises performance with notable improvements of +2.27, +10.18, and +3.37 in Object A@1, Predicate A@1, and Triplet A@50, which illustrates the effectiveness of enriching the intuitive volumetric information, especially in recognizing the relationship of instance pair. Additionally, excluding the convex hull features results in a -0.66 drop in Triplet A@50, suggesting the positive impact of more precise volumetric representation on granularity-awareness. Besides, we conducted an interesting experiment where predictions were made using only the attribute information. Surprisingly, it still maintained a certain level of object classification and relationship prediction capability, highlighting the importance of granularity attribute embeddings.

Granularity Distribution: We then report the ablation study of the multi-granularity features employed in our GrT and CGCT blocks, reflecting the impact of the granularity of point cloud features on scene graph prediction. Numbers before ';' denote the features used in GrT, while the subsequent numbers denote the features applied in CGCT. Beginning with a single granularity, we gradually increased the features perceived by both modules. The results shown in Table IX illustrate that the optimal performance is achieved when the Granularity1-4 and Granularity2-5 are employed in instance category identification and relationship prediction, respectively. When the receptive range for granularity-awareness is too small, appropriate features might not be captured. Conversely, an expanded granularity receptive range might cause a more dispersed weight distribution, potentially affecting prediction stability. To better illustrate this finding, we provide the average weight distribution of the GrT module for small, medium, and large instances under the configurations of Granularity 1-4 and all five granularities. As shown in Figure 10, it can be observed that the weight proportion of Granularity5, remains nearly constant. This suggests that the information carried by this large receptive field feature may only serve as background context for the instance identification task, failing to yield substantial improvements while simultaneously diluting the contribution of other more beneficial features.

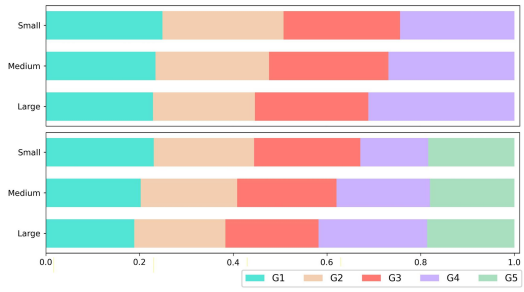


Fig. 10. Average weight distribution of the Granularity Transformer module when employing four-granularity (upper) and all five granularities (lower).

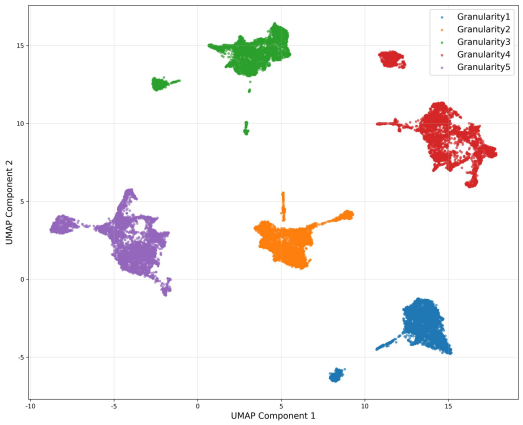


Fig. 11. UMAP visualization of instance features across five granularities.

Meanwhile, we conduct an experiment where the features seen by the GrT and CGCT modules are swapped. The performance drops significantly, further confirming that instance prediction relies on fine-grained features, while relationship judgment depends more on coarse-grained features.

On the other hand, we conducted additional experiments to investigate whether it is necessary to introduce an additional learnable granularity-specific embedding in the GranSSG to explicitly distinguish features across different granularities, thereby preventing potential performance loss caused by the inability of the GrT and CGCT modules in discerning the correspondence between features and granularities. As shown in Table IX, it is clear that introducing additional embeddings does not lead to a significant change. For example, after introducing the embedding, the performance fluctuates slightly by -0.18% and +0.01% in Triplet A@50/100. To better explain this, we provide UMAP visualization plots for instance features of each granularity. From Figure 11, it is evident that the features of various granularities encoded by GranSSG exhibit excellent separability, allowing the algorithm to easily discern the latent associations between features and their corresponding granularities, thus rendering additional specific embeddings unnecessary.

