Granular3D: Delving into multi-granularity 3D scene graph prediction

Kaixiang Huang [a,b] , Jingru Yang [a,b] , Jin Wang [a,b,f,*] , Shengfeng He [c] , Zhan Wang [d] , Haiyan He [e] , Qifeng Zhang [a,b] , Guodong Lu [a,b]

a Zhejiang University, State Key Laboratory of Fluid Power & Mechatronic Systems, Hangzhou, 310027, Zhejiang, China
b Zhejiang University, Robotics Institute of Zhejiang University, Hangzhou, 310027, Zhejiang, China
c Singapore Management University, 178903, Singapore
d Zhejiang Energy Digital Technology Co., Ltd, Department of Artificial Intelligence and Robotics, Hangzhou, 310027, Zhejiang, China
e Zhejiang Baima Lake Laboratory Co., Ltd., Hangzhou, 310000, Zhejiang, China
f Jinhua Key Laboratory of Robot Intelligent Welding Technology, Jinhua, 321000, Zhejiang, China

Published in Pattern Recognition (2024) 153, 110562. DOI: 10.1016/j.patcog.2024.110562

Abstract: This paper addresses the significant challenges in 3D Semantic Scene Graph (3DSSG) prediction, essential for understanding complex 3D environments. Traditional approaches, primarily using PointNet and Graph Convolutional Networks, struggle with effectively extracting multi-grained features from intricate 3D scenes, largely due to a focus on global scene processing and single-scale feature extraction. To overcome these limitations, we introduce Granular3D, a novel approach that shifts the focus towards multi-granularity analysis by predicting relation triplets from specific sub-scenes. One key is the Adaptive Instance Enveloping Method (AIEM), which establishes an approximate envelope structure around irregular instances, providing shape-adaptive local point cloud sampling, thereby comprehensively covering the contextual environments of instances. Moreover, Granular3D incorporates a Hierarchical Dual-Stage Network (HDSN), which differentiates and processes features of instances and their pairs at varying scales, leading to a targeted prediction of instance categories and their relationships. To advance the perception of sub-scene in HDSN, we design a Gather Point Transformer structure (GaPT) that enables the combinatorial interaction of local information from multiple point cloud sets, achieving a more comprehensive local contextual feature extraction. Extensive evaluations on the challenging 3DSSG benchmark demonstrate that our methods provide substantial improvements, establishing a new state-of-the-art in 3DSSG prediction, boosting the top-50 triplet accuracy by ＋2.8%.

Keywords: 3D point cloud, 3D semantic scene graph prediction, Gather point transformer, Multi-granularity

1. Introduction

Understanding complex 3D real-world environments is pivotal for various applications, including robotics, AR/VR, and navigation [1]. In this context, the prediction of 3D Semantic Scene Graphs (3DSSG) [2] has garnered significant attention. As illustrated in Fig. 1, predicting the categories of instances within a 3D point cloud environment, associated with class-agnostic instance masks, and determining their potential relationships forms the basis of relation triplets (e.g., <chair, close by, table>). Extending this approach to every possible instance pair in the environment results in a structured and comprehensive representation of 3D scenes.

However, despite its importance, the task of 3DSSG prediction faces substantial challenges. Unlike 2D semantic scene graph prediction [3], [4], 3D indoor scenes often present complex and densely populated instances, as exemplified by the room shown in Fig. 1. Current 3DSSG prediction methods, which predominantly focus on global scene processing [2], struggle to simultaneously identify all instances and their potential relationships within these intricate scenes. This process necessitates feature extraction from extensive point cloud patches, making it challenging to provide comprehensive and targeted multi-granularity feature extraction for instance identification and relationship prediction that forms the semantic scene graph. In particular, the primary limitations stem from three factors: (1) the point cloud patches sampled by the 3D bounding box are coarse and uneven, leading to incomplete contextual environmental information for feature extraction; (2) the reliance on single-scale point cloud feature extraction, which is incapable of discerning the fine-grained local patterns required for instance identification and the broader receptive field needed for relationship prediction; and (3) the employed PointNet [5] is deficient in perceiving

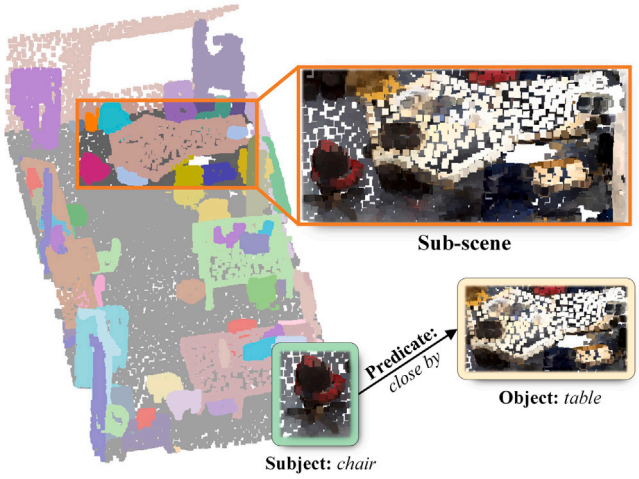

Fig. 1. Illustration of a sub-scene within a complex environment, mirroring human observational tendencies. It focuses on a localized area encompassing target instances, enabling the recognition of instance categories and the assessment of their relationships, predicting the relationship triplet <*subject*, *predicate*, *object*> independent of other instances in the environment. The emphasis is on the effective extraction of multi-grained information from the sub-scene.

the complex environment as it struggles to capture the local contextual information.

To date, few studies have deviated from the established 3DSSG prediction architecture to focus on effective multi-granularity analysis. Specifically, the pioneering work of SGPN [2] and its successors, such as SGG_{point} [6] and SGFN [7], have continued to employ a combination of PointNet [5] and Graph Convolutional Networks (GCNs) [8] with coarse 3D bounding box sampling method to understand entire complex scenes simultaneously. Frustratingly, this original single-scale global scene processing restricts the introduction of multi-scale and comprehensive feature extraction for instances and inter-instance relationships within the complete contextual environment, thereby directly limiting the effectiveness of 3DSSG prediction. Even recent advancements like VL-SAT [9], which acknowledges the need for enhanced sub-scene analysis, have not altered the fundamental architecture of point cloud feature extraction. On the contrary, as shown in Fig. 1, focusing on multi-granularity analysis from specific sub-scenes with complement contextual environment information should be an effective approach for predicting relationship triplets in the scene graph, while also better reflecting human observation patterns.

In contrast, although current 3DSSG prediction methods struggle in the single-scale analysis, hierarchical 3D point cloud feature extraction has seen considerable progress [10]. With the prior work PointNet++ [11] that utilizes set abstraction blocks for capturing local point cloud features at different granularity, the subsequent research has focused on enhancing local context perception, where Transformer-based networks show promising results. For instance, the Point Transformer [12] employs self-attention blocks for local information aggregation, while the Stratified Transformer [13] introduces a long-range local point cloud sampling strategy. However, these advancements have predominantly concentrated on single local point cloud set interactions, overlooking the potential benefits of combinatorial interactions among different local sets, a critical aspect of point cloud perception [5].

In this paper, we address the challenges posed by single-granularity feature extraction for global scenes in current 3DSSG prediction architectures. We present a novel solution, Granular3D, which diverges from the traditional architectures, reorganizing the environment information sampling method, network architecture, and contextual feature extraction structure for effective multi-granularity analysis. This strategy more accurately reflects human perceptual processes and emphasizes the prediction of individual relation triplets within a sub-scene to construct a comprehensive 3DSSG. Unlike the current coarse 3D bounding box partitioning strategy, a central to Granular3D is the Adaptive Instance Enveloping Method (AIEM), employing an approximate envelope around irregular instances to establish shape-adaptive local point cloud sampling. As a foundation of the subsequent multi-granularity analysis, this method allows Granular3D to construct a sub-scene with comprehensive contextual environment information. Furthermore, compared to the single-granular structure, our Hierarchical Dual-Stage Network (HDSN) within Granular3D identifies fine-grained local patterns within instances, as well as captures features with a larger receptive field for accurate prediction of inter-instance relationships. This hierarchical approach is instrumental in extracting multi-granularity information from the sub-scene. To promote the feature extraction of each granular in Granular3D, we introduce the Gather Point Transformer structure (GaPT) as the encoder of HDSN. In terms of innovatively gathering features from multiple local sets, this component facilitates the combinatorial interaction of local information, resulting in a more comprehensive contextual understanding across different granularities. Extensive evaluations demonstrate that Granular3D significantly outperforms existing 3DSSG prediction methods, establishing a new state-of-the-art with a remarkable +2.8% increase in top-50 triplet accuracy.

In summary, our main contributions are fourfold:

1. We propose Granular3D, comprising an Adaptive Instance Enveloping Method which establishes a shape-adaptive sampling around irregular instances, thereby constructing sub-scenes with comprehensive contextual environment information for the subsequent multi-granularity analysis.
2. Additionally, Granular3D incorporates a Hierarchical Dual-Stage Network to process and differentiate features of instances and their associated pairs across multiple scales. The multi-granularity analysis significantly enhances the accuracy of predicting instance categories and their interrelations, thereby achieving superior performance.
3. We design the Gather Point Transformer for Granular3D. Through the interaction of features across multiple point cloud sets, this unique multi-local feature extraction results in a more comprehensive contextual feature extraction, markedly enhancing the perception capability under various granularity.
4. Extensive experimental evaluations demonstrate the significant advancements brought by Granular3D. Notably, our approach establishes a new state-of-the-art in 3DSSG prediction, evidencing the substantial potential of multi-granularity analysis in this field.

2. Related work

Scene Graph Prediction in Point Cloud: Building upon the foundation laid by 2D image-based scene graph prediction [14–19], recent endeavors in the 3D point cloud domain [2,6,7,9] seek to replicate this success, which aims to identify instances and their interrelationships within a 3D scene to construct the scene graphs. The seminal work [2] introduces the 3DSSG benchmark, along with the SGPN model, which pioneered the use of the PointNet [5] and GCN [8] architecture in this context. Following this, SGG_{point} [6] enhances the GNN module with an edge-oriented scene graph reasoning method. Zhang et al. [20] introduces a graph auto-encoder network to automatically learn and inject per-learned knowledge into the 3DSSG prediction network. Additionally, SGFN [7] explores incremental 3DSSG prediction in dynamically created RGB-D scenes. Despite these advancements, current methods still adhere to the foundational SGPN architecture that focuses on global scene processing, lacking in providing comprehensive contextual environment information while exhibiting limitations in point cloud feature extraction. The recent work VL-SAT [9] while recognizing the deficiency of current architecture in sub-scene information extraction,

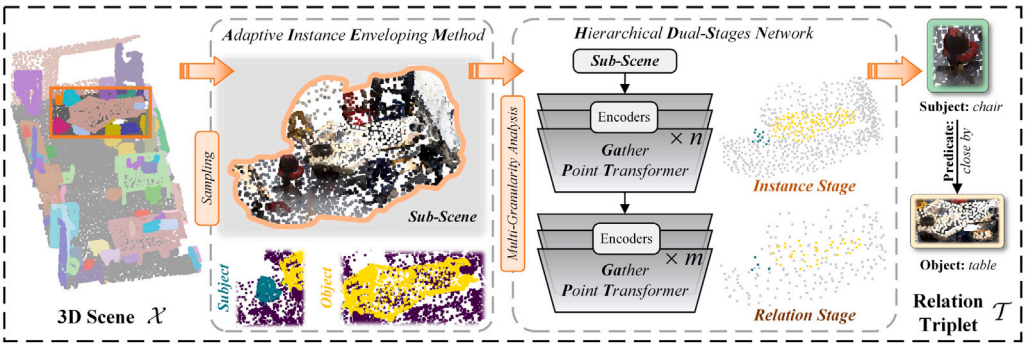

Fig. 2. Overview of the proposed Granular3D. For complex 3D indoor environments, we first apply the proposed Adaptive Instance Enveloping Method (AIEM) to sample sub-scenes, ensuring comprehensive coverage of the contextual environments of instances. Building upon the AIEM, we introduce a Hierarchical Dual-Stage Network (HDSN) for targeted multi-granularity analysis of instance identification and inter-instance relationship prediction, thereby obtaining the implied relation triplets that form the scene graph. Furthermore, in both stages of the HDSN, we employ the proposed Gather Point Transformer (GaPT) for effective feature extraction from point clouds at various granularities. Altogether, the Granular3D provides multi-granularity feature extraction, thereby significantly elevating the 3DSSG prediction performance.

frustratingly relies on supplemental 2D image and natural language descriptions rather than breaking the established architecture.

3D Scene Understand with Hierarchical Network: Hierarchical networks, due to their effective perception of 3D point cloud environments, have been widely studied and applied. Mirroring the success of hierarchical feature learning in 2D fields [21,22], PointNet++ [11] first introduces the set abstraction block to capture fine-grained 3D point cloud structures and features with a broader receptive field at varying scales. Following the paradigm of PointNet++, various researches with hierarchical network are emerged [10,23–29]. For example, in 3D point cloud recognition, PatchAugNet [29] employs a pyramid point cloud transformer network for multi-stage feature capture and enhances discrimination. Additionally, in 3D object detection, HVNet [26] also extracts and aggregates the voxel features from different scales and achieves competitive results. However, the effective hierarchical network has not been applied in current 3DSSG prediction methods, even though the multi-grained information provided by the hierarchical network fits the targeted recognition of instance categories and their relationships.

Local Point Cloud Feature Perception: Analogous to the set abstraction block in PointNet++ [11], local point feature perception is the crucial component in hierarchical network and has attracted plentiful attention [10–12,30–32]. Specifically, PointWeb [32] proposes an adaptive feature adjustment module that uses trainable indicators to distribute point features in the local set. Moreover, RandLA Net adopts attentive pooling with attention scores in local feature perception. Meanwhile, driven by the remarkable success of Transformer and self-attention mechanisms in NLP and 2D image [33–36], many approaches investigate the application of Transformer in point cloud feature perception [12,13,37]. For example, Point Transformer [12] leveraged vector attention for contextual feature aggregation. Stratified Transformer [13] then explores the memory-efficient transformer with an enlarged local set receptive field. However, these methods primarily focus on single local sets, overlooking the combinatorial interactions among different local sets, which is a main attribute of 3D point cloud perception [5]. In contrast, our proposed Gather Point Transformer innovatively merges features from multiple local sets, enabling the extraction of more comprehensive contextual features.

3. Methodology

Predicting 3D Semantic Scene Graphs (3DSSG) is pivotal in comprehending complex 3D environments. It involves identifying categories of instances and the relationships between linked subject-object pairs, thereby constructing a structured scene graph. To delineate our proposed Granular3D, this section is organized as follows. In Section 3.1, we first formulate the 3DSSG prediction problem in point cloud, while presenting an overview of our methods. In Section 3.2, we detailly introduce our proposed Granular3D, including the Adaptive Instance Enveloping Method (AIEM), Hierarchical Dual-Stages Network (HDSN), and Gather Point Transformer (GaPT).

3.1. Problem formulation and method overview

Suppose a 3D scene $\mathcal{X}$ consisting of a set of points, where each point $x = (\boldsymbol{p}, \boldsymbol{f}, m)$ encompasses not only the essential position information $\boldsymbol{p} \in \mathbb{R}^3$, attribute information $\boldsymbol{f} \in \mathbb{R}^n$ (e.g., color) but also appends a class-agnostic instance mask $m \in \{1, \dots, k\}$ to distinguish individual instance $\{\mathcal{M}_1, \dots, \mathcal{M}_k\}$ within $\mathcal{X}$. Following the principles outlined by SGPN [2], the ultimate goal is to construct a scene graph $\mathcal{G} = (\mathcal{C}, \mathcal{R})$, where the predictions of instances categories $\mathcal{C}$ and their inner relationships $\mathcal{R}$ depict the nodes and edges of the scene graph. Within graph $\mathcal{G}$, the relation triplet <*subject*, *predicate*, *object*>, denoted as $\mathcal{T}$, is the basic component. Concretely, for a pair of instances $\mathcal{M}_i$ and $\mathcal{M}_j$ in $\mathcal{X}$, the categories of instance c_i and c_j serve as the *subject* (head node) and *object* (tail node), respectively, while their potential relationship r_{ij} forms the *predicate* (edge).

In contrast to prevailing methods that focus on global scene processing and single-scale feature extraction, methodology pivots towards the multi-granularity analysis, which efficiently predicts relation triplets within each sub-scene, thereby organizing the 3D scene graph. As demonstrated in Fig. 2, the proposed AIEM is initially employed to sample sub-scene containing subject-object instance pair, aiming to adaptively extract comprehensive contextual environment information, serving as the foundation for subsequent multi-granularity analysis. Next, the HDSN differentiates and processes features of instances and the instance pair at varying scales, providing specific features for the targeted prediction of the relation triplet. Moreover, as a crucial component of HDSN, we introduce the innovative GaPT as the encoder to extract features from point clouds at various granularities, enhancing the ability of local contextual feature perception.

3.2. Granular3D

3.2.1. Adaptive instance enveloping method

Mirroring human observation patterns, sampling target instances from complex global scenes while ensuring the completeness of their contextual environmental information forms the foundation for subsequent multi-granularity feature extraction, particularly crucial in instance identification. As the example shown in Fig. 3, observing the adjacent *table* intuitively enhances the accuracy and credibility of the *chair* identification. Therefore, for a pair of subject-object instances $\mathcal{M}_i$ and $\mathcal{M}_j$, the first challenge is to construct an informative sub-scene S_{ij} from the global scene.

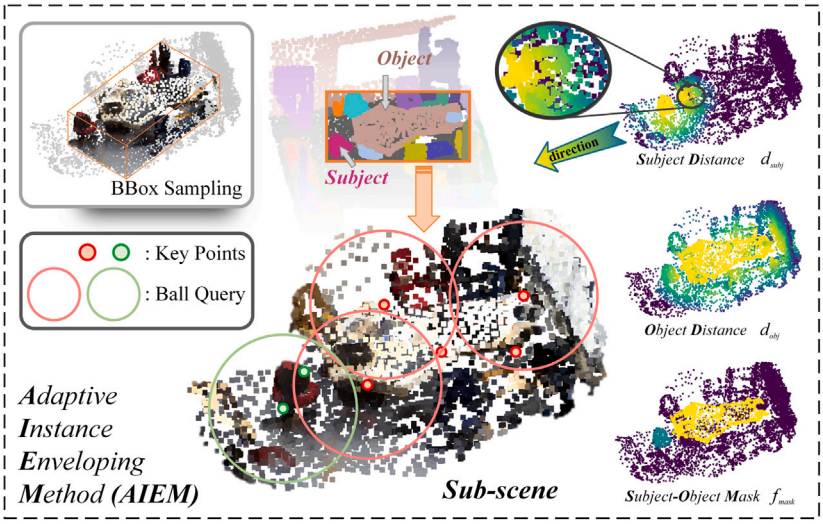

Fig. 3. Illustration of the proposed Adaptive Instance Enveloping Method (AIEM). Contrasting with traditional BBox Sampling, AIEM first evenly samples key points within target instances and employs Ball Query around these points to establish an approximate envelope structure, achieving adaptive surrounding environment sampling, especially for irregular instances. Meanwhile, AIEM further offers the distance features to the subject/object instance, enriching the contained information.

Noticing the current 3D bounding box sampling method coarsely defines the border of sub-scene by the range of instances, providing incomplete contextual environmental information, we then propose the Adaptive Instance Enveloping Method (AIEM) to adaptively sample the neighborhood point cloud by creating an approximate envelope structure around each instance.

Denotes the targeted subject as $\mathcal{M}_i$, since the 3D envelop structure can be approached by a finite number of spheres [38], the initial step in AIEM involves selecting a set of key points $\mathcal{P}_i$ from $\mathcal{M}_i$, which uniformly distributes over the irregular instance, especially including protruding structures. For this, we innovatively implement the Farthest Point Sampling (FPS) algorithm that keeps the selected $\mathcal{P}_i$ evenly covering the whole set [39]. Furthermore, we consider each key point $\boldsymbol{x}_i^n$ in $\mathcal{P}_i$ as centroid, then perform Ball Query to identify all scene points $\mathcal{S}_i$ which are within a radius of the key point. Finally, by combining both scene points in instance pair $\mathcal{M}_i$ and $\mathcal{M}_j$, the entire sub-scene $\mathcal{S}_{ij}$ is generated. In summary, the neighborhood point cloud sampled via AIEM is expressed as follows:

$$\begin{aligned} \mathcal{S}_i &= \left\{ B\left(\boldsymbol{x}_i^n, R\right) | \forall \boldsymbol{x}_i^n \in F\left(\mathcal{M}_i\right) \right\} \\ \mathcal{S}_{ij} &= \left\{ \mathcal{S}_i, \mathcal{S}_j \right\} \end{aligned} \tag{1}$$

where, $B(\cdot)$ and $F(\cdot)$ represent the Ball Query and FPS algorithm correspondingly, and R represents the radius used for the ball query searching.

The incorporation of an approximate envelope structure in the sampling of the surrounding environment allows AIEM to adaptively construct sub-scenes. As depicted in Fig. 3, AIEM surpasses traditional BBox sampling methods by capturing a more comprehensive neighborhood context, such as including *bookshelves* and *walls* around the *table*. Furthermore, the congruence of sub-scene edges with the irregular shapes of instances, like the desktop, further demonstrates the adaptability of our AIEM to irregular instances.

Moreover, in addition to the shape-adaptive sub-scene point cloud sampling, the AIEM introduces subject/object distance features, enriching the contained information. Specifically, for a point $\boldsymbol{x}_s^m$ within $\mathcal{S}_{ij}$, the subject distance feature d_{subj}^m is directly calculated as the normalized nearest distance to the subject instance, as follows:

$$d_{subj}^m = \max\left(\left(R - Dist\left(x_s^m, \mathcal{M}_i\right)\right), 0\right) / R \tag{2}$$

where, the $Dist(\cdot)$ represents the nearest neighbor distance calculation. Analogously, the object distance feature d_{obj}^m is determined by calculating the distance to the object instance.

As illustrated in Fig. 3, the d_{subj} and d_{obj} features visually indicate the proximity of each point to the corresponding instance. Meanwhile, within a local point set, the trend of distance features indirectly characterizes the relative spatial orientation between the point and target instances, thus aiding in the prediction of spatial relationships, such as *left*, *right*. Finally, the AIEM also integrates the subject-object mask f_{mask} and distinctly labels the subject instance, object instance, and neighboring points as m_{subj}, m_{obj}, and m_{nei}, for identification.

In summary, the innovative approach to shape-adaptive neighborhood environment sampling of our AIEM, coupled with the introduction of additional features, enables the construction of sub-scenes with comprehensive contextual information. This forms a robust foundation for the subsequent multi-granularity feature extraction and analysis.

3.2.2. Hierarchical dual-stages network

In a complex 3D indoor environment, identifying instances requires fine-grained local patterns, while predicting their potential relationship needs a broader receptive field. This conflict renders single-granularity 3D scene extraction incapable of achieving targeted 3DSSG prediction. Therefore, we further proposed the Hierarchical Dual-Stages Network (HDSN) within the Granular3D, aiming to comprehensively extract the features at multiple granularities. As visualized in Fig. 4, during the hierarchical feature encoding and down-sampling process of sub-scene, our HDSN divides two specific stages dedicated to classifying instances and relationships respectively, eventually leading to the multi-granularity analysis.

Backbone Structure. Generally, the backbone of our proposed HDSN is organized by multiple feature encoders that process the progressively down-sampled sub-scene. Specifically, the backbone is bifurcated into two stages, each targeting a different granularity of feature extraction: the instance and the relationship stages. The down-sampling rates for the first stage are [1, 4, 4, 2], thus generating an instance-level feature map $\boldsymbol{F}_{ins}$, reducing the original point count to $N/16$, where N is the number of sub-scene points. Building upon $\boldsymbol{F}_{ins}$, the HDSN transitions to the second stage, employing down-sampling rates of [2, 2], and thereby forms the relationship-level feature map $\boldsymbol{F}_{rel}$, further reducing the point count to $N/64$. The instance-level feature map $\boldsymbol{F}_{ins}$, with a relatively lower down-sampling rate, preserves more local structural details pertinent to categorizing instances. Conversely, the relationship-level feature map $\boldsymbol{F}_{rel}$, derived from deeper network layers, benefits from a higher down-sampling rate, resulting in an expanded receptive field conducive to discerning relationships between instances.

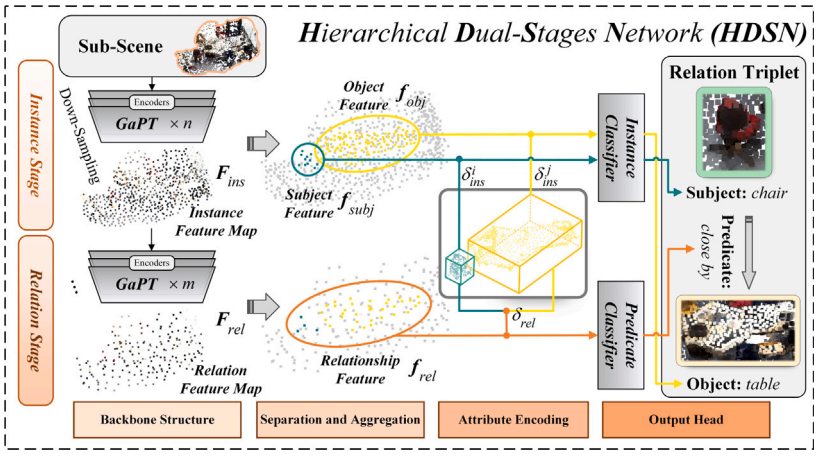

Fig. 4. Illustration of the proposed Hierarchical Dual-Stage Network (HDSM). Focus on the multi-granularity analysis, HDSM hierarchically processes the sub-scene, encoding the point cloud at multiple scales to extract feature maps at both the instance and relationship stages. Next, features with different granularity are separated and aggregated according to the instance mask, with attribute encoding supplementing basic shape and spatial information before decoding.

Separation and Aggregation. After obtaining the targeted feature maps, the subsequent stage involves minimizing the interference of background features. Therefore, we adopt a feature separation and aggregation method, which isolates the essential point features by tracking the subject-object mask f_{mask} during down-sampling. As depicted in Fig. 1, through the mask, our HDSN leverages this mask to segregate the point clouds corresponding to the subject instance, object instance, and neighboring points, thus refining the multi-grained features for analysis. Detailly, at the instance level, HDSN ought to aggregate the features specific to subject and object instances respectively. While at the relationship level, the HDSN amalgamates the features of both instances to comprehensively perceive their correlation. The global average pooling is performed across pointwise features during the aggregation process. The mathematical formulation of this feature separation and aggregation is described as follows:

$$\begin{aligned} \boldsymbol{f}_{subj} &= AvgPooling\left(\left\{\boldsymbol{f}^{p}_{ins} \in \boldsymbol{F}_{ins} | f^{p}_{mask} = m_{subj}\right\}\right) \\ \boldsymbol{f}_{rel} &= AvgPooling\left(\left\{\boldsymbol{f}^{p}_{rel} \in \boldsymbol{F}_{rel} | f^{p}_{mask} \neq m_{nei}\right\}\right) \end{aligned} \tag{3}$$

where, the $\boldsymbol{f}_{subj}$ and $\boldsymbol{f}_{rel}$ represent the aggregated subject feature and relationship feature respectively, $\boldsymbol{f}^{p}_{ins}$ and $\boldsymbol{f}^{p}_{rel}$ represent a sampling point in $\boldsymbol{F}_{ins}$ and $\boldsymbol{F}_{rel}$ which fits the screening mask. And by replacing the screening conditions of $\boldsymbol{f}_{subj}$ to m_{obj}, the aggregated object feature $\boldsymbol{f}_{obj}$ is constructed.

Attribute Encoding. The attributes of instances, including the geometry center $\boldsymbol{g} = (g_x, g_y, g_z)$, size $\boldsymbol{s} = (s_x, s_y, s_z)$, and volume $v = s_x s_y s_z$, play a significant role in representing the primitive shape and spatial information of each instance [7]. Therefore, except for the aggregated features in instance and relationship level, the proposed HDSN further introduces the trainable attribute encodings for both instance features and relationship features via multi-layer perceptron (MLP) layers. Specifically, the instance-specific attribute encoding δ_{ins} focuses only on the attributes of the corresponding single instance, while the relation-specific attribute encoding δ_{rel} measures the difference between the subject instance and object instance, as follows:

$$\begin{aligned} \delta^{i}_{ins} &= MLP_{ins}\left(Cat\left(\boldsymbol{g}_i, \boldsymbol{s}_i\right)\right) \\ \delta_{rel} &= MLP_{rel}\left(Cat\left(\boldsymbol{g}_i - \boldsymbol{g}_j, \boldsymbol{s}_i - \boldsymbol{s}_j, \ln\left(v_i / v_j\right)\right)\right) \end{aligned} \tag{4}$$

where, the $MLP_{ins}(\cdot)$ and $MLP_{rel}(\cdot)$ denote the mapping functions, and the $Cat(\cdot)$ denotes the concatenation operation.

Subsequently, attribute encodings are directly added with the corresponding instance feature and relationship feature to complement the primitive shape information and relative spatial relationship.

Output Head. In the final phase, to respectively focus on instance features and relational features with different granularities, we implement two MLP layers. These layers independently decode the instance categories and their relationships, resulting in the formation of relation triplet $\mathcal{T}_{ij} = \{c_i, r_{ij}, c_j\}$ which describes the sub-scene. By systematically predicting all possible instance pairs in the scene through HDSN, we then obtain the comprehensive scene graph.

Losses. We perform a joint loss $\mathcal{L}_{tri}$ when training the HDSN, which optimizes both the instance classification loss $\mathcal{L}_{ins}$ and the relationship classification loss $\mathcal{L}_{rel}$, as follows:

$$\mathcal{L}_{tri} = \lambda_{ins}\mathcal{L}_{ins} + \lambda_{rel}\mathcal{L}_{rel} \tag{5}$$

where, λ_{ins} and λ_{rel} are the respective weighting factors, with values set to 0.3 and 1. Given that a pair of instances can simultaneously exhibit multiple relationships [2], we utilize per-class binary cross entropy when calculating $\mathcal{L}_{rel}$. Conversely, multi-class cross entropy is used to calculate $\mathcal{L}_{ins}$.

In summary, focusing on instances and subject-object instance pairs, the proposed Hierarchical Dual-Stage Network (HDSN) in our Granular3D innovatively applies a multi-granularity feature extraction structure. This enables targeted prediction of instance categories and their potential relationship in complex environments, significantly enhancing the effectiveness of scene graph prediction.

3.2.3. Gather point transformer

For each scale of HDSN, the local point cloud feature encoder is utilized to capture the corresponding point cloud structures and patterns through local point set feature perception, which directly impacts the perception capability of complex environments. However, existing methods isolate the various local features of point clouds. Therefore, as a pivotal component of our Granular3D, we innovatively propose the Gather Point Transformer structure (GaPT), designed to enable combinatorial interaction among point features across multiple local sets, leading to more comprehensive contextual information extraction through the interaction of multiple local contextual features.

POS Encoding. Acknowledging that the fine-grained point information may lost in the deep network layers [40], we first introduce a POS Encoding to complement not only the **p**osition information which represents the local structure but also includes the **s**ubject/**o**bject distance features, d_{subj} and d_{obj}, designed in AIEM that describe the correlation to subject/object instance. Specifically, for each point $\boldsymbol{x}_i$ within the local point cloud set $\mathcal{N}(\boldsymbol{x}_a)$ that centered around $\boldsymbol{x}_a$, the POS encoding δ^{i}_{POS} is formulated as:

$$\delta^{i}_{POS} = MLP_{POS}\left(Cat\left(\boldsymbol{p}_a - \boldsymbol{p}_i, d^{a}_{subj} - d^{i}_{subj}, d^{a}_{obj} - d^{i}_{obj}\right)\right) \tag{6}$$

where, the $MLP_{POS}(\cdot)$ is the mapping function. Subsequently, through a straightforward vector addition, $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \delta^{i}_{POS}$, a more informative local set $\mathcal{N}(\hat{\boldsymbol{x}}_a)$ is generated.

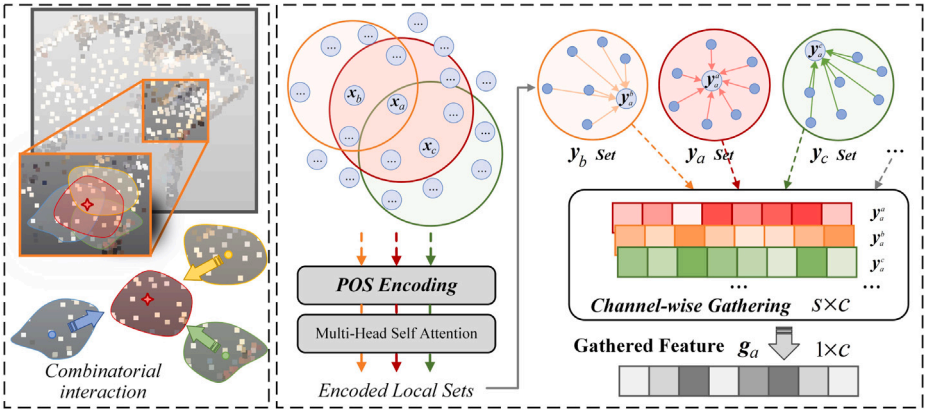

Fig. 5. Illustration of the proposed Gather Point Transformer (GaPT), where each point cloud patch or circle indicates a local set. The left part shows the combinatorial interaction of point features from multiple local sets. The right part demonstrates the application of our GaPT. For each point, the gathering of its features among multiple encoded local sets enables the interaction of extracted local structures and patterns, leading to a more comprehensive contextual information perception.

Local Set Self-Attention. With the enhanced local set feature $\mathcal{N}(\hat{\boldsymbol{x}}_a) \in \mathbb{R}^{n\times(h\times d)}$, GaPT then employs the original multi-head self-attention to encode the local set. Formally, n represents the point number within the local set, h and d represent the head number and head dimension, and $c = h \times d$ is the dimension of the point feature in $\mathcal{N}(\hat{\boldsymbol{x}}_a)$. Consequently, the multi-head self-attention in the local set is defined as follows:

$$\begin{aligned} &Q = W_q\left(\mathcal{N}(\hat{\boldsymbol{x}}_a)\right), K = W_k\left(\mathcal{N}(\hat{\boldsymbol{x}}_a)\right), V = W_v\left(\mathcal{N}(\hat{\boldsymbol{x}}_a)\right) \\ &\mathcal{N}(\boldsymbol{y}_a) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{7}$$

where, the W_q, W_k, and W_v are the trainable weight matrixes, $\mathcal{N}(\boldsymbol{y}_a) \in \mathbb{R}^{n\times(h\times d)}$ represents the encoded local set point features, and $\sqrt{d_k}$ is a scaling factor.

Feature Gathering. Through the multi-head self-attention, each point feature in the encoded local set $\mathcal{N}(\boldsymbol{y}_a)$ can attend to all point features in the $\mathcal{N}(\hat{\boldsymbol{x}}_a)$, thereby weighting the correlation of different points in the set. Consequently, for the central point $\boldsymbol{x}_a$, this mechanism extracts local contextual features, resulting in the generation of $\boldsymbol{y}_a$. When this local set encoding is applied across the entire sub-scene, each point has the capability to aggregate contextual features from its neighboring points. However, in this case, the feature extraction for each local set is typically isolated, while serving only the respective center point. As shown in Fig. 5, even if the nearest local set $\mathcal{N}(\boldsymbol{x}_b)$ which is centered around $\boldsymbol{x}_b$ contains $\boldsymbol{x}_a$, the feature in $\boldsymbol{y}_a^b \in \mathcal{N}(\boldsymbol{y}_b)$ remains unutilized. To address this, our GaPT introduces the feature gathering block which fully utilizes the feature extraction of self-attention across different local sets, thereby facilitating the interaction between multiple local features.

Specifically, for a sampled point $\boldsymbol{x}_a$, the primary objective is to consider not only its basis egocentric local set $\mathcal{N}(\boldsymbol{x}_a)$, but also the surrounding local sets that include $\boldsymbol{x}_a$, such as the $\mathcal{N}(\boldsymbol{x}_b)$ and $\mathcal{N}(\boldsymbol{x}_c)$ in Fig. 5. To be specific, before encoding local sets, we first apply the k-nearest neighbor (kNN) search in $\boldsymbol{x}_a$, yielding a set of neighboring points $\mathcal{K}(\boldsymbol{x}_a) = \{\boldsymbol{x}_a, \boldsymbol{x}_b, \ldots, \boldsymbol{x}_n\}$. Next, after the feature encoding through multi-head self-attention, the example point feature $\boldsymbol{y}_a^b \in \mathcal{N}(\boldsymbol{y}_b)$ then symbolizes the correlation between $\boldsymbol{x}_a$ and the pattern in $\mathcal{N}(\boldsymbol{x}_b)$, essentially capturing contextual information from an alternate perspective. Parallelly, the features set $\mathcal{F}(\boldsymbol{y}_a) \in \mathbb{R}^{s\times c}$ of $\boldsymbol{x}_a$ in multiple local sets is constructed, where the s represents the number of searched local sets, inclusive of $\mathcal{N}(\boldsymbol{x}_a)$ itself. Furthermore, the GaPT employs a series of SharedMLP [5] for the channel-wise gathering of these features, culminating in the formation of the aggregated feature $\boldsymbol{g}_a$. Generally, the mathematical expression for feature gathering in the GaPT is expressed as follows:

$$\begin{aligned} &\mathcal{K}(\boldsymbol{x}_a) = \text{kNN}(\boldsymbol{x}_a, \mathcal{S}) \\ &\mathcal{F}(\boldsymbol{y}_a) = \left\{\boldsymbol{y}_a^a, \boldsymbol{y}_a^b, \boldsymbol{y}_a^c, \ldots \,|\forall \boldsymbol{y}_a^k \in \mathcal{N}(\boldsymbol{y}_k) \text{ and } \boldsymbol{x}_k \in \mathcal{K}(\boldsymbol{x}_a)\right\} \\ &\boldsymbol{g}_a = \text{Ga}\left(\mathcal{F}(\boldsymbol{y}_a)\right) \end{aligned} \tag{8}$$

where, the $\mathcal{S}$ represents the sub-scene, and the $\text{Ga}(\cdot)$ represents the channel-wise gathering function.

To summarize, the proposed Gather Point Transformer (GaPT) extends beyond mere feature aggregation within the local neighborhood of a sampling point. It innovatively captures the point's connections with multiple surrounding neighborhoods, effectively establishing combinatorial interactions between various local sets. This approach allows GaPT to perceive complex local structures and patterns from a multitude of perspectives, contributing to a more comprehensive contextual information perception across different granularities in the HDSN, and thereby achieving improved 3DSSG prediction performance.

4. Experiment

In this section, to evaluate the benefits of our proposed Granular3D, we conduct intensive experiments on the large-scale 3D point cloud semantic scene graph prediction dataset, 3DSSG [2]. Furthermore, we also conduct extensive ablation experiments to validate the efficacy of each component within our methods.

4.1. Setups and implementation details

Dataset: We chose to conduct experiments on the 3DSSG dataset [2] to validate the effectiveness of the proposed method, which is a large-scale 3D dataset that expands upon the 3RScan [41] with detailed semantic scene graph annotations. The 3DSSG dataset encompasses 1553 real-world indoor point cloud scenes, where each point possesses coordinates, RGB information, and a class-agnostic instance mask. For the various instances in the complex indoor scene, the 3DSSG dataset includes semantic labels from 160 semantic categories and 26 types of predicates to represent their potential relationships. We evaluate our proposed methods in the same training/validation split as applied in the 3DSSG dataset [2].

Evaluation Metrics: Since the key to scene graph prediction lies in classifying instances and their potential relationship, we then respectively adopt the top-k accuracy metric to evaluate the prediction performance of the instance and predicate, denoted as Object A@k and Predicate A@k as defined in [9]. This allows us to finely evaluate the performance of our Granular3D, demonstrating the effectiveness of multi-granularity analysis in the scene graph prediction process.

Table 1
Quantitative results (%) on 3DSSG [2] validation set. The **bold** denotes the best performance.

Model	Encoder	Object			Predicate			Triplet	
		A@1	A@5	A@10	A@1	A@3	A@5	A@50	A@100
SGPN [2]	PointNet [5]	48.28	72.94	82.74	91.32	98.09	99.15	87.55	90.66
SGG_{point} [6]		51.42	74.56	84.15	**92.40**	97.78	98.92	87.89	90.16
SGFN [7]		53.67	77.18	85.14	90.19	98.17	99.33	89.02	91.71
VL-SAT [9]		55.66	78.66	85.91	89.81	**98.45**	**99.53**	90.35	92.89
Granular3D	**GaPT (Ours)**	**66.02**	**86.13**	**91.60**	91.29	98.35	99.45	**93.15**	**95.13**
Granular3D	PointNet [5]	57.48	81.37	88.67	89.56	97.78	99.11	90.90	93.36
Granular3D	PointNet++ [11]	64.09	84.75	90.16	90.62	98.23	99.42	92.52	94.58
Granular3D	Point Transformer [12]	65.23	85.81	91.26	91.31	98.11	99.36	92.87	94.91

Meanwhile, as an essential component of scene graphs, which comprehensively represents the prediction effect, we also apply the top-k accuracy metric to evaluate the predicted relationship triplets $\mathcal{T} = \langle subject, predicate, object \rangle$. Following the methodology from [9], we first multiply the prediction scores of *subject*, *predicate*, and *object* to derive a combined relation triple score, subsequently ranked in order. Additionally, in this standard, the relation triple is ultimately considered correct only if all components (*subject*, *predicate*, and *object*) are accurately predicted, thereby effectively measuring the superiority of different algorithmic comprehensive performances.

Further, we also follow the Zhang et al. [20] and report the performance of our methods in two specific tasks to assist in evaluating method effectiveness: (1) Scene Graph Classification (SGCls) and (2) Predicate Classification (PredCls). Specifically, the SGCls task assesses the accuracy of the entire triplet, while the PredCls task focuses solely on predicate accuracy, given the ground-truth labels of the instance entities. Therefore, the SGCls task reflects the comprehensive prediction effect of the scene graph, while PredCls evaluates the accuracy of predicate prediction. As defined in Zhang et al. [20], the recall at the top-k (R@k) triplets is considered as the metric of the two tasks, where the triplet is considered correct when the subject, predicate, and object are all correct.

Baseline methods: We compare our method with a list of advanced methods, including SGPN [2], SGG_{point} [6], SGFN [7], Co-Occurrence [20], KERN [14], Schemata [42], Zhang et al. [20] and the current state-of-the-art method VL-SAT [9]. In addition, considering the flexibility of our proposed Granular3D in terms of applying different point cloud encoding methods in the HDSN, we also evaluated the performance under various point cloud encoders, including PointNet [5], PointNet++ [11] and Point Transformer [12].

Implementation details: Empirically, considering the spatial relationships (e.g. *left* and *right*) strongly correlate with the coordinates, we standardize all 3D scenes to a common coordinate system during training and testing, to maintain the spatial relation predicates unambiguous. Our experiments are conducted on a computational platform equipped with Intel(R) Xeon(R) CPU E5- 2680v3 @2.50 GHz CPU x2, 128G memory, and RTX 3090 GPU x8. All models are trained using a single RTX 3090 GPU.

4.2. Evaluation on 3DSSG

To verify the effectiveness of our proposed methods, we first compare our methods with other leading approaches [6,7,9] on the 3DSSG validation dataset. Evaluations are conducted in terms of object (instance), predicate, and triplet.

As shown in Table 1, the proposed Granular3D outperforms current methods by a large margin, indicating the effectiveness of comprehensive multi-granularity analysis. Notably, the Granular3D (with GaPT as encoder) demonstrates superior performance in the relation triplet prediction than all other approaches, boots the current state-of-the-art method VL-SAT by a significant improvement of +2.8 and +2.24 in Triplet A@50/100. Meanwhile, in comparison with other 3D encoders, our Granular3D+GaPT also exhibits competitive performance. As the essential component of semantic scene graphs, these results demonstrate that Granular3D achieves remarkable comprehensive performance in 3DSSG prediction. In terms of the instance classification evaluation, the proposed Granular3D particularly exhibits dominant performance, with improvements of +10.36, +7.44, and +5.96 at Object A@1/5/10 compared to the VL-SAT, respectively. This highlights the advantages of providing targeted and comprehensive granularity analysis in identifying instances within complex scenes. While in the prediction of relationships, our method falls slightly short of fully surpassing VL-SAT, with +1.42%, −0.1%, and −0.08% in Predicate A@1/3/5. This could be attributed to VL-SAT leveraging the CLIP model [43] to introduce natural language features for expressing object relationships, which bring additional performance. Meanwhile, due to the introduction of object relation feature attention in the SGG_{point} model, our Granular3D lags in the Predicate A@1 metric, while still achieving +0.57 and +0.53 at A@3/5.

Moreover, when the PointNet [5] is equally utilized as the 3D point cloud encoder, our methods (Granular3D+PointNet) already show a marked improvement, with a +0.55 increase in Triplet A@50 metrics compared to VL-SAT. The experimental results further suggest the rationality and effectiveness of the proposed architecture. Integrating Granular3D with advanced encoders like PointNet++ [11] and Point Transformer [12] leads to notable overall performance gains. Typically, in comparison to the PointNet, employing PointNet++ or Point Transformer as the encoder shows improvement of +1.62 and +1.97 on Triplet A@50 respectively, while our GaPT achieves the best performance. This highlights the importance of enhancing the ability to extract local contextual features from point clouds for multi-granularity analysis.

For a more comprehensive evaluation of our proposed method, we additionally report the results of our proposed methods under the SGCls and PredCls tasks, followed by [20]. As shown in Table 2, our proposed Granular3D significantly outperforms current methods, particularly in the SGCls task, which is considered a more challenging testing scenario [9] (+7.9, +10.5 and +12.2 in R@20/50/100 compare with the VL-SAT). In the PredCls task, which focuses solely on predicate accuracy, the proposed method also maintains a similar level to the VL-SAT, which incorporates CLIP knowledge. Once again, the results of SGCls and PredCls tasks subsequently confirm the effectiveness of our proposed methods through the innovative multi-granularity analysis.

To more intuitively demonstrate the 3DSSG prediction ability of our proposed Granular3D, we select several typical indoor scenes and demonstrate in Fig. 6, including a dining room (left part in 1st row), a laundry room (right part in 1st row), a kitchen (2nd row) and a bathroom (3rd row). The results depicted in Fig. 6 exhibit that our proposed methods maintain the efficiency of instance category identification and their internal relationship prediction across a variety of real-world indoor environments.

Remark. The experimental results with our state-of-the-art performance continuously demonstrate the effectiveness of multi-granularity analysis in 3D semantic scene graph prediction, with particularly outstanding performance in instance identification. Additionally, the comparison of various point cloud feature encoders also showcases the

Table 2
Quantitative results (%) of the SGCls and PredCls tasks, with and without graph constraints.

Model	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 [20]	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 [14]	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]	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]	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. [20]	28.5	30.0	30.1	59.3	65.0	65.3	29.8	34.3	37.0	62.2	78.4	88.3
SGFN [7]	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 [9]	32.0	33.5	33.7	67.8	**79.9**	**80.8**	33.8	41.3	47.0	**70.5**	85.0	**92.5**
Granular3D (Ours)	**43.4**	**53.0**	**59.2**	**73.2**	79.4	79.5	**41.7**	**51.8**	**59.2**	69.5	**86.1**	91.3

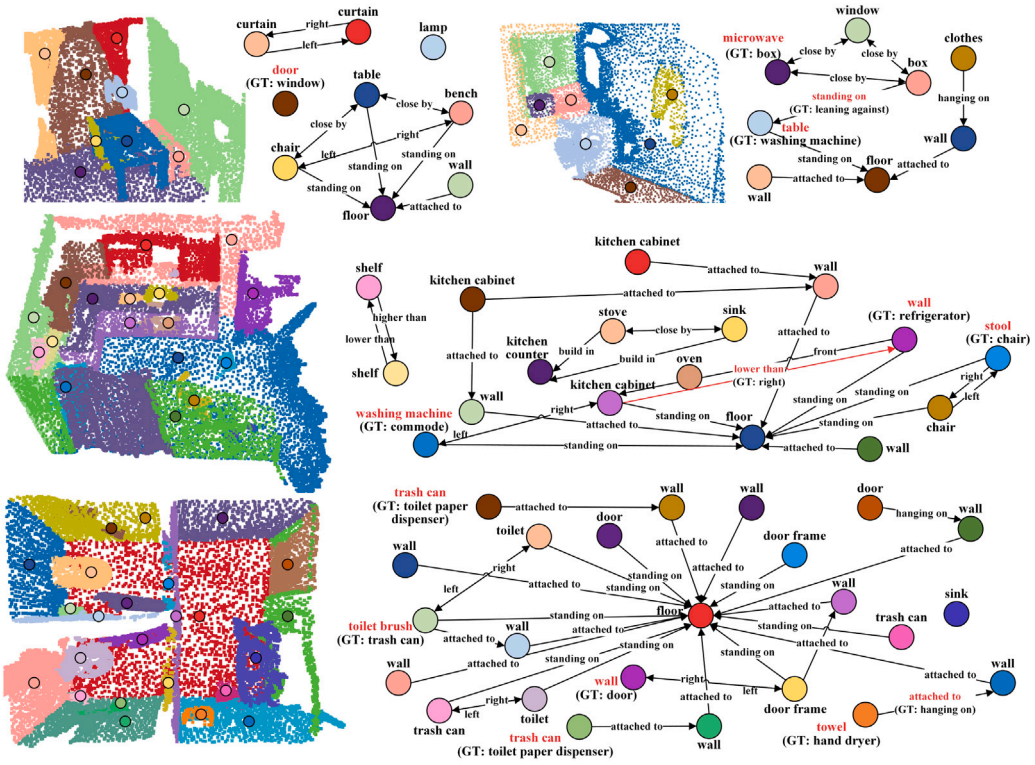

Fig. 6. Visualization of some typical 3D semantic scene graph prediction results on the 3DSSG validation set. The small colored circle represents an instance in the scene. And the red denotes misclassified instances or relationships. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Ablation study of the Adaptive Instance Enveloping Method (%).

Env.	Dist.	Object		Predicate		Triplet	
		A@1	A@5	A@1	A@3	A@50	A@100
		63.33	85.17	90.55	98.14	92.38	94.70
√		65.89	**86.55**	91.03	98.26	93.10	95.02
	√	63.81	85.11	90.81	98.29	92.45	94.77
√	√	**66.02**	86.13	**91.29**	**98.35**	**93.15**	**95.13**

Table 4
Ablation study of the down-sampling scales in our Hierarchical Dual-Stages Network (%).

Down-sampling scales	Object		Predicate		Triplet	
	A@1	A@5	A@1	A@3	A@50	A@100
[1,2,2,2; 2,2]	63.76	85.40	91.11	98.35	92.43	94.60
[1,4,4,2; 2,2] (Ours)	**66.02**	**86.13**	**91.29**	**98.35**	**93.15**	**95.13**
[1,4,4,4; 4,2]	63.50	85.60	91.26	98.29	92.42	94.65
[1,4,4,2] + [1,4,8,4]	64.83	85.39	90.54	98.26	92.93	95.02

compatibility and stability of our Granular3D, while highlighting the advantages of the proposed GaPT in extracting local contextual features at different granularities.

4.3. Ablation study

To thoroughly evaluate the individual contributions of the various components within our proposed Granular3D, we then conducted several ablation studies on the 3DSSG validation set.

Adaptive Instance Enveloping Method: We initially report the ablation study of our proposed Adaptive Instance Enveloping Method

Table 5
Ablation study of the attribute encoding in our HDSN (%).

Ins. Enc.	Rel. Enc.	Object		Predicate		Triplet	
		A@1	A@5	A@1	A@3	A@50	A@100
		65.24	86.09	83.77	95.49	91.78	94.02
√		**66.03**	86.10	82.92	95.23	91.80	94.03
	√	65.45	86.02	91.06	**98.36**	92.83	94.93
√	√	66.02	**86.13**	**91.29**	98.35	**93.15**	**95.13**

Table 6
Ablation study of weighting factors (%).

Ins. F.	Rel. F.	Object		Predicate		Triplet	
		A@1	A@5	A@1	A@3	A@50	A@100
0.1	1.0	63.25	84.90	**91.32**	98.23	91.40	93.68
0.3	1.0	**66.02**	86.13	91.29	**98.35**	**93.15**	**95.13**
0.7	1.0	65.40	**86.85**	90.30	98.04	93.08	94.99
1.0	1.0	65.87	86.66	90.61	98.03	93.00	95.02

Table 7
Ablation study of the GaPT method (%).

POS	Ins. Ga.	Rel. Ga.	Object		Predicate		Triplet	
			A@1	A@5	A@1	A@3	A@50	A@100
			65.06	85.23	90.61	98.16	92.46	94.71
√			65.13	85.27	90.74	98.15	92.57	94.82
√	√		65.91	**86.36**	90.67	98.21	93.01	95.01
√		√	65.10	85.74	**91.35**	98.25	93.00	94.94
√	√	√	**66.02**	86.13	91.29	**98.35**	**93.15**	**95.13**

Table 8
Ablation study of the channel-wise gathering method used in our GaPT (%).

Gather method	Object		Predicate		Triplet	
	A@1	A@5	A@1	A@3	A@50	A@100
Max pooling	64.58	85.53	**91.36**	98.33	92.89	95.09
Average pooling	65.21	**86.45**	91.17	98.30	93.02	95.03
Vector attention	65.11	85.71	91.30	98.34	92.93	95.04
GaPT (Ours)	**66.02**	86.13	91.29	**98.35**	**93.15**	**95.13**

Table 9
Ablation study of model efficiency.

Model	Encoder	Params. (M)	Inference time (sec)	Triplet A@50 (%)
VL-SAT [9]	PointNet	25.06	74.99	90.35
Granular3D	PointNet	**1.49**	**68.32**	90.90
	GaPT	2.08	316.56	**93.15**

Table 10
Ablation study of module design (%).

ID	Granular3D			Object		Predicate		Triplet	
	AIEM	HDSN	GaPT	A@1	A@5	A@1	A@3	A@50	A@100
①				56.79	79.51	82.30	95.31	88.51	91.54
②	√			57.01	79.83	83.56	95.27	88.97	91.71
③		√		62.59	84.64	89.24	97.88	91.92	94.25
④			√	57.77	80.06	83.44	95.38	89.06	92.00
⑤		√	√	63.33	85.17	90.55	98.14	92.38	94.70
⑥	√		√	58.49	80.46	83.96	95.39	89.13	91.93
⑦	√	√		65.06	85.23	90.61	98.16	92.46	94.71
⑧	√	√	√	**66.02**	**86.13**	**91.29**	**98.35**	**93.15**	**95.13**

(AIEM), which constructs the sub-scene with comprehensive contextual environment information and is a vital component of our proposed Granular3D. In this ablation study, we respectively examine the impact of the enveloping sampling method denoted as 'Env.', and the distance features, denoted as 'Dist.', in our AIEM. As shown in Table 3, in general, the application of our proposed AIEM significantly raises the performance with notable improvements of +2.69, +0.74, and +0.77 in Object A@1, Predicate A@1, and Triplet A@50, which illustrates the effectiveness of enriching the contextual information, especially in the recognition of instance categories. Specifically, after incorporating the enveloping sampling method, the 'Env.' component overtakes the baseline model with +2.56 and +0.48 in Object A@1 and Predicate A@1, substantiating the potency of our shape-adaptive contextual information sampling method. Moreover, the superior performance of 'Dist.' compares with the baseline model and 'Env. + Dist.' compares 'Env.' consistently affirming the value of incorporating distance features in our AIEM. Overall, the results suggest that our proposed AIEM, by providing comprehensive contextual information for the subsequent multi-granularity feature analysis, substantially improves instance identification and relationship prediction in Granular3D.

Down-sampling Scales: We then investigate the setting of down-sampling scales of our Hierarchical Dual-Stages Network (HDSN) in Granular3D, which determines the grained of point cloud feature encoding in both instance and relation stages. Each number represents the scale of a down-sampling in the hierarchical HDSN, while the ';' denotes the dividing line of the two stages. The results shown in Table 4 illustrate that the optimal performance is achieved when the point cloud is down-sampled 16 times in the instance stage, while is down-sampled 64 times in the relation stage. When the down-sampling scales are smaller, the network may not have sufficient context for the predictions of both instances and their relationships. On the contrary, the larger scales may limit the network in capturing local patterns, thus reducing the performance. Meanwhile, in addition to the gradual down-sampling across two stages, we also perform a study that splits the prediction of instance and relationship into two independent branches with the same overall down-sampling scales. And the result drops significantly, underscoring the importance of the cooperation between the two stages.

Attribute Encoding: We report the ablation study of the attribute encoding in the HDSN, which intends to complement the primitive shape information and relative spatial relationship. In this part, we comprehensively compared the impact of introducing instance-specific encoding, denoted as 'Ins. Enc.' and relation-specific encoding, denotes as 'Rel. Enc.' on model performance. The results in Table 5 demonstrate that the 'Ins. Enc.' and 'Rel. Enc.' directly influences the performance of their corresponding parts. Specifically, after integrating instance-specific encoding, the 'Ins. Enc.' gains an improvement of +0.79 in Object A@1. More obviously, after introducing relation-specific encoding, the 'Rel. Enc.' significantly boosts Predicate A@1 by +7.29. Overall, the 'Ins. Enc. + Rel. Enc.' approach outperforms the model without attribute encoding by +1.37 in Triplet A@50. In conclusion, these results highlight the importance of attribute encoding, especially in complementing relative spatial relationships to assist in predicting predicates.

Weighting Factors: We also investigate the setting of the weighting factors applied in the joint loss of our HDSN, which determines the sensitivity during training to instance identification and relationship prediction. The results are shown in Table 6, where the best performance is achieved when the instance classification loss factor, denoted as 'Ins. F.' is set to 0.3 and the relationship classification loss factor, denoted as 'Rel. F.' is set to 1. Intuitively, when 'Ins. F. diminishes, and the performance of instance identification will decrease. Conversely, when the relative proportion of 'Ins. F. increases, it will affect the accuracy of relationship prediction. This study underscores the importance of balancing the two tasks in the 3DSSG prediction task.

Gather Point Transformer: We then show the performance of our proposed Gather Point Transformer (GaPT) in Granular3D. In this ablation study, we first evaluate the POS encoding in the GaPT, denoted as 'POS'. As displayed in Table 7, the encoding without distance features is less effective than 'POS'. Additionally, we investigate the influence of applying our GaPT in both instance and relation stages, denoted as 'Ins. Ga.' and 'Rel. Ga.' respectively. Detailly, the results indicate that compared with the model using only POS encoding, after respectively introducing GaPT in the two stages, the prediction performances are correspondingly enhanced (+0.85 in Object A@1 by 'Ins. Ga.', and +0.61 in Predicate A@1 by 'Rel. Ga.'), which reflects the positive impact of the GaPT. Moreover, the complete integration of GaPT leads

to a further enhancement of +0.69 in Triplet A@50 compared to the baseline model. These findings, as summarized in Table 7, conclude that the design of our proposed GaPT excels in perceiving the sub-scene in the hierarchical network by establishing the combinatorial interaction between multiple local sets.

Channel-wise Gathering Method: We also study the channel-wise gathering method used in our Gather Point Transformer (GaPT). Initially, we compare our method with basic feature fusion methods: Max Pooing and Average Pooling. As illustrated in Table 8, compared with the channel-wise MLP used in our GaPT, the performance declines by −0.26 and −0.13 in Triplet A@50 when using Max and Average Pooling, respectively. On the other hand, we further apply the vector attention to gather the features from multiple local sets, which also results in a performance decrease of −0.22 in Triplet A@50.

Method Efficiency: Subsequently, we investigate the efficiency of our proposed Granular3D. As displayed in Table 9, our method showcases a significant advantage, with a reduction of over 10 times in parameter consumption compared to the advanced VL-SAT, which contains a GCN network with over 20M parameters, illustrating the benefit of allocating resources for the critical multi-granularity 3D point cloud feature extraction. However, Granular3D (with GaPT as encoder) exhibits diminished inference efficiency on the 3DSSG validation set. The primary reason for this situation may be the complex encoder structure. Therefore, we further present the time consumption when utilizing the same PointNet encoder as VL-SAT. Notably, Granular3D+PointNet then surpasses VL-SAT in both speed and accuracy. Overall, this study demonstrates that the balance between accuracy and efficiency is also worth further exploration in our future research endeavors.

Module Design: Finally, we summarize the influence of model performance after introducing our proposed methods. As shown in the experiments ①–④ of Table 10, the incorporation of each individual component within Granular3D consistently yields positive effects on the baseline model. Particularly noteworthy is the Hierarchical Dual-Stages Network (HDSN), which contributes the most significant improvement (+3.41 in Triplet A@50 compared to the baseline), highlighting the benefits of multi-granularity feature extraction in the complex 3D indoor environment. This trend continues in experiments ⑥ and ⑧, where HDSN distinctly improves the prediction of instances and relationships in their respective stages, posting a significant positive impact (+4.02 in Triplet A@50). Meanwhile, comparing experiments ⑤ and ⑧ allows us to assess how our Adaptive Instance Enveloping Method (AIEM) enhances instance prediction by +2.69 in Object A@1, demonstrating its effectiveness in providing sufficient contextual environmental information for multi-granularity learning. Lastly, the comparison between experiments ⑦ and ⑧ indicates that the introduction of our Gather Point Transformer (GaPT) further boosts overall performance by facilitating the combinatorial interaction between local sets during encoding, resulting in a +0.69 improvement in Triplet A@50.

Remark. Table 10 demonstrates the effectiveness of the components within the proposed Granular3D. Due to superior environment information construction, network architecture, and contextual feature extraction, Granular3D achieves comprehensive multi-granularity analysis, thereby ultimately benefiting the overall 3DSSG prediction performance. Among all the proposed components, the HDSN, which is directly related to multi-granularity analysis, exhibits the most significant improvement in the experimental results, further underscoring its importance.

4.4. Discussion and limitation

While our proposed Granular3D achieves optimal performance in the 3D Scene Graph Prediction task and significantly enhances instance identification, it does not exhibit outstanding performance in predicting inter-instance relationships.

Specifically, Granular3D provides detailed multi-granularity analysis, extracting potential relationships between instance pairs at a targeted stage within the hierarchical network, resulting in competitive predicate prediction performance compared to current state-of-the-art methods. However, the complexity of indoor environments leads to the coexistence of many similar relationships between instance pairs, such as “standing on”, “leaning against” and “attached to” in the 3DSSG dataset. The discernment of these subtle differences may limit the performance of our method, as shown in the visualization of Fig. 6. On the other hand, as reported in Table 9, our Granular3D also has certain deficiencies in the balance between accuracy and efficiency.

Therefore, after comprehensive multi-granularity point cloud feature extraction, achieving the distinction of approximate inter-instance relationships may require additional effort. For instance, designing targeted loss functions, employing extra attention mechanisms, or incorporating natural language prior features could be considered. Thus, exploring more effective methods for predicting relationships between object pairs while maintaining efficiency should be a meaningful research direction for our further work.

5. Conclusion

In this paper, we introduce the Granular3D as a novel approach to overcome the limitations of existing 3D Semantic Scene Graph (3DSSG) prediction methods, particularly in the multi-granularity point cloud feature extraction. The Granular3D comprises an Adaptive Instance Enveloping Method (AIEM) a Hierarchical Dual-Stages Network (HDSN), and a Gather Point Transformer structure. These approaches cover the sub-scene construction, network architecture, and local point cloud feature extraction of the multi-granularity analysis, improving the 3D environment perception and further boosting the overall 3DSSG prediction performance by +2.8% in the top-50 triplet accuracy. Our comprehensive experimental analysis demonstrates that our proposed methods achieve new state-of-the-art performance in the 3DSSG dataset. These results not only validate the effectiveness of our Granular3D but also underscore the critical role of multi-granularity point cloud feature extraction in 3DSSG prediction.

CRediT authorship contribution statement

Kaixiang Huang: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jingru Yang:** Writing – original draft, Supervision, Methodology, Data curation. **Jin Wang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Shengfeng He:** Writing – review & editing, Supervision. **Zhan Wang:** Writing – review & editing, Supervision, Funding acquisition. **Haiyan He:** Supervision, Resources. **Qifeng Zhang:** Writing – original draft, Investigation, Data curation. **Guodong Lu:** Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank the editors and reviewers for their comments, which led to the improvement of this paper. This work is supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2023C01180, 2023C01070), National Natural Science Foundation of China (52175032), Yuyao Science and Technology project (No. 2023J H03010019), Robotics Institute of Zhejiang University under Grant K11808 and K11811.

References

[1] K. Huang, Y. Han, J. Wu, F. Qiu, Q. Tang, Language-driven robot manipulation with perspective disambiguation and placement optimization, IEEE Robot. Autom. Lett. 7 (2) (2022) 4188–4195.

[2] J. Wald, H. Dhamo, N. Navab, 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, pp. 3961–3970.

[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (2017) 32–73.

[4] D. Xu, Y. Zhu, C.B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5410–5419.

[5] C.R. Qi, H. Su, K. Mo, 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, pp. 652–660.

[6] C. Zhang, J. Yu, Y. Song, 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, pp. 9705–9715.

[7] S.-C. Wu, J. Wald, K. Tateno, N. Navab, 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, pp. 7515–7525.

[8] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, ICLR, 2017, pp. 1–14.

[9] Z. Wang, B. Cheng, L. Zhao, D. Xu, Y. Tang, 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, pp. 21560–21569.

[10] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, A. Markham, Randla-net: Efficient semantic segmentation of large-scale point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11108–11117.

[11] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, Adv. Neural Inf. Process. Syst. 30 (2017).

[12] H. Zhao, L. Jiang, J. Jia, P.H. Torr, V. Koltun, Point transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16259–16268.

[13] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, J. Jia, Stratified transformer for 3D point cloud segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8500–8509.

[14] T. Chen, W. Yu, R. Chen, L. Lin, Knowledge-embedded routing network for scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6163–6171.

[15] K. Tang, H. Zhang, B. Wu, W. Luo, W. Liu, Learning to compose dynamic tree structures for visual contexts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6619–6628.

[16] A. Zareian, S. Karaman, S.-F. Chang, Bridging knowledge graphs to generate scene graphs, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, Springer, 2020, pp. 606–623.

[17] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: Scene graph parsing with global context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5831–5840.

[18] M. Zhao, Y. Kong, L. Zhang, B. Yin, Class correlation correction for unbiased scene graph generation, Pattern Recognit. (2023) 110221.

[19] Z. Wang, X. Xu, Y. Luo, G. Wang, Y. Yang, Hypercomplex context guided interaction modeling for scene graph generation, Pattern Recognit. 141 (2023) 109634.

[20] S. Zhang, A. Hao, H. Qin, Knowledge-inspired 3d scene graph prediction in point cloud, Adv. Neural Inf. Process. Syst. 34 (2021) 18620–18632.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[22] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).

[23] C. Choy, J. Gwak, S. Savarese, 4D spatio-temporal convnets: Minkowski convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3075–3084.

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

[25] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, B. Ghanem, Pointnext: Revisiting pointnet++ with improved training and scaling strategies, Adv. Neural Inf. Process. Syst. 35 (2022) 23192–23204.

[26] M. Ye, S. Xu, 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, pp. 1631–1640.

[27] H. Shao, J. Bai, R. Wu, J. Jiang, H. Liang, FGPNet: A weakly supervised fine-grained 3D point clouds classification network, Pattern Recognit. 139 (2023) 109509.

[28] B. Lu, Y. Sun, Z. Yang, R. Song, H. Jiang, Y. Liu, HRNet: 3D object detection network for point cloud with hierarchical refinement, Pattern Recognit. (2024) 110254.

[29] X. Zou, J. Li, Y. Wang, F. Liang, W. Wu, H. Wang, B. Yang, Z. Dong, PatchAugNet: Patch feature augmentation-based heterogeneous point cloud place recognition in large-scale street scenes, ISPRS J. Photogramm. Remote Sens. 206 (2023) 273–292.

[30] J. Choe, C. Park, F. Rameau, J. Park, I.S. Kweon, Pointmixer: Mlp-mixer for point cloud understanding, in: European Conference on Computer Vision, Springer, 2022, pp. 620–640.

[31] H. Thomas, C.R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, L.J. Guibas, Kpconv: Flexible and deformable convolution for point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6411–6420.

[32] H. Zhao, L. Jiang, C.-W. Fu, J. Jia, Pointweb: Enhancing local neighborhood features for point cloud processing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5565–5573.

[33] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[37] J. Yang, Q. Zhang, B. Ni, L. Li, J. Liu, M. Zhou, Q. Tian, Modeling point clouds with self-attention and gumbel subset sampling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3323–3332.

[38] N. Kruithof, G. Vegter, Envelope surfaces, in: Proceedings of the Twenty-Second Annual Symposium on Computational Geometry, 2006, pp. 411–420.

[39] J. Sankaranarayanan, H. Samet, A. Varshney, A fast k-neighborhood algorithm for large point-clouds, in: PBG@ SIGGRAPH, 2006, pp. 75–84.

[40] S. Qiu, S. Anwar, N. Barnes, Pnp-3d: A plug-and-play for 3d point clouds, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2021) 1312–1319.

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

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

[43] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

Kaixiang Huang received the 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 with the School of Zhejiang University, Hangzhou, China. His research interests include 3D computer vision and human–robot interaction.

Jingru Yang received the 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 with the School of Zhejiang University, Hangzhou, China. His research interests include 2D and 3D computer vision.

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

Shengfeng He is an associate professor of Singapore Management University. He serves as the lead guest editor of IJCV, the associate editor of IEEE TNNLS, Visual Intelligence, and Neurocomputing. He also serves as the area chair/senior program committee of ICML, AAAI and IJCAI. His research interests include computer vision.

Zhan Wang received the B.Sc. degree from Wuhan university of technology, China, in 2012, the M. Eng. degree from university of the Chinese Academy of Sciences, China, in 2015 and the Ph.D. degree of robotics from University of Paris-Saclay, France, in 2018. His research interests include robotic localization and computer vision.

Haiyan He received her B.Eng. degree in Material Science and Engineering from Zhejiang University, China, in 2007, and received the Ph.D. degree in Material Science and Engineering from Zhejiang University, China, in 2013. Her research interests include the Photovoltaic system and intelligent operation and maintenance.

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

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 with Zhejiang University, Hangzhou, China. His research interests are CAD, CG, and robotics.