Complete 3D Relationships Extraction Modality Alignment Network for 3D Dense Captioning

Aihua Mao, Zhi Yang, Wanxin Chen, Ran Yi, Yong-jin Liu, *Senior Member, IEEE*

Abstract—3D dense captioning aims to semantically describe each object detected in a 3D scene, which plays a significant role in 3D scene understanding. Previous works lack a complete definition of 3D spatial relationships and the directly integrate visual and language modalities, thus ignoring the discrepancies between the two modalities. To address these issues, we propose a novel complete 3D relationship extraction modality alignment network, which consists of three steps: 3D object detection, complete 3D relationships extraction, and modality alignment caption. To comprehensively capture the 3D spatial relationship features, we define a complete set of 3D spatial relationships, including the local spatial relationship between objects and the global spatial relationship between each object and the entire scene. To this end, we propose a complete 3D relationships extraction module based on message passing and self-attention to mine multi-scale spatial relationship features and inspect the transformation to obtain features in different views. In addition, we propose the modality alignment caption module to fuse multi-scale relationship features and generate descriptions to bridge the semantic gap from the visual space to the language space with the prior information in the word embedding, and help generate improved descriptions for the 3D scene. Extensive experiments demonstrate that the proposed model outperforms the state-of-the-art methods on the ScanRefer and Nr3D datasets.

Index Terms—3D dense captioning, multi-modal learning, 3D spatial relationship, modality alignment.

1 Introduction

DEEP learning has achieved great progresses in computer vision and graphics in the last decade, including the field of object detection and scene understanding. Particularly, scene understanding performs higher level functions than object recognition in analyzing objects in context with respect to the spatial and semantic relationships between objects of the scene. The rapid research progress has further promoted the cross-modal learning tasks of vision and language in the fields of visual question answering and semantic captioning. However, most of the current works are based on 2D image data, and the cross-modality tasks of 3D scenes and language are a new direction and have been rarely explored.

Recently, 3D scene understanding becomes a crucial research topic for robotics, augmented reality and autonomous vehicles. Cross-modal tasks for 3D scenes semantic captioning have gained great attention with the creation of datasets that express 3D object localization via natural language, such as ScanRefer [1] and Nr3D [2]. Chen et al. [1] were the first to develop *3D visual grounding*, which takes two separate modalities (i.e., point cloud and language expression) as input to locate and identify target objects in 3D scenes. Users can not only locate objects in 3D scenes, but also generate concrete descriptions. The researchers then introduced the *3D dense captioning* task [3], which takes the point cloud of a 3D scene as input, performs 3D object detection and generates natural language description for each detected object. 3D dense captioning offers a meticulous semantic description of individual objects in a 3D scene and has significant value for 3D scene understanding.

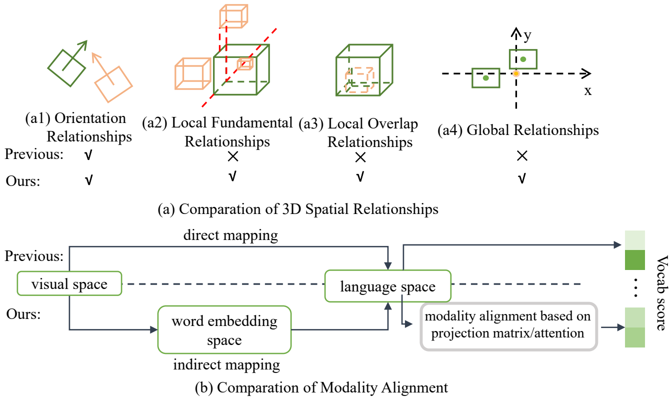

(a1) Orientation Relationships
(a2) Local Fundamental Relationships
(a3) Local Overlap Relationships
(a4) Global Relationships
y
x
Previous: √ × × ×
Ours: √ √ √ √
(a) Comparation of 3D Spatial Relationships
direct mapping
Previous:
visual space
language space
Ours:
word embedding space
modality alignment based on projection matrix/attention
indirect mapping
Vocab score
(b) Comparation of Modality Alignment

Fig. 1. (a) Comparison of 3D spatial relationships: Previous methods [3], [4] only extract orientation relationships (a1), while we define a complete set of 3D spatial relationships (a1-a4), which increases the authenticity of the description. (b) Comparison of modality alignment: Previous methods directly maps from the visual space to language space, while we leverage the prior information in the word embedding to bridge the inter-modality discrepancies.

The captioning task requires the collection of high-level semantic information from visual data (such as images and point clouds) and the creation of one or more natural language sentences to automatically describe the scene. Given the sparse, irregular, and disordered nature of point clouds,

- *A. Mao, Z. Yang, W. Chen are with School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. E-mails: ahmao@scut.edu.cn, 202021044575@mail.scut.edu.cn, adrien.chenwx@gmail.com.*
- *R. Yi is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. Email: ranyi@sjtu.edu.cn.*
- *Y.-J. Liu is with BNRist, MOE-Key Laboratory of Pervasive Computing, Department of Computer Science and Technology, Tsinghua University, Beijing, China. E-mail: liuyongjin@tsinghua.edu.cn.*

Corresponding authors: Aihua Mao and Ran Yi

3D dense captioning is more challenging than image captioning, which deals with regular image data, because it is necessary to not only generate a description for each object, but also deal with disordered point cloud data and solve more complicated spatial relationships. Scan2Cap [3] and X-Trans2Cap [4] are the pioneering works of deep learning network for 3D dense captioning. However, the previous methods suffer from the following weaknesses: (1) these methods only extract orientation relationships (shown in Fig.1(a1)) and lack the complete definition of 3D spatial relationships; (2) these methods directly integrate the visual and language modalities (shown in the upper part of Fig.1(b)), ignoring the discrepancies between the two modalities; (3) Transformer can gather object features well but needs to learn many parameters during the training phase.

To address these issues, we propose a novel 3D dense captioning approach with a complete 3D relationships extraction modality alignment network (REMAN) that enhances the accuracy of description sentence generation. REMAN consists of a 3D object detection backbone, a complete 3D relationships extraction module (REM) and a modality alignment caption module (MACM). First, we define a complete set of 3D spatial relationships (illustrated in Fig.1(a)), including orientation relationships, local fundamental, local overlap, and global relationships. Second, we propose REM to extract both local (between objects) and global (between each object and the global scene) relationship features, which increases the authenticity and diversity of the description. Third, we propose MACM to fuse the relationship features and generate descriptions, which leverages the prior information in the word embedding to bridge the semantic gap from the visual to the language space (the lower part of Fig.1(b)). Notably, the proposed modality alignment approach does not increase the parameters of network learning, allowing it to be easily applied to various tasks.

To completely evaluate the performance gain of REMAN, we conduct extensive experiments on the ScanRefer [1] and Nr3D [2] datasets, training them in both end-to-end and non end-to-end approaches. The results of the BLEU-4 [5], CIDEr [6], ROUGE [7], and METEOR [8] metrics on different datasets show the effectiveness of our proposed approach, when compared with the state-of-the-art methods.

In summary, our main contributions are threefold:

- We propose a novel point cloud-based 3D dense captioning network (REMAN) that can accurately learn the local and global relationships of objects in a 3D scene.
- To better locate the target object in a 3D scene, we define a comprehensive set of 3D spatial relationships that consider both local and global relationships and propose REM to extract various relationship features.
- We propose MACM for natural language description generation, which bridges the semantic gap between the vision and language modalities by leveraging the prior information in the word embedding and improves the captioning without increasing the parameters.

2 Related Works

2.1 Image Captioning

Although the research on the 3D scene semantic caption generation task has just begun, many solid works [9], [10], [11], [12], [13], [14], [15] on image captioning have been conducted. Anderson et al. [16] proposed a bottom-up and top-down attention mechanism based on the faster region-CNN [17] object detection framework, which directs the network to encode visual features and output captions through LSTM [18]. To further encode the relationship between regions, Li et al. proposed ReGAT [19] learn the relationship features by constructing the region relation graph. With the popularity of the Transformer [20] architecture, several image captioning models have begun to use the attention mechanism. Huang et al. [21] augments the traditional self-attention mechanism by addressing irrelevant attention concerns. Cornia et al. [22] employs mesh connection to learn low- and high-level features in the decoding step, using learnt prior knowledge to model the multilevel representation of the relationship between regions. To selectively use visual information and execute multimodal reasoning, Pan et al. proposed X-LAN [23]and it uses a bi-linear pooling module.