Module Design: We summarize the impact of our proposed methods on model performance. As shown in experiments ①-④ in Table X, each individual component within GranSSG positively affects the baseline model to varying degrees. Specifically, the baseline model employs the same backbone as GranSSG for scene encoding, while utilizing conventional

TABLE X
ABLATION STUDY OF MODULE DESIGN (%).

ID	GranSSG			Object		Predicate		Triplet		Triplet		
	VP	GrT	CGCT	A@1	A@5	A@1	A@3	A@50	A@100	A@50-s/m/l		
①				61.09	82.49	91.46	98.28	91.17	93.50	89.95	91.01	93.01
②	✓			62.53	83.80	91.59	98.24	91.95	93.28	90.40	91.32	93.70
③		✓		62.59	83.38	91.59	98.35	91.97	94.20	90.51	91.62	93.78
④			✓	62.03	82.78	91.88	**98.52**	91.76	94.01	90.31	91.21	93.75
⑤	✓	✓		63.04	83.80	91.52	98.30	92.01	94.07	90.03	91.26	93.82
⑥		✓	✓	62.68	83.69	91.77	98.42	92.03	94.25	90.02	91.74	93.77
⑦	✓		✓	62.42	83.72	91.99	98.36	91.88	94.10	90.17	91.44	93.84
⑧	✓	✓	✓	**63.14**	**83.89**	**92.11**	98.51	**92.64**	**94.64**	**90.74**	**91.99**	**94.35**

TABLE XI
COMPARISON OF MODEL EFFICIENCY. THE MODEL WITH ∗ PREDICTS A SINGLE RELATION TRIPLET AT A TIME. THE MODEL WITH † REMOVED CANDIDATE PAIR FILTERING

Model	Params. (M)	Inference time (sec)	Training time (h)	Triplet A@50 (%)
VL-SAT	25.06	74.99	22.6	90.35
GranSSG†	3.55	13.51	7.5	92.60
GranSSG	3.55	12.27	7.4	92.64
Granular3D*	2.08	316.56	39.1	93.15
GranSSG*	3.93	248.37	41.4	93.49

pooling and MLP layers to capture instance features for each granularity. For prediction, the baseline simply relies on single-granularity features, where Granularity3 is used for instance identification and Granularity5 for relationship prediction. Specifically, the VP improves the overall network performance by +0.78 in Triplet A@50 by supplementing and balancing the information at different granularities. Additionally, the GrT and CGCT blocks enhance instance classification and instance pair relationship prediction, with improvements of +1.5 in Object A@1 and +0.42 in Predicate A@1, respectively. This underscores the benefits of adaptive multi-granularity feature analysis in complex 3D scenes where instances with significant size discrepancies coexist. Furthermore, the effects of combining module pairs, as shown in experiments ⑤-⑦, continue to demonstrate positive trends. Overall, GranSSG enhances the baseline model by +1.47 in Triplet A@50. The size-based metric also shows significant positive impacts on scene graph prediction for instances of different granularities, with improvements of +0.79, +0.98, and +1.34 in Triplet A@50-s/m/l, highlighting the importance of granularity-awareness ability in the 3DSSG prediction task.

E. Further Analysis

Model Efficiency: We further examine the efficiency of the proposed GranSSG, while all experiments are under the same platform and the inference time reported here specifically denotes the pure network forward pass on the 3DSSG validation set. As illustrated in Table XI, our method demonstrates significant advantages. Compared to VL-SAT, GranSSG achieves nearly a 10-fold reduction in parameters and a 6× improvement in inference speed. Meanwhile, the results indicate that while GranSSG is inherently efficient, the candidate pair filtering further enhances inference speed. And from the comparison of training time, we can also observe that the proposed GranSSG consumes only 33% of the time required by VL-SAT. Simultaneously, under the type when distributed prediction a single relation triplet at a time, our GranSSG also exhibits a 20% speed advantage over Granular3D. Overall, this study demonstrates that GranSSG successfully balances accuracy and efficiency.

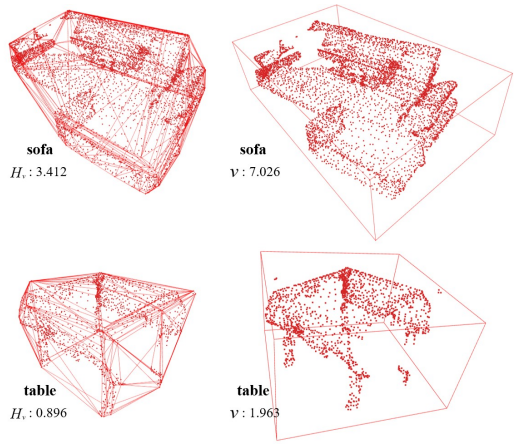


Fig. 12. Visualization of volume estimation using BBox and convex hull.

Robustness Study: To evaluate the anti-interference capability of our method, we then measure the robustness by applying a variety of perturbations. In addition to standard transformations such as scaling and jitter, we further assess the model under more challenging scenarios: 1) modifying the positions and sizes of instance bounding boxes, denoted as 'BBox bias', and 2) simulating imperfect instance segmentation results by removing peripheral regions of instance, denoted as 'Incomplete mask'. As shown in Table XII, the introduction of scale and jitter results in less than 0.5% drop in Triplet A@50, indicating that our method is robust to common perturbations. When introducing a 15% BBox bias, which directly affects the computation of granularity embeddings, GranSSG still maintains strong scene graph prediction performance, with only a -0.62 drop in Predicate A@1. Especially, in the 'Incomplete mask' condition, where original instance information is explicitly reduced, the method is most significantly affected. While GranSSG still delivers competitive results, it exhibits a performance drop of -0.48 and -1.17 in Triplet A@50. Overall, these comparisons demonstrate that our multi-granularity analysis, coupled with the VP approach, provides strong robustness against various types of perturbations.

Granularity Descriptor: We then provide a case study to explore the performance of convex hull volume as a main granularity descriptor. As illustrated in the upper part of Figure 12, for structurally non-convex and complex instances like the *L-shaped sofa*, traditional Bbox creates a loose envelope, capturing substantial empty space in the corner. On the other hand, for the instance with large internal space, such as *table*, while the points only exist on the tabletop and legs, human perception naturally considers the volume of the table to include the functional space it encloses. Matching this perception, the convex hull successfully approximates this subjective volume, whereas the BBox remains overly coarse for around 2 times and is more vulnerable to environmental outlier noise points. In summary, the convex message serves as an excellent granularity descriptor, providing a more stable and semantically meaningful representation for GranSSG.

Generalization Analysis: To further evaluate the generalization capability of our proposed method, we conduct

TABLE XII
ROBUSTNESS STUDY RESULTS (%) ON 3DSSG [2] VALIDATION SET. THE **BOLD** DENOTES THE BEST PERFORMANCE.

Methods	Object						Predicate						Triplet				
	A@1/5/10			A@1-s/m/l			A@1/3/5			A@1-s/m/l			A@50	A@100	A@50-s/m/l		
BBox bias (5%)	63.08	83.37	89.55	32.98	55.10	76.68	92.03	98.35	99.29	92.40	90.00	92.23	92.35	94.39	90.65	91.77	93.92
BBox bias (15%)	63.03	83.59	89.41	32.25	**55.45**	76.37	91.49	98.20	99.26	91.45	89.73	92.43	92.13	94.35	90.25	91.45	93.78
Incomplete mask (10%)	62.36	83.13	88.47	32.72	54.30	76.64	91.68	98.37	99.32	91.13	89.96	92.06	92.16	94.23	90.18	91.35	93.83
Incomplete mask (20%)	61.91	82.54	88.08	31.95	53.73	75.36	91.31	97.83	98.91	90.66	89.23	92.02	91.47	93.70	89.93	90.43	93.07
Scale (×0.8)	62.57	83.48	89.17	33.01	54.62	**76.83**	91.69	98.47	99.36	92.28	90.95	93.22	92.23	94.40	90.52	91.60	94.17
Scale (×1.2)	63.02	83.84	89.51	32.58	54.53	75.75	92.07	98.54	99.45	**92.61**	91.19	93.14	92.26	94.41	90.21	91.74	94.10
Jitter	62.97	83.74	**89.65**	32.30	53.90	76.54	91.96	**98.56**	99.46	92.48	91.23	93.23	92.44	94.61	90.71	91.95	94.31
None	**63.14**	**83.89**	89.45	**34.00**	55.34	76.45	**92.11**	98.51	**99.47**	92.48	**91.26**	**93.24**	**92.64**	**94.64**	**90.74**	**91.99**	**94.35**

TABLE XIII
QUANTITATIVE RESULTS (%) ON RELATIONSCANNET [50] VALIDATION SET. THE **BOLD** DENOTES THE BEST PERFORMANCE.

	Methods	SGCls		PredCls	
		R@20	mR@20	R@20	mR@20
Semantic	SGPN [2]	43.6	44.0	61.2	58.8
	SGFN [9]	64.7	56.0	73.5	65.3
	MonoSSG [45]	70.1	63.3	77.7	69.3
	SGGPC [50]	52.8	55.5	67.5	70.0
	HEDSGP [47]	86.9	83.6	93.7	87.1
	GranSSG	**89.2**	**84.5**	**96.7**	**93.6**
	GranSSG-O	78.4	72.1	88.0	83.9
Geometric	SGPN [2]	35.8	36.1	48.2	48.9
	SGFN [9]	55.8	49.9	62.0	58.6
	MonoSSG [45]	64.7	55.3	70.1	67.7
	SGGPC [50]	42.5	42.6	51.9	52.5
	HEDSGP [47]	79.2	78.7	87.3	**84.8**
	GranSSG	**82.9**	**80.1**	**87.9**	84.5
	GranSSG-O	70.7	67.9	76.0	72.8

experiments on the RelationScanNet dataset [50], which is derived from ScanNet [51] and includes 24 entity classes and 12 relation classes. Following the experimental settings in [50], we adopt Scene Graph Classification (SGCls) and Predicate Classification (PredCls) as evaluation metrics, and report performance individually for semantic and geometric relations. As shown in Table XIII, GranSSG achieves comparable results to HEDSGP on the PredCls task, while obtaining +2.3 and +3.7 improvements on R@20 in the SGCls task for semantic and geometric relations respectively, demonstrating the effectiveness of the proposed multi-granularity analysis. Additionally, to assess the impact of removing predefined class-agnostic instance masks, we also report the performance where instance masks are generated by a state-of-the-art instance segmentation method OneFormer3D [52], denoted as 'GranSSG-O'. The comparison in Table XII highlights the significant effect of instance mask accuracy on 3DSGG performance. Nonetheless, the competitive results of GranSSG-O suggest that our framework is still capable of real-world applications.

F. Limitation

While our proposed GranSSG demonstrates optimal performance in the 3DSSG prediction task through adaptive multi-granularity analysis, the accuracy in identifying smaller instances remains weaker. This bottleneck may reflect the challenges of relying solely on point clouds for complex 3D scene understanding, as point clouds are sparser than images and less effective in representing texture information.

To further analyze this limitation, we conducted a case study in Figure 13. As small instances struggle to clearly characterize differences in their geometric and textural attributes within sparse point clouds, this inevitably increases the difficulty of categorical judgment, even with advanced multi-granularity features. Furthermore, when instances exhibit similar sizes and geometric features, such as the observed confusion between a *box* and a *suitcase*, it results in identification confusion. Efficient fusion of multi-modal information presents a meaningful research direction for future research.

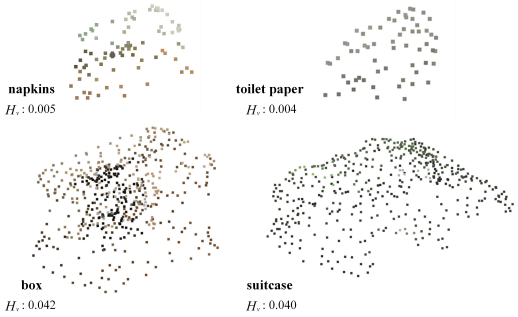


Fig. 13. Visualization of challenging instances. Sparse point distributions in small instances and the high geometric similarity between different categories represent the primary bottlenecks.

Additionally, current 3DSSG prediction methods depend on predefined class-agnostic instance masks. Integrating instance semantic segmentation with 3DSSG prediction could eliminate this dependency, enhancing practical applications.

V. CONCLUSION

An effective 3D perception method is crucial for advanced 3DSSG prediction, with granularity-awareness ability being indispensable. In this paper, we introduce GranSSG, a novel approach addressing the limitations of multi-granularity feature analysis in current 3DSSG prediction methods, particularly the coexistence of instances with significant size discrepancies in complex 3D scenes. GranSSG comprises a Volumetric Pooling block, a Granularity Transformer block, and a Cross-Granularity Correlation Transformer block. These components balance the information contained in various granularity features, adaptively shunt multi-granularity instance features, and comprehensively assess and fuse cross-granularity instance pair features, significantly enhancing the perception in real-world 3D scenes. Extensive experimental analysis validates the positive impact of the proposed methods, demonstrating that GranSSG achieves new state-of-the-art performance in the 3DSSG prediction task. Our code is available at *https://github.com/Hermione-HKX/GranSSG*.

REFERENCES

[1] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An embodied generalist agent in 3d world," in *ICLR 2024 Workshop: How Far Are We From AGI*, 2024, Conference Proceedings. 1

[2] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, Conference Proceedings, pp. 3961–3970. 1, 2, 4, 8, 9, 10, 16

[3] L. Chen, X. Wang, J. Lu, S. Lin, C. Wang, and G. He, "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*, 2024, Conference Proceedings, pp. 27 863–27 873. 1, 2, 10

[4] K. Huang, J. Yang, J. Wang, S. He, Z. Wang, H. He, Q. Zhang, and G. Lu, "Granular3d: Delving into multi-granularity 3d scene graph prediction," *Pattern Recognition*, p. 110562, 2024. 1, 2, 6, 7, 8, 10

[5] Z. Wang, B. Cheng, L. Zhao, D. Xu, Y. Tang, and L. 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*, 2023, Conference Proceedings, pp. 21 560–21 569. 1, 2, 4, 8, 9, 10

[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, Conference Proceedings, pp. 818–833. 1

[7] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, Conference Proceedings, pp. 821–830. 1, 3

[8] C. Zhang, J. Yu, Y. Song, and W. 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*, 2021, Conference Proceedings, pp. 9705–9715. 1, 2, 10

[9] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, Conference Proceedings, pp. 7515–7525. 1, 2, 4, 6, 10, 16

[10] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, Conference Proceedings, pp. 652–660. 1, 2, 4

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *In International Conference on Learning Representations (ICLR)*, 2016, Conference Proceedings. 1, 2

[12] S. Koch, N. Vaskevicius, M. Colosi, P. Hermosilla, and T. Ropinski, "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*, 2024, Conference Proceedings, pp. 14 183–14 193. 2, 10

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, Conference Proceedings, pp. 8748–8763. 2, 10

[14] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017. 2, 3, 4

[15] L. Hui, H. Yang, M. Cheng, J. Xie, and J. Yang, "Pyramid point cloud transformer for large-scale place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, Conference Proceedings, pp. 6098–6107. 2, 3

[16] M. Wang, T. Ma, X. Zuo, J. Lv, and Y. Liu, "Correlation pyramid network for 3d single object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, Conference Proceedings, pp. 3215–3224. 2

[17] P. Xiang, X. Wen, Y.-S. Liu, H. Zhang, Y. Fang, and Z. Han, "Retro-fpn: Retrospective feature pyramid network for point cloud semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, Conference Proceedings, pp. 17 826–17 838. 2, 3

[18] M. Ye, S. Xu, and T. Cao, "Hvnet: Hybrid voxel network for lidar based 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, Conference Proceedings, pp. 1631–1640. 2, 3

[19] X. Zou, J. Li, Y. Wang, F. Liang, W. Wu, H. Wang, B. Yang, and Z. Dong, "Patchaugnet: Patch feature augmentation-based heterogeneous point cloud place recognition in large-scale street scenes," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 206, pp. 273–292, 2023. 2, 3

[20] Z. Li and P. Xu, "Cspformer: A cross-spatial pyramid transformer for visual place recognition," *Neurocomputing*, vol. 580, p. 127472, 2024. 3

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, Conference Proceedings, pp. 8759–8768. 3

[22] X. Wang and Y. Yuan, "Mnat-net: Multi-scale neighborhood aggregation transformer network for point cloud classification and segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2024. 3

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, Conference Proceedings, pp. 2117–2125. 3

[24] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," *Advances in neural information processing systems*, vol. 35, pp. 23 192–23 204, 2022. 3

[25] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268. 3

[26] T. Zhang, H. Yuan, L. Qi, J. Zhang, Q. Zhou, S. Ji, S. Yan, and X. Li, "Point cloud mamba: Point cloud learning via state space model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 121–10 130. 3

[27] T. Guan, J. Wang, S. Lan, R. Chandra, Z. Wu, L. Davis, and D. Manocha, "M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, Conference Proceedings, pp. 772–782. 3

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015. 3

[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890. 3

[30] K. Zhiheng and L. Ning, "Pyramnet: Point cloud pyramid attention network and graph embedding module for classification and segmentation," *arXiv preprint arXiv:1906.03299*, 2019. 3

[31] Y. Li, J. Wen, R. Gong, B. Ren, W. Li, C. Cheng, H. Liu, and N. Sebe, "Pvafn: Point-voxel attention fusion network with multi-pooling enhancing for 3d object detection," *Expert Systems with Applications*, vol. 281, p. 127608, 2025. 3

[32] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022. 3

[33] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *2023 IEEE International Conference On Robotics and Automation*, 2023, Conference Proceedings, pp. 8216–8223. 4

[34] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, Conference Proceedings, pp. 16 259–16 268. 4

[35] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. 4

[36] J. Sankaranarayanan, H. Samet, and A. Varshney, "A fast k-neighborhood algorithm for large point-clouds," in *PBG@ SIGGRAPH*, 2006, Conference Proceedings, pp. 75–84. 5

[37] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, Conference Proceedings, pp. 10 076–10 085. 6

[38] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Nießner, "Rio: 3d object instance re-localization in changing indoor environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, Conference Proceedings, pp. 7658–7667. 8

[39] S. Zhang, A. Hao, and H. Qin, "Knowledge-inspired 3d scene graph prediction in point cloud," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 620–18 632, 2021. 9, 10

[40] S. Koch, P. Hermosilla, N. Vaskevicius, M. Colosi, and T. Ropinski, "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*, 2024, Conference Proceedings, pp. 3404–3414. 10

[41] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, Conference Proceedings, pp. 6163–6171. 10

[42] S. Sharifzadeh, S. M. Baharlou, and V. Tresp, "Classification by attention: Scene graph classification with prior knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, Conference Proceedings, pp. 5025–5033. 10

[43] Y. Ma, H. Liu, Y. Pei, and Y. Guo, "Heterogeneous graph learning for scene graph prediction in 3d point clouds," in *European Conference on Computer Vision*. Springer, 2024, pp. 274–291. 10

[44] M. Feng, H. Hou, L. Zhang, Z. Wu, Y. Guo, and A. 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*, 2023, Conference Proceedings, pp. 9182–9191. 10

[45] S.-C. Wu, K. Tateno, N. Navab, and F. Tombari, "Incremental 3d semantic scene graph prediction from rgb sequences," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5064–5074. 10, 16

[46] M. Feng, C. Yan, Z. Wu, W. Dong, Y. Wang, and A. Mian, "History-enhanced 3d scene graph reasoning from rgb-d sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 10

[47] ——, "Hyperrectangle embedding for debiased 3d scene graph prediction from rgb sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 10, 16

[48] S. Wu, H. Fei, and T.-S. Chua, "Universal scene graph generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 158–14 168. 10

[49] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 313–19 322. 10

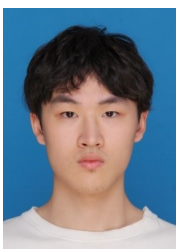
[50] W. Wei, P. Wei, J. Qin, Z. Liao, S. Wang, X. Cheng, M. Liu, and N. Zheng, "3d scene graph generation from point clouds," *IEEE Transactions on Multimedia*, vol. 26, pp. 5358–5368, 2023. 16

[51] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839. 16

[52] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, and D. Rukhovich, "Oneformer3d: One transformer for unified point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 943–20 953. 16

VI. Biography Section

Kaixiang Huang received a B.Eng. degree in Machine Design & Manufacturing and Automation from Sichuan University, Chengdu, China. He is currently pursuing his Ph.D. degree in Mechanical Engineering at the School of Zhejiang University, Hangzhou, China. His research interests include 3D computer vision and human-robot interaction.

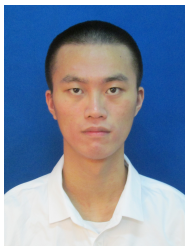


Qifeng Zhang received his B.Eng. degree in Mechanical Engineering from Zhejiang University. He is currently pursuing his M.Eng. degree in Mechanical Engineering at the School of Zhejiang University, Hangzhou, China. His research interests include 3D computer vision and human-robot interaction.

Jin Wang received a B.Eng. degree in mechatronic engineering and a Ph.D. degree in mechanical design and theory from Zhejiang University, Zhejiang, China, in 2003 and 2008, respectively. He is currently a professor at Zhejiang University. His research interests include computer vision.

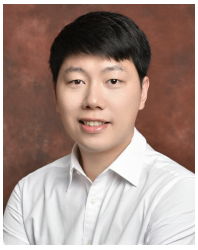
Jingru Yang is an incoming Postdoc at the School of Computer Science, Carnegie Mellon University. He received a B.Eng. degree in Mechanical Engineering from Sichuan University, Chengdu, China. He obtained a Ph.D. degree in Mechanical Engineering from the School of Zhejiang University, Hangzhou, China. His research interests include 2D and 3D computer vision.

Yang Zhou received the B.Eng. degree in vehicle engineering from Anhui Agricultural University, Hefei, China. He is currently pursuing his M.Eng. degree in Mechanical Engineering at the School of Zhejiang University, Hangzhou, China. His research interests include computer vision.



Jiao Yi received his B.Eng. degree in Mechanical Engineering from Zhejiang University. He is currently pursuing his M.Eng. degree in Mechanical Engineering at the School of Zhejiang University, Hangzhou, China. His research interests include Imitation learning and artificial intelligence.

Guodong Lu received the B.S. degree, M.Eng. degree, and the Ph.D. degree in Applied Mathematics from Zhejiang University, Zhejiang, China. He is currently a Professor at Zhejiang University, Hangzhou, China. His research interests are CAD, CG, and robotics.



Shengfeng He (Senior Member, IEEE) is an associate professor in the School of Computing and Information Systems, Singapore Management University. He was on the faculty of the South China University of Technology, from 2016 to 2022. He obtained B.Sc. and M.Sc. degrees from Macau University of Science and Technology in 2009 and 2011 respectively, and a Ph.D. degree from City University of Hong Kong in 2015. His research interests include computer vision and generative models. He is a senior member of IEEE and CCF. He serves as the lead guest editor of the IJCV, and the associate editor of IEEE TNNLS, IEEE TCSVT, Visual Intelligence, and Neurocomputing. He also serves as the area chair/senior program committee of ICML, AAAI, IJCAI, and BMVC.