With the development of pre-training models like BERT [24] in natural language processing, many vision-language models with pre-training and fine-tuning strategies have been proposed, such as VL-BERT [25], ViL-BERT [26], and LXMERT [27], which can be applied to image captioning.

However, all the works above focus on image data and are difficult to be applied for the 3D scene. Our work turns the focus on the semantic captioning of 3D cloud point and further explores the reasoning of 3D spatial relationships.

2.2 3D Scene Captioning

With the easy acquisition of 3D data thanks to the development of sensor technology, an increasing number of researchers are shifting their attention from 2D vision to 3D vision and attempt to utilize the language information, solving tasks such as 3D question answering [28], [29], [30], 3D Captioning [3], [31], [32], and 3D Visual Grounding [1], [2], [33], [34], [35]. Chen et al. [3] first introduced the 3D dense captioning task, which aims to create dense descriptions for a 3D scene, namely, one semantic sentence for each detected object. The overall task is similar to that of image captioning, except that the visual information is derived from 3D data. Subsequently, Scan2cap [3] is designed for the generation of 3D scene dense captioning for point cloud. It consists of three primary modules: object detection, spatial relationship graph reasoning, and context-aware caption generation. The network can output the detected object bounding boxes and the related natural language caption. X-Trans2Cap [4] is another recent work, which is based on Transformer and the teacher–student design in [36]. During the training step, the teacher network receives input of both 2D and 3D features and guides the student network using solely 3D features. These works begin the research for 3D dense captioning, however, are lacking in the definition of 3D spatial relationships of objects in a 3D scene. Thus, we define a comprehensive set of spatial relationships and propose a

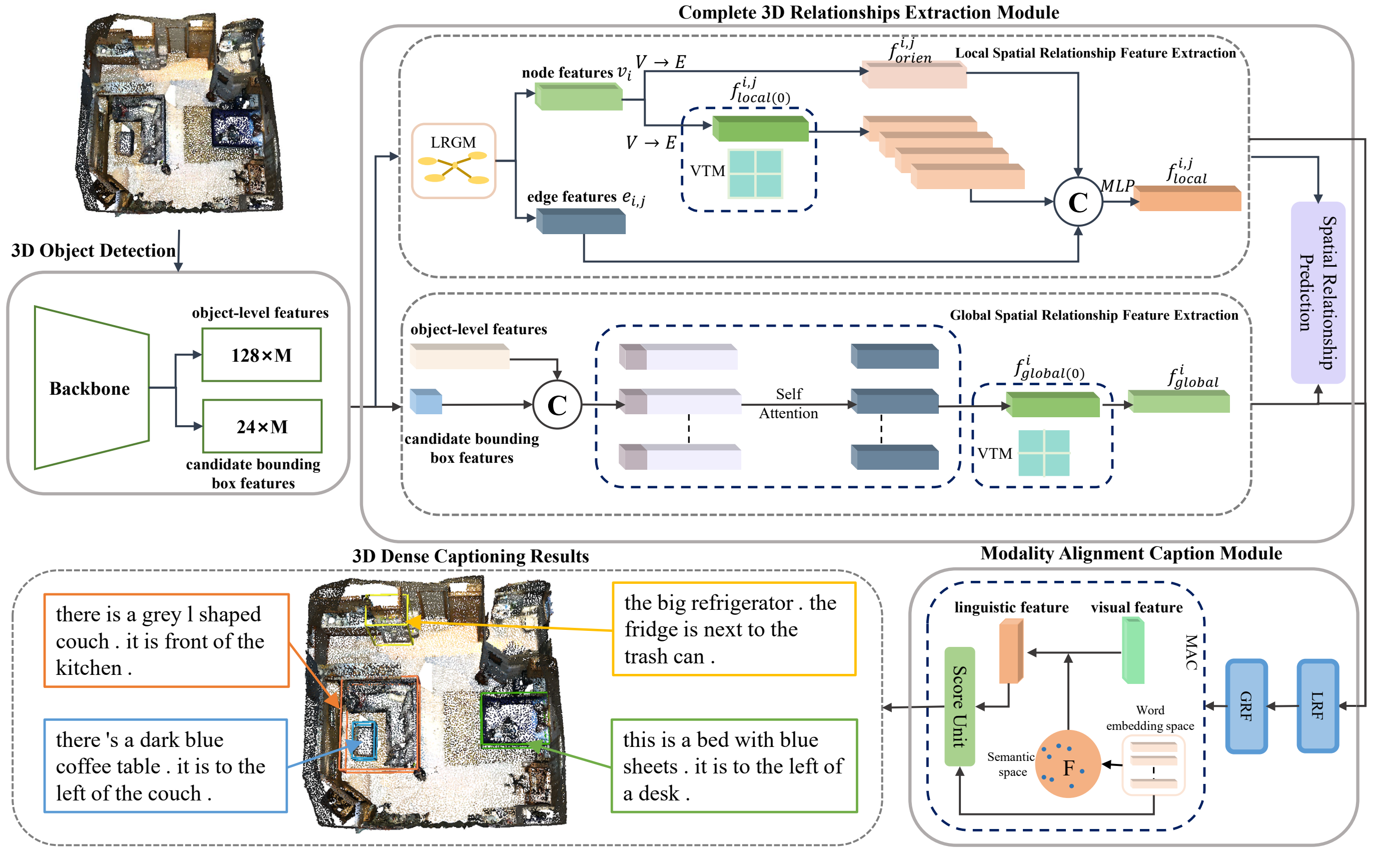

Complete 3D Relationships Extraction Module
Local Spatial Relationship Feature Extraction
node features v_i
$V \rightarrow E$
$f^{i,j}_{orien}$
$f^{i,j}_{local(0)}$
LRGM
$V \rightarrow E$
VTM
edge features $e_{i,j}$
MLP
C
$f^{i,j}_{local}$
Spatial Relationship Prediction
3D Object Detection
Backbone
object-level features
128×M
24×M
candidate bounding box features
Global Spatial Relationship Feature Extraction
object-level features
C
candidate bounding box features
Self Attention
$f^{i}_{global(0)}$
VTM
f^{i}_{global}
3D Dense Captioning Results
there is a grey l shaped couch . it is front of the kitchen .
there 's a dark blue coffee table . it is to the left of the couch .
the big refrigerator . the fridge is next to the trash can .
this is a bed with blue sheets . it is to the left of a desk .
Modality Alignment Caption Module
linguistic feature
visual feature
MAC
Score Unit
Word embedding space
Semantic space
F
GRF
LRF

Fig. 2. An overview of our REMAN. First, given the point cloud of a 3D scene, the 3D object detection backbone extracts object-level features and candidate bounding box features. Then the two types of features are fed into the Complete 3D Relationships Extraction Module to extract a newly defined complete set of 3D spatial relationships. The extracted relationship features are then sent into the Modality Alignment Caption Module, which leverages the prior information in the word embedding space to bridge the semantic gap between visual and linguistic features and generates a description for each detected object.

REM to extract various relationship features, which increase the authenticity and diversity of the description.

2.3 Visual Relationship Learning

In vision-language multi-modal tasks, the spatial relationship between objects is extremely crucial. ReGAT [19] is a graph convolution model that has been applied to visual question answering. This model creates an object relationship graph using 11 different types of relative position relationships and adopts a spatial relationship encoder to model them. Kant et al. [37] also developed an object relationship graph, but learned the relationship using the multi-head attention mechanism in the Transformer architecture, with each attention head responsible for learning a distinct type of relationship. Luo et al. [38] proposed an image captioning method based on the Transformer architecture. This method does not generate an object relationship graph individually, but encodes the relative and absolute positions of objects instead and then uses the attention mechanism to perform fusion learning. Dual-GCN [39] captures both the object-to-object spatial relation in a single image, and the feature information provided by related images. Dong et al. [40] also considered rich spatial relationship information, including the Intersection over Union (IoU) and relative orientation angle among the objects, when constructing the object relationship graph. By improving the learning of these auxiliary relationships, the extracted relationship features are highly accurate in describing the spatial relationship. GPaS [41] constructs a graph from semantic words and apply GCN [42] and LSTM [18] for dense video caption. Different from the above methods, our model focuses on complete spatial relationship modeling of 3D scenes.

2.4 Feature Alignment

Feature alignment is an important work for cross-modality tasks of vision and language, such as image-text retrieval and image-text matching. It usually needs to transform the extracted features to align the visual and linguistic features, and eliminate the enormous gaps in feature distribution caused by different modalities. TIRG [43] considers the inter-modality gaps between visual and linguistic features, projects them into the same feature space, and learns two types of gated and residual features to improve the feature fusion. Oscar [44] represents each input image-text as a triplet (w, q, v) of words, category labels, and object feature. Then, the category label q is employed as an anchor to help the visual features in aligning the regions and the text. CAMERA [45] first computes the cosine similarity between the visual features of each view in the image and the linguistic features of the caption, and then adopts triple loss to reduce the gap between visual and linguistic features with the same semantic meaning, thereby improving feature alignment. MFA [46] aligns regions and sentences in word and phrase level. EMAF [47] designs the alignment module based on the improved dynamic routing algorithm and has achieved competitive results in fake news detection. Different from the above methods, our method bridges the gap between the visual features extracted from 3D scenes and the linguistic features by leveraging the prior information in the word embedding.

3 METHOD

In this section, we introduce our complete 3D REMAN for 3D dense captioning. Given the input data (point cloud), our model follows the steps of 3D object detection and generation of semantic description for each detected object. As shown in Fig.2, the pipeline consists of three steps. (1) First, we feed the input point cloud data into a 3D object detection backbone such as Votenet [48] to obtain object-level features $f_{obj} \in \mathbb{R}^{128\times M}$ and the corresponding bounding boxes $\in \mathbb{R}^{24\times M}$ where M is the numbers of proposals, which will be leveraged to extract spatial relationship features and generate descriptions. (2) Then, we define a complete set of spatial relationships suitable for 3D scenes and propose the complete 3D REM to model local and global spatial relationships. Considering that the relationship changes with the view, we propose a view transformation module (VTM) to obtain features under different views. (3) Finally, considering the gaps between vision and language, we propose two modality alignment classifiers in MACM to fuse multi-scale features and bridge inter-modality gaps, allowing for a more seamless transition from visual to linguistic features.

3.1 Definition of 3D spatial relationships

The orientation relationships in Scan2Cap [3] illustrated in Fig.1(a1) only describe the angular deviations between two objects. In this section, we propose the following three new 3D spatial relationships to form a complete set of 3D spatial relationships.

Local Fundamental Relationships. The detected bounding boxes enable the objects in the scene to establish six fundamental relative relationships: up, down, left, right, front, and back. Fig.1(a2) shows the 3D schematic of the neighboring left, upper, and front bounding boxes (red) for the green target bounding box.

Local Overlap Relationships. Additionally, to address the coverage phenomena, we define two types of local overlap relationships, namely, inside and coverage. As shown in Fig 1(a3), two objects in the scene are marked by green and red bounding boxes in which the green one covers the red one, and the red one is inside the green one. Two solid 3D objects will not intersect, so no intersection relationship exists.

We utilize a 0-1 discrete vector $y^{i,j}_{local} \in \mathbb{R}^8$ to represent the local fundamental and overlap relationships between the i-th and the j-th objects. The first six dimensions indicate the fundamental orientation relationships, while the last two dimensions indicate the overlap relationships. If $y^{i,j,k}_{local} = 1$, then the two objects have the k-th spatial relationship.

Global Spatial Relationship. Simply employing local spatial relationships is insufficient, since a scene has a wealth of global information that can be helpful for description generation, such as “a table in the corner of the room” or “a cabinet on the left side of the room’s entrance.” To improve the spatial relationship representation in the 3D scene, we further define the relationship between the object and the entire scene. As seen in Fig.1(a4), for the entire scene, considering the $x-y$ plane’s global relationship is sufficient to augment the description. The relationship between them can be classified into four categories on the basis of the center position relationship, namely, front, back, left, and right. We represent the relationship between the i-th object and global scene as a 0-1 discrete vector $y^{i}_{global} \in \mathbb{R}^4$.

3.2 Complete 3D Relationships Extraction Module

Given that the object detection features extracted from the 3D object detection backbone only include object-level information and exclude spatial relationship features, we propose the complete 3D REM to extract the 3D spatial relationships. To enrich the properties of the local spatial relationship between objects, we apply graph convolution network to extract the eight local relative relationships defined in Section 3.1 and apply the self-attention mechanism to the model global spatial relationship features between objects and scenes.

3.2.1 *View Transformation Module*

The visual relationship between two objects is difficult to determine without a stable view because the visual relationship changes with the view. To address this issue, we propose a view transformation module (VTM) to obtain features in different views and use this module in both local and global spatial relationship feature extraction. Our VTM employs a 2D mesh interpolation approach. Given that the transformation of the view points occurs only in the $x-y$ plane, we mainly address four distinct views: left, right, front, and back.

As shown in Fig.3, for each object, the $1 \times D$ input visual feature vector is first duplicated, with the number of copies equals to the number of view types. In our case (the number of view types is 4), the dimension of the output features is $4 \times D$. (2) Second, this approach employs a standard 2D mesh to represent four distinct visual views, and utilizes the 2D mesh to guide the visual features to generate spatial relationship features from several views. The 2D mesh is a 2×2 grid and each with a 2-bit code, which can simply express ”00 01 10 11” from left top to right bottom, as shown in Fig. 3. Finally, visual features are concatenated with the 2D mesh following the way in FoldingNet [49], which uses a fixed 2D mesh to distinguish points in a point cloud. And then we feed them into a FC layer for fusion and dimensional alignment.

We consider the construction of visual features during view transformation to be a form of feature expansion and introduce a 2D mesh to accomplish this task. In our approach, the feature representations under different views can be sufficiently distinctive, and the entire network can ultimately learn visual features independent of views because the visual features utilized for stitching are consistent, e.g, object A is on the left of object B in the front view, while object A is on the right to B in the back view.

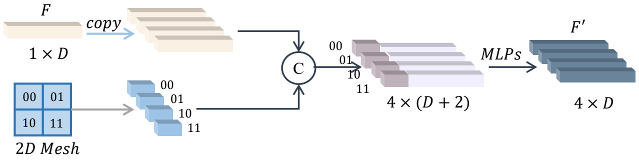


Fig. 3. View Transformation Module (VTM) for obtaining features in different views.

3.2.2 Local Spatial Relationship Feature Extraction

First, given the output candidate bounding boxes $\in \mathbb{R}^{24\times M}$ from 3D object detection backbone (where M is the number of bounding boxes), we compute the adjacency matrix and construct the spatial relation graph $G(V, E)$. The nodes of the graph are initialized as the candidate object features $\{v_i\}_{i=1}^{M} \in \mathbb{R}^{d\times M}$ (where d indicates the dimension of the candidate object features and is set to 128). The edge features of the graph $\{e_{i,j}\}_{i=1,j=1}^{M,M} \in \mathbb{R}^{d\times M\times M}$ are initialized by using the K nearest neighbor node features of the current node.

We adopt the local relational graph module (LRGM) proposed in Scan2Cap [3] which is effective in extracting node and edge features, to obtain the local spatial features (Fig.4). LRGM uses the traditional message passing mechanism to enhance node features and extract local spatial features. Note that local spatial relationship here refers to the relative distance between two objects, but doesn't include orientation and overlap information. The module consists of two graph convolution layers. And the graph convolution operations from node to edge and edge to node can be stated as follows:

$$\begin{aligned} V \rightarrow E &: e_{i,j}^{\tau+1} = f^{\tau}\left(\left[v_i^{\tau}, v_j^{\tau} - v_i^{\tau}\right]\right) \\ E \rightarrow V &: v_i^{\tau+1} = \sum_{j=1}^{K} e_{i,j}^{\tau+1} \end{aligned} \tag{1}$$

where $v_i^{\tau} \in \mathbb{R}^{128}$ and $v_j^{\tau} \in \mathbb{R}^{128}$ represent the node features of the i-th and j-th nodes in the τ-th iteration, respectively, and $e_{i,j}^{\tau+1} \in \mathbb{R}^{128}$ represents the features transmitted from the j-th node the to i-th node in the $(\tau+1)$-th iteration of the graph. $f^{\tau}(\cdot)$ is a non-linear function that can be learned, and it is implemented by MLP. The final node and the edge features are produced after multiple iterations of graph convolution.

The final node features v_i are then input into two other graph convolution layers ($V \rightarrow E$), and the outputs of the two layers are the orientation angle feature $f_{orien}^{i,j}$ and local spatial relationship feature $f_{local(0)}^{i,j}$. $f_{local(0)}^{i,j}$ is further sent into the VTM (Section 3.2.1), and the output features are concatenated with edge features $e_{i,j}$ and $f_{orien}^{i,j}$ to obtain the final enhanced local spatial relationship features $f_{local}^{i,j}$.

3.2.3 Global Spatial Relationship Feature Extraction

For global spatial relationships, we propose a self-attention module which employs non-local operations [50] to extract the relationship between the i-th object and the global scene. The input of this module is the i-th object feature $f_{obj}^{i} \in \mathbb{R}^{128}$ and its corresponding bounding box features $p_i \in \mathbb{R}^{24}$ output from the object detection stage. We concatenate the

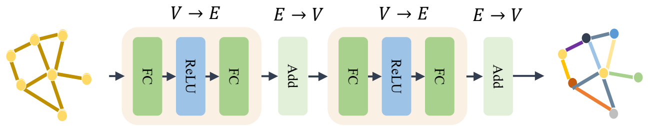


Fig. 4. Local Relational Graph Module (LRGM) for enhancing local spatial features in Local Spatial Relationship Feature Extraction.

object feature and the corresponding bounding box features to obtain feature vector $x_i \in \mathbb{R}^{(128+24)}$. Then, the enhanced object feature ${f_{obj}^{i}}' \in R^{128}$ is obtained by feeding feature vector x_i into the self-attention module:

$${f_{obj}^{i}}' = \sum_{\forall j} f\left(\theta\left(x_i\right), \varphi\left(x_j\right)\right) g\left(x_j\right) \tag{2}$$

where $\theta(\cdot), \varphi(\cdot)$ and $g(\cdot)$ are activation functions, and $f(\cdot,\cdot)$ is the dot-product similarity of the two feature vectors. Finally, the spatial relationship is predicted by spatial relationship prediction (SRP, described in 3.2.4) according to the generated features as an auxiliary task.

The non-local operation is a non-local weighted average computation, which calculates the similarity between a given feature vector and all other feature vectors and uses the similarity as a weight. Then, the sum of all feature vectors is obtained using weighted summation. This computation approach can capture long-distance dependencies and is highly suitable for extracting the global scale feature representation. We then feed the enhanced object feature ${f_{obj}^{i}}'$ output from the attention module into a FC layer and obtain the global spatial relationship feature $f_{global(0)}^{i}$, which is then sent to the VTM (Section 3.2.1) to obtain the final global relationship features f_{global}^{i}.

3.2.4 Spatial Relationship Prediction

At the end of this module, we extend the auxiliary orientation prediction module proposed in Scan2cap [3] and propose the SRP module. The SRP module predicts spatial relationships from the extracted spatial relationship features to assist the learning of feature extraction. In this subtask, according to extracted relationship features f_{orien}, f_{local} and f_{global}, we set up three prediction heads on the basis of MLPs to predict the relative orientation angle, local spatial relationship, and global spatial relationship, respectively. Note that we predict local spatial relationship and calculate ground truth for each view (front, back, left, right). The predicted local and global spatial relationships under different views are respectively divided into eight and four categories according to the spatial relationship defined in Section 3.1. As for the relative orientation angle between two objects, given that the relative orientation angle between objects ranges from 0° to 180°, we divide it into six intervals, each of which is 30°, and regard it as a six-classification task. SRP is trained by comparing the differences between the predicted and the ground-truth relationship/relative orientation angle using the spatial relationship loss (introduced in Section 3.4).

3.3 Modality Alignment Caption Module

In Section 3.2, we obtain the global and local spatial relationship features. However, we should further integrate them to generate the final description. For the above purpose and considering the differences among different modalities, we propose the MACM, which mainly includes three parts: local relationship fusion (LRF), global relationship fusion (GRF), and modality alignment classifier (MAC). Fig.5 describes the structure of the MACM in detail.

3.3.1 Local Relationship Fusion (LRF)

In the LRF, we model the local context features of objects as follows to supplement the description generation of the relationship between objects.

$$u_t^i = \text{concat}\left(v_i, h_{t-1}^{(2)}, w_i\right) \tag{3}$$

$$h_{\text{t}}^{(1)} = GRU\left(FC\left(u_t^i\right), h_{\text{t}-1}^{(1)}\right) \tag{4}$$

where v_i is the node feature output by LRGM, $h_{t-1}^{(1)}$ and $h_{t-1}^{(2)}$ are the previous hidden states, and w_i is Glove [51] word embedding.

Then, we use the attention mechanism to fuse node and edge features to obtain local description features f_{local}^{τ}:

$$v_r^i = v_i + \sum_{j=1}^{K} f_{local}^{i,j} \tag{5}$$

$$\text{scores}^{\tau} = \text{softmax}\left(\text{FC}\left(\tanh\left(h_{\text{t}}^{(1)} + v_r^i\right)\right)\right) \tag{6}$$

$$f_{\text{local}}^{\tau} = \sum_{i=1}^{M} \text{scores}_i^{\tau} \cdot v_r^i \tag{7}$$

where $f_{local}^{i,j}$ is the enhanced local spatial relationship features (Section 3.2.2).

3.3.2 Global Relationship Fusion (GRF)

In the GRF, the global relationship features f_{global}^i extracted in the previous stage (Section 3.2.3) are further integrated with local description features f_{local}^{τ} and hidden state $h_{\text{t}}^{(1)}$ (extracted in LRF) as follows:

$$h_{\text{t}}^{(2)} = GRU\left(FC\left(\text{concat}\left(f_{\text{local}}^{\tau}, f_{\text{global}}^{i}, h_{\text{t}}^{(1)}\right)\right)\right) \tag{8}$$

We then feed the fused features into the MAC to generate word embedding.

3.3.3 Modality Alignment Classifier (MAC)

Visual tasks can be guided by a wealth of semantic prior knowledge, whereas cross-modality tasks should consider the interplay and complementarity of knowledge across modalities. However, prior works seldom addressed the discrepancies between modalities, impairing the final results. To address the issue of inter-modality discrepancies, we propose two modality alignment classifier 1 (A1) and classifier 2 (A2), in which A1 is a naive version of our idea verifying the feasibility of transforming visual features into linguistic subspace. We further improve A1 to A2, which leverages the word embedding feature as a prior to conduct an intermediate mapping.

Modality Alignment Classifier 1 based on Projection Matrix. The overview of this approach is illustrated in Fig.6(a). First, we project the hidden state $h_t^{(2)} \in \mathbb{R}^{1\times D}$ (output from the visual space in GRF) into the word embedding space using learnable projection matrix $P \in \mathbb{R}^{D\times D}$, and the resulting linguistic feature is denoted as s_i:

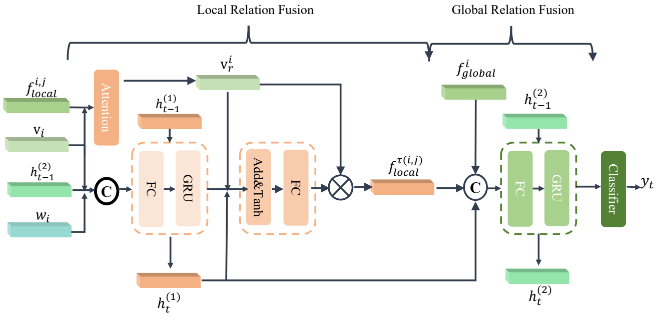


Fig. 5. Modality Alignment Caption Module for generating caption, which mainly includes three parts: local relationship fusion (LRF), global relationship fusion (GRF), and modality alignment classifier (MAC).

$$s_i = h_{\text{t}}^{(2)} P \tag{9}$$

Then, we calculate the matrix multiplication of linguistic feature $s_i \in \mathbb{R}^{1\times D}$ and word embedding space matrix $W_s \in \mathbb{R}^{N_W \times D}$ (where N_W is the number of word vectors) to obtain the final classification result:

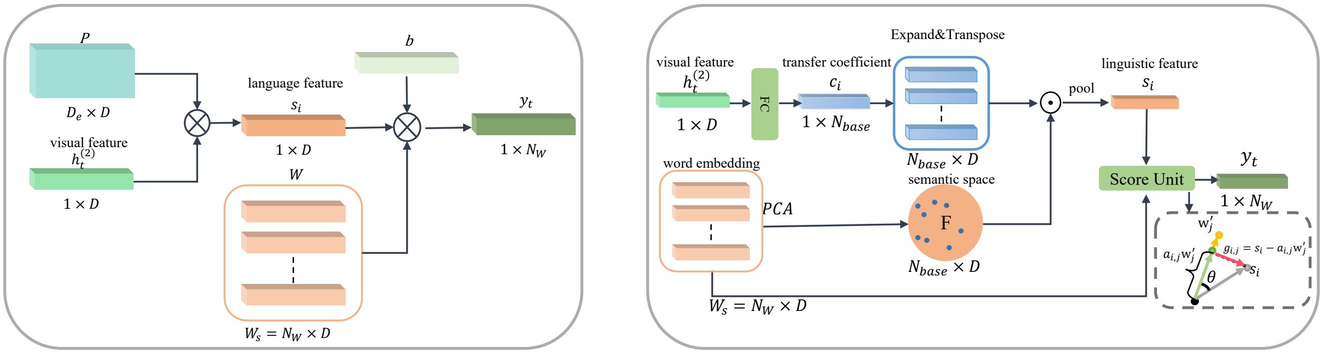
$$scores = s_i W_s^T + b \tag{10}$$

For each word embedding vector $w_j \in \mathbb{R}^{1\times D}$, this calculation equals to the inner-product similarity between s_i and w_j. Given that s_i has prior knowledge of the language space, it can indirectly solve the token classification.

Modality Alignment Classifier 2 based on Attention. A1 which is based on projection matrix largely relies on classification loss, and transforms the modality alignment process into the learning of projection matrix P. However, it is not efficient to learn the projection matrix P directly, and there is no direct supervision for training. To address this issue, we improve A1 to A2 based on attention by leveraging semantic guidance. We use the word embedding feature as a prior to conduct an intermediate mapping. MAC module first transforms the visual features into word embedding space to obtain a linguistic feature, and then uses a score unit to predict the caption word based on the similarity between the linguistic feature and each word embedding vector. Compared with the way that directly classifies visual features into word categories in the vocabulary through linear layers, our model with MAC has obvious improvement in modal alignment thanks to the intermediate mapping to word embedding space. This modality alignment strategy can improve the token classifier's performance without an increase on the complexity the original network.

As illustrated in Fig.6(b), the MAC based on attention aims to execute pre-processing on the word embedding space, and further specifies the problem to strengthen the guiding role in the learning process. For the first part, we need use the corpus as prior information, while the corpus is quite large due to it has about 140,000 words. Thus we use PCA to reduce the dimensionality of the word embedding space. Concretely, we use principal components analysis (PCA) to compute the N_{base} principal components from word embedding space $W_s = [w_1; \ldots; w_{N_W}] \in \mathbb{R}^{N_W \times D}$

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3279204



(a) Modality Alignment Classifier based on Projection Matrix

(b) Modality Alignment Classifier based on Attention

Fig. 6. Modality Alignment Classifier based on Projection Matrix and Attention for aligning visual and linguistic features.

(where N_W is the number of word vectors, N_{base} is the number of eigenvectors), and obtain $F = [f_1; \ldots; f_{N_{base}}] \in \mathbb{R}^{N_{base} \times D}$, which is a simplified representation of word embedding space. Then we generate the new linguistic feature space S, by jointly integrating the processed word embedding F and visual feature $h_t^{(2)}$as follows:

$$c_i = Expand\left(FC\left(h_t^{(2)}\right)\right), \tag{11}$$

$$s_i = pool(F \odot c_i) \tag{12}$$

where $Expand$ is expansion operation, which duplicates the feature vector of dimension $1 \times N_{base}$ with the number of copies equals to D; and $pool$ is pooling operation.

To obtain the final classification result, we must perform a mapping from linguistic feature s_i to classification vector y_t. We develop a score unit to compute the category scores to reduce the distance between ground-truth word vector w_{gt_i} and predicted linguistic feature s_i. The score computation method estimates the classification score between each linguistic feature s_i and each word embedding vector w_j, which can be perceived as the attention mechanism, as follows:

$$\begin{aligned} \text{scores}_{1,i,j} &= -\left[\frac{(a_{i,j} - \|w_j\|)}{\|w_j\|}\right]^2 \\ \text{scores}_{2,i,j} &= -\left\|s_i - a_{i,j}w_j'\right\| \end{aligned} \tag{13}$$

where $a_{i,j} = \|s_i\| \cdot \cos\theta_{i,j}$, and $\theta_{i,j}$ is the angle between s_i and w_j, the scalar $a_{i,j}$ represents the length of projection of s_i on direction w_j. $w_j' = \frac{w_j}{\|w_j\|}$ is the unit vector in w_j direction.The purpose of score unit is to make the vector s_i and its corresponding vector w_{gt} as similar as possible in the semantic space, which is controlled by the relative angle and distance between them. In Eq.13, $score_1$ and $score_2$ measures the relative angle and distance between vectors. Smaller angles and shorter distance will lead to higher score of $score_1$ and $score_2$. Then we will choose the word with the highest score as the prediction answer.

The final classification score between s_i and each word embedding vector w_j is calculated as:

$$\text{scores}_{i,j} = \lambda_1 \,\text{scores}_{1,i,j} + \lambda_2 \,\text{scores}_{2,i,j} \tag{14}$$

Then the final classification result for the i-th object is $y_t = [\text{scores}_{i,1}, \ldots, \text{scores}_{i,N_W}]$.

3.4 Loss Function

We first define two basic loss functions for the subsequent spatial relationship loss calculation. Since the local and global relationships change along with the views, we need to supervise the local and global spatial relationships under different views. We denote the predicted spatial relationship between the i-th and j-th objects under the v-th view as $\hat{y}_r^{i,j,v}$, which is a 0-1 discrete vector, and the ground truth spatial relationship[1] as $y_r^{i,j,v}$, where $r \in \{\text{local}, \text{global}\}$ represents the type of spatial relationship. For the global spatial relationship between the i-th object and the global scene under the v-th view, we represent the predicted and true relationship as $\hat{y}_r^{i,g,v}$ and $y_r^{i,g,v}$, respectively.

- Embedding similarity loss: This loss uses cosine similarity to measure the distance between the predicted and true spatial relationship:

$$L_{sim}^r(i,j,v) = 1 - \cos\left(\hat{y}_r^{i,j,v}, y_r^{i,j,v}\right) \tag{15}$$

- Mutually exclusive relational loss: Given that the fundamental spatial relationships we establish are mutually exclusive, e.g., if object obj_i is on the left side of another object obj_j, then it cannot be on the right side. For the network to learn the reciprocal exclusion of this relationship, we add the mutually exclusive relational loss as follows:

$$L_{mut}^r(i,j,v) = \sum_k (1 - \sqrt{\left(\hat{y}_r^{i,j,v,k} - y_r^{i,j,v,k}\right)^2}) \tag{16}$$
$$k = 0,2,4,6; v = 1,2,3,4$$

where k represents the k-th spatial relationship, and v represents under the v-th view. $y_r^{i,j,v,k} = 1$ indicates that the k-th spatial relationship truly exists, while $\hat{y}_r^{i,j,v,k} = 1$ indicates that the k-th spatial relationship is predicted.

Spatial relationship loss. We propose the spatial relationship loss to supervise the training of the SRP (Section 3.2.4), which predicts the local, global spatial relationship

[1] The ground truth of local spatial relationship between objects A and B is determined by the 3D directionality from A to B, which can be calculated by connecting their center points or bounding boxes under different view direction (front, back, left, right). The view is distinguished by 2D mesh "00 01 10 11".

TABLE 1
Comparison with other STATE-OF-THE-ARTS methods on the ScanRefer and Nr3D Dataset with End-to-End Training ‡(the bounding boxes are generated by 3D Detection backbone) under different threshold of IoU.[†] denotes reproduction in our local environment.[‡] denotes teacher-student design in [4]. S means using the REM only without MAC, and A1 means using the MAC based on the projection matrix. A2 means using the MAC based on attention.

Dataset	Method	DIM	Detection	B-4@0.25	C@0.25	R@0.25	M@0.25	B-4@0.5	C@0.5	R@0.5	M@0.5
ScanRefer	2D-3D Proj [3]	2D	M R-CNN	10.27	18.29	33.63	16.67	2.31	8.31	25.93	12.54
	3D-2D Proj [3]	2D	VoteNet	17.86	19.73	40.68	19.83	8.56	11.47	31.65	15.73
	VoteNetRetr [3]	3D	VoteNet	18.09	15.12	38.99	19.93	13.38	10.18	33.22	17.14
	Scan2Cap [3]	3D	VoteNet	34.18	56.82	55.27	26.29	23.32	39.08	44.78	21.97
	Scan2Cap [3]†	3D	VoteNet	33.06	54.82	54.19	25.88	23.13	39.21	44.93	21.99
	TransCap [4]	3D	VoteNet	35.04	60.04	54.46	26.27	24.25	43.12	44.72	22.15
	X-Trans2Cap [4] ‡	3D	VoteNet	35.65	61.83	54.70	26.61	25.05	43.87	45.28	22.46
	Ours (S)	3D	VoteNet	34.27	57.46	55.58	26.67	23.67	39.12	45.47	22.12
	Ours (S+A1)	3D	VoteNet	34.23	59.26	55.4	26.36	24.34	41.85	45.82	22.34
	Ours (S+A2)	3D	VoteNet	**36.37**	**62.01**	**56.25**	**26.76**	**26.31**	**45.00**	**46.96**	**22.67**
Nr3D	3D-2D Proj [3]	2D	VoteNet	7.49	8.57	44.95	18.83	4.21	3.93	41.24	16.68
	Scan2Cap [3]	3D	VoteNet	24.43	42.24	55.88	25.07	15.01	24.10	47.95	21.01
	TransCap [4]	3D	VoteNet	25.79	45.06	55.55	25.22	19.09	33.45	50.00	22.24
	X-Trans2Cap [4] ‡	3D	VoteNet	27.62	51.43	56.46	25.75	19.29	33.62	50.00	22.27
	Ours (S)	3D	VoteNet	25.93	48.67	56.58	25.97	17.74	29.42	49.57	22.19
	Ours (S+A1)	3D	VoteNet	27.78	49.68	57.01	26.06	19.38	32.71	50.63	22.69
	Ours (S+A2)	3D	VoteNet	**28.61**	**52.39**	**57.67**	**26.45**	**20.37**	**34.81**	**50.99**	**23.01**

under different views and the relative orientation angle from the extracted relationship features. Our spatial relationship loss consists of three parts. The local spatial loss is computed using the embedding similarity loss and the mutually exclusive relational loss defined above:

$$\begin{aligned} L_{\text{local}} &= \operatorname{Avg}\left(L_{sim}^{\text{local}}(i,j,v)\right) + \operatorname{Avg}\left(L_{mut}^{\text{local}}(i,j,v)\right), \\ & i=1,\ldots,M; j=N(i); v=1,2,3,4 \end{aligned} \tag{17}$$

Similarly, we define the global spatial relationship loss as follows:

$$\begin{aligned} L_{\text{global}} &= \operatorname{Avg}\left(L_{sim}^{\text{global}}(i,g,v)\right) + \operatorname{Avg}\left(L_{mut}^{\text{global}}(i,g,v)\right), \\ & i=1,\ldots,M; v=1,2,3,4 \end{aligned} \tag{18}$$

We then introduce the relative angle loss to supervise the prediction of the relative orientation angle:

$$L_{\text{angle}} = L_{\text{cross-entropy}}\left(y_{\text{angle}}^{i,j}, \hat{y}_{\text{angle}}^{i,j}\right) \tag{19}$$

where $y_{\text{angle}}^{i,j}$ is the true class label of relative orientation angle between the i-th and the j-th objects (0° to 180° divided into six intervals), and $\hat{y}_{\text{angle}}^{i,j}$ is the predicted class label.

The final spatial relationship loss is defined as follows:

$$L_{\text{spatial}} = \alpha L_{\text{local}} + \beta L_{\text{global}} + \gamma L_{\text{angle}} \tag{20}$$

Caption loss. To train the MACM, we define the caption loss as the cross-entropy loss function:

$$L_{\text{cap}} = L_{\text{cross-entropy}}\left(y_t, y_{gt}\right) \tag{21}$$

where y_t is the probability of generating each word, y_{gt} is the one-hot encoding for the ground-truth token.

3D Object Detection Loss. For end-to-end training, we apply the 3D object detection loss L_{dect} to optimize the object detection backbone. As this loss is common, please refer to [48] for the specific definition.

Total loss. On the basis of the losses defined above, our final loss function is summarized as follows:

$$L_{total} = L_{\text{cap}} + L_{\text{spatial}} + \delta L_{\text{dect}} \tag{22}$$

4 Experiment

In this section, we evaluate and analyze the results of the proposed REMAN. This section first introduces the datasets, the hardware and software configurations, and the metrics that are used for the experiments. We then qualitatively and quantitatively compare the proposed REMAN with existing state-of-the-art methods. Ablation experiments are then carried out to verify the effectiveness of the key modules in our method. The experiment results demonstrate that the proposed network is accurate and diverse in representing the spatial relationships between objects in a scene.

4.1 Datasets and Metrics

ScanRefer dataset: Chen et al. proposed the ScanRefer [1] dataset, which enhances the ScanNet [52] dataset with extensive descriptive information. The dataset includes 800 3D indoor scenes and 51,583 descriptive annotations for 11,046 objects. These descriptions primarily consist of information about the object's appearance and spatial relationship, such as the object's color and material, and the object's position relative to its neighbors.

Nr3D dataset: Natural Reference in 3D Scenes (Nr3D) [2] is similar to ScanRefer in that it builds a dataset by annotating ScanNet [52] objects with natural language descriptions. The dataset contains 41,503 description annotations that were annotated interactively using manual question-and-answer games. It should be underlined that the description annotation does not use the template generation method, which results in a diversified and less redundant sentence pattern, which makes this dataset more challenging than ScanRefer. Additionally, we do not undertake experiments on the Sr3D dataset because it is entirely generated by machine templates and hence has a lower degree of comparability.

Metrics: The quantitative metrics utilized in our experiments are BLEU-4 [5], CIDEr [6], ROUGE [7], and METEOR [8], which are denoted by acronyms B-4, C, R, and M in the table. The higher the values of the aforementioned indicators are, the better the network performance is.

4.2 Implementation Details

We conduct experiments under two different training strategies: (1) End-to-end training employs point cloud data as input and uses the 3D object detection backbone to generate candidate bounding boxes. (2) Non end-to-end training uses the ground-truth object bounding boxes as position information and further extracts object features for dense captioning. Non end-to-end training performed on the REMAN without passing through the 3D object detection backbone, which can be regarded as the upper limit of our model.

For end-to-end training, we employ VoteNet [48] as the 3D object detection backbone, and train the network for 50 epochs using the Adam optimizer with the batch size set to 8. For non end-to-end training, we employ PointNet++ [53] to extract object-level features, and train the network for 20 epochs with the batch size set to 32. We set the hyperparameters in Eq.(14) as $\lambda_1 = 1, \lambda_2 = 2$, and for loss terms in Eq. (20) and Eq. (22) as $\alpha = 0.1, \beta = 0.1, \gamma = 0.1$ and $\delta = 10$. The initial learning rate is 1e-3, and the weight decay factor is 1e-5. We implement our architecture in PyTorch and train on a single RTX2080Ti GPU.

4.3 Quantitative Comparison with State-of-the-art

This section compares the proposed REMAN to existing state-of-the-art models, including 2D-3D Proj [3], 3D-2D Proj [3], VoteNetRetr [54], Scan2Cap [3], OracleRetr3D [3], OracleCap3D [3], TransCap[2] and X-Trans2Cap [4].

Quantitative comparisons are conducted under both end-to-end and non end-to-end training. For fair comparison, we also re-trained Scan2Cap† [3], Oracle2Cap3D†, TransCap† and X-Trans2Cap† [4] in the local experimental context. In Table 1-3, ours(S), ours(S+A1) and ours(S+A2) respectively represent using only REM without MAC (namely, using linear layer as a classifier), using REM and MAC(A1), and using REM and MAC(A2). The comparison of ours(S) is just used to validate the function of MAC, and ours(S+A2) could be regarded as our complete model.

Comparison under End-to-end training. To explore the impact of IoU, we carried out a quantitative comparison under the IoU thresholds of 0.25 and 0.5 and the results are shown in Table 1.

When the IoU threshold is 0.25, our model outperforms Scan2Cap [3] on the metrics of BLEU-4, CIDEr, ROUGE, and METEOR by 2.19, 5.19, 0.98, 0.47 on Scanrefer, and by 4.18, 10.15, 1.79, 1.38 on Nr3D, respectively. Furthermore, the improvement of our model beyond Scan2Cap [3] on all metrics is greater than that of X-Trans2Cap [4] which represents the best performance of SOTA works. Even if we compare the X-Trans2Cap [4] based on the teacher-student design in [36], we still outperform it without the design.

To further validate the network's performance, we additionally set the IoU threshold to 0.5. Table 1 shows that our proposed network also significantly outperforms other methods. Compared with Scan2Cap [3], metrics BLEU-4, CIDEr, ROUGE and METEOR are improved by 2.99, 5.92, 2.18 and 0.7 on Scanrefer and 5.36, 10.71, 3.04, 2.00 on Nr3D respectively. Our model also outperforms TransCap and X-Trans2Cap [4], with all metrics increased.

TABLE 2
Comparison with other STATE-OF-THE-ARTS methods on the ScanRefer Dataset with Non End-to-End training (the bounding boxes are generated by human marking).† denotes reproduction in our local environment.‡ denotes teacher-student design in [36]

Method	B-4	C	R	M
OracleRetr3D [3]	33.03	23.36	52.99	25.80
Oracle2Cap3D [3]	41.49	67.95	63.66	29.23
Oracle2Cap3D [3]†	40.38	64.98	63.36	29.03
TransCap [4]†	41.70	**86.03**	62.31	30.22
X-Trans2Cap [4]†‡	45.19	82.49	65.67	29.89
Ours (S)	42.66	67.10	64.23	29.36
Ours (S+A1)	42.59	70.89	64.12	29.01
Ours (S+A2)	**45.30**	78.34	**66.28**	**30.28**

TABLE 3
Comparison of parameters between end-to-end and non end-to-end training of various models. M denotes one million parameters.

Method	Parameters	
	End-to-End	Non End-to-End
Scan2Cap [3]	6.1M	5.2M
TransCap [4]	21.9M	21.0M
X-Trans2Cap [4]	40.9M	40.0M
Ours (S)	6.7M	5.7M
Ours (S+A1)	6.7M	5.7M
Ours (S+A2)	**5.1M**	**4.2M**

To summarize, in both IoU thresholds of 0.25 or 0.5, our model takes the lead over the state-of-the-art methods on both Scanrefer and Nr3D, demonstrating the superiority of the proposed method.

Comparison under non end-to-end training. Under non end-to-end training, no IoU threshold for comparison is required because the position information is provided by the ground-truth bounding boxes. As the author has not made the model and training code for TransCap [4] open source, we reproduced them according to the paper. Table 2 summarizes the experimental results.

From the comparison of results in Table 2, it can see that the effect of human-marked bounding boxes is obviously better than that of the predicted bounding boxes from the 3D object detection backbone. Except for CIDEr, our model (S+A2) achieves state-of-the-art in all other three metrics (BLUE-4, ROUGH and METEOR) and greatly exceeds the baseline model Scan2Cap [3].

Comparison of Parameters. We also compared the parameters of the different models under end-to-end and non end-to-end trainings as shown in Table 3. Under the end-to-end training strategy, compared with the Scan2Cap model, we only increased the parameters by 10% after applying the S and A1 modules. After replacing A1 with A2, the parameter quantity decreased by 28% and is only 80% of Scan2Cap. Although the CIDEr of X-Trans2Cap [4] achieves the state-of-the-art under the non end-to-end training strategy, its parameter quantity is nine times higher than our REMAN. The parameter quantity of TransCap (without the student-teacher design in [36], and we only measure the student module) is still four times higher than our REMAN. In general, our REMAN has made amazing achievements in

[2] Note that TransCap is X-Trans2Cap [4] without teacher-student design in [36].

TABLE 4
Ablation Study for REM on ScanRefer under 0.25IoU where L_f, L_o,L and G denotes local fundamental, local overlap, complete local and global spatial relationship, respectively. Our baseline model is Scan2Cap†.

Config	B-4@0.25	C@0.25	R@0.25	M@0.25
baseline w/o LRGM	32.49	53.29	52.93	25.37
baseline	33.06	54.82	54.19	25.88
baseline+L_f	33.72	55.61	55.16	26.24
baseline+L_o	33.37	55.06	54.41	25.97
baseline+L	33.97	55.93	55.39	26.44
baseline+G	33.74	56.03	54.73	26.54
baseline+L+G (S)	**34.27**	**57.46**	**55.58**	**26.67**

TABLE 5
Ablation Study with different modality alignment methods on ScanRefer under 0.25IoU. A1 means the Modality Alignment Classifier based on projection matrix. A2 means the Modality Alignment Classifier based on attention. Our baseline model is Scan2Cap†.

Config	B-4@0.25	C@0.25	R@0.25	M@0.25
baseline	33.06	54.82	54.19	25.88
baseline+A1	34.68	58.32	55.21	26.03
baseline+A2	**35.56**	**60.72**	**55.96**	**26.49**

most of the metrics with minimum parameters.

4.4 Ablation Study

4.4.1 *Analysis of REM module*

We first perform ablation studies on our REM module, and summarize the results in Table 4. We compare ours with six ablated versions, where the baseline is Scan2Cap† and our complete model is the baseline + L + $G(s)$ using REM. It is seen that when LRGM is removed from the baseline model, all the metrics declines in varying degrees. And when different scales of spatial relationship features are used (L_f, L_o,L and G), the description ability of the network is improved to some extent. After using complete 3D spatial relationship features (L + $G(s)$), the description ability of the network is further superimposed. The purpose of introducing complete 3D spatial relationships is to help the network develop rich semantic descriptions and guarantee that all reasonable spatial relationship descriptions are complete.

Our proposed REM module exhibits a small performance improvement when the threshold of IoU is 0.5 on the ScanRefer dataset as shown in Table 1. The main reason is that the description on ScanRefer dataset is rich, involving many objects. When the IoU threshold is 0.5, many objects are filtered out, resulting in the model being unable to learn enough spatial relationship, which will be quite different from ground-truth. When the IoU threshold is 0.25, more objects are detected, which is beneficial to the REM model. For the Nr3D dataset, the IoU setting is not extremely sensitive because the description is relatively simple.

4.4.2 *Analysis of MAC module*

The network performance is greatly boosted with the addition of a MAC (A1 / A2), as shown in Table 1 and Table 2. We also conduct an ablation study on the two modality alignment methods in Table 5. The results show that compared with adding REM, our novel MAC can significantly improve the network performance because we make use of prior knowledge (word embedding) from other spaces to aid learning.

The impact of the modality alignment based on attention by leveraging semantic guidance (A2) is clearly superior to that of the projection matrix approach (A1). The following are the key reasons: (1) The modality alignment based on attention (A2) simplifies the problem and reduces the network's learning difficulty. In this approach, PCA is first used to filter unwanted features and results in a simplified representation of the word embedding space, and then the network reconstructs the target word embedding using a set of features, with a precise learning objective; (2) A score unit is designed to measure the reconstruction effect, which transfers the reconstructed word embedding to the classifier's output. The difference between the reconstructed and intended target word vectors is actually encoded throughout this procedure, helping the network supervise the reconstruction process well.

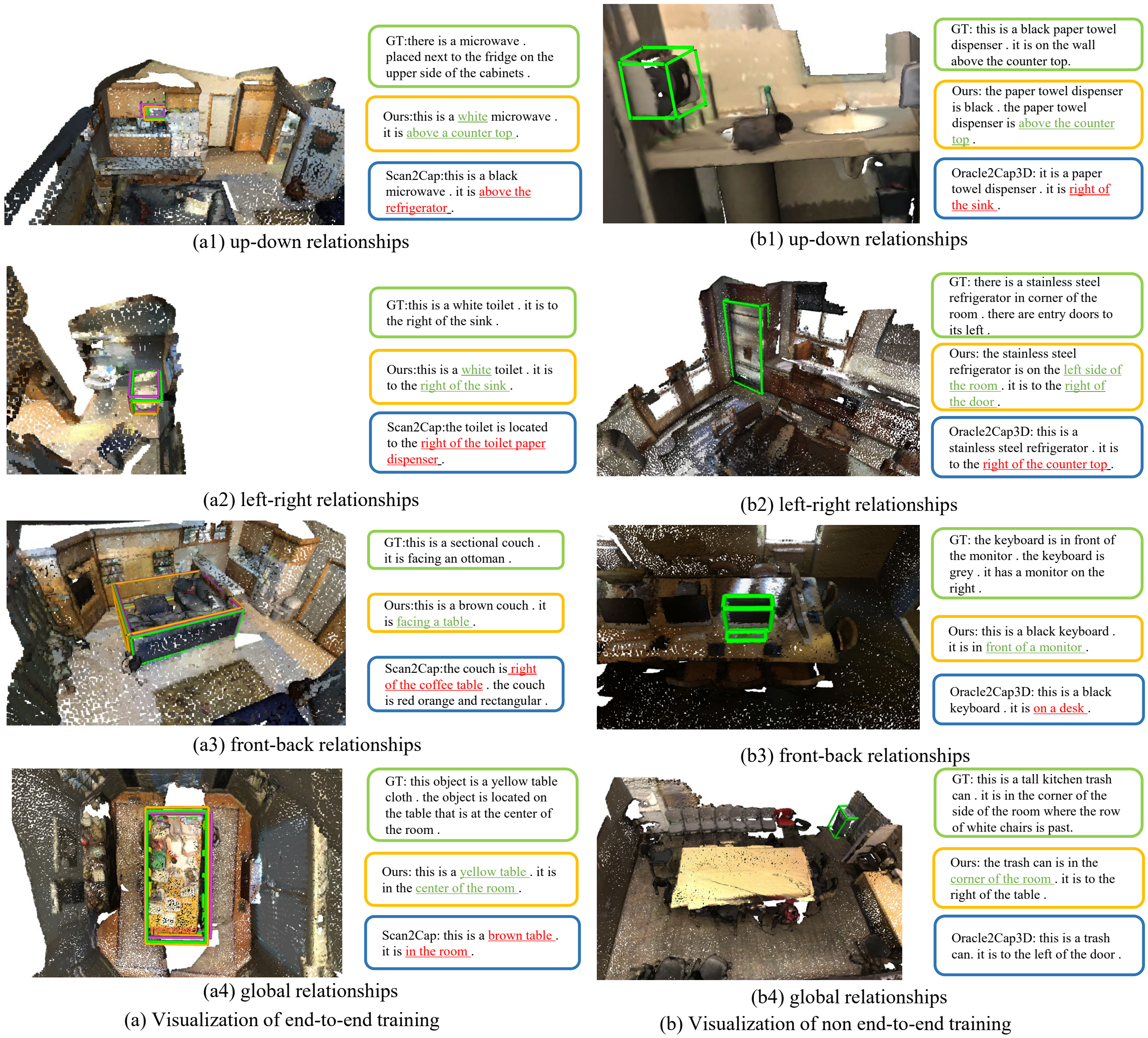
4.5 Visualization

To intuitively compare this method with Scan2Capp [3], the experiment results are visualized for supplementary explanation and analysis.

Visualization for end-to-end training. The visual comparison results of Scan2Cap [3], REMAN under end-to-end training and GT are shown in Fig.7(a). As Fig.7(a1) shows that REMAN generates the description of the up-down relationship between the microwave oven and the counter top, while Scan2Cap [3] incorrectly describes the relationship as it is on the refrigerator. Actually, Scan2Cap [3] does not supervise the specific relationship between objects, and only relies on the constraint of the relative orientation angle, which cannot completely cover all possible spatial relationships, so Scan2Cap [3] is relatively weak in correctly describing the spatial relationships. Similarly, REMAN generates the description of the left-right relationship between the toilet and the sink in (a2), describes the table in front of the couch in (a3); and successfully describes the global relationship (center) between the table and the global scene in (a4). REMAN has a good ability to correctly describe various spatial relationships and appearance characteristics such as color, while Scan2Cap [3] often makes mistakes or obtains meaningless description statements due to the lack of REM and MACM.

Visualization for Non End-to-End training. Fig.7(b) shows the visualization of the outputs of REMAN, Oracle2Cap3D [3] under non end-to-end training and the corresponding GT. Both methods adopt the human-marked real bounding boxes as input, and generate the descriptions. The scene (b1) describes the up-down spatial relationship between the paper towel dispenser and the counter top. Scene (b2) shows REMAN can correctly describe the left-right relationship between the refrigerator and the door, and the overall relationship between the refrigerator and the room. Scene (b3) demonstrates that REMAN can correctly describe the relationship between a monitor and the keyboard placed in front of it. Scene (b4) shows that REMAN successfully describes the global relationship (corner), whereas Scan2Cap cannot. REMAN can cover all the different spatial spatial

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3279204



GT:there is a microwave . placed next to the fridge on the upper side of the cabinets .

Ours:this is a white microwave . it is above a counter top .

Scan2Cap:this is a black microwave . it is above the refrigerator .

(a1) up-down relationships

GT:this is a white toilet . it is to the right of the sink .

Ours:this is a white toilet . it is to the right of the sink .

Scan2Cap:the toilet is located to the right of the toilet paper dispenser .

(a2) left-right relationships

GT:this is a sectional couch . it is facing an ottoman .

Ours:this is a brown couch . it is facing a table .

Scan2Cap:the couch is right of the coffee table . the couch is red orange and rectangular .

(a3) front-back relationships

GT: this object is a yellow table cloth . the object is located on the table that is at the center of the room .

Ours: this is a yellow table . it is in the center of the room .

Scan2Cap: this is a brown table . it is in the room .

(a4) global relationships

(a) Visualization of end-to-end training

GT: this is a black paper towel dispenser . it is on the wall above the counter top.

Ours: the paper towel dispenser is black . the paper towel dispenser is above the counter top .

Oracle2Cap3D: it is a paper towel dispenser . it is right of the sink .

(b1) up-down relationships

GT: there is a stainless steel refrigerator in corner of the room . there are entry doors to its left .

Ours: the stainless steel refrigerator is on the left side of the room . it is to the right of the door .

Oracle2Cap3D: this is a stainless steel refrigerator . it is to the right of the counter top .

(b2) left-right relationships

GT: the keyboard is in front of the monitor . the keyboard is grey . it has a monitor on the right .

Ours: this is a black keyboard . it is in front of a monitor .

Oracle2Cap3D: this is a black keyboard . it is on a desk .

(b3) front-back relationships

GT: this is a tall kitchen trash can . it is in the corner of the side of the room where the row of white chairs is past.

Ours: the trash can is in the corner of the room . it is to the right of the table .

Oracle2Cap3D: this is a trash can. it is to the left of the door .

(b4) global relationships

(b) Visualization of non end-to-end training

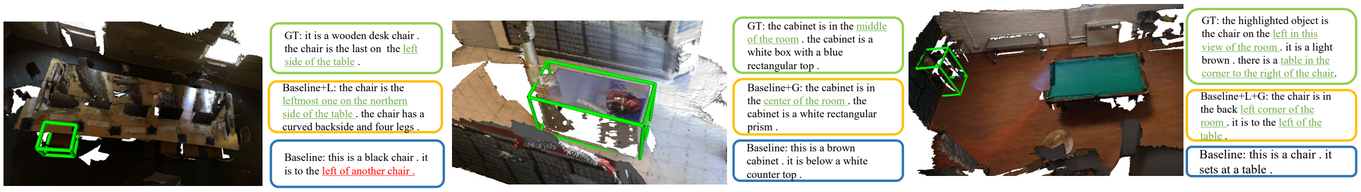
Fig. 7. Qualitative comparisons on ScanRefer dataset. The left illustrates the results under end-to-end training, and the right part presents the results under non end-to-end training. Best viewed in color.

relationships that are common in the scene, and generate descriptions similar to GT.

Visualization for ablation of spatial relationships. Fig. 8(a), (b) and (c) respectively depicts the visualization results of ablation study on local spatial relationship module, global spatial relationship module, and both local and global spatial relationship modules. As seen from Fig. 8(a), our model (baseline +L) successfully describes the local spatial relationship of the chair on the left of table; In Fig. 8(b), our model (baseline +G) successfully describes the global spatial relationship of cabinet which is in the center of the room; and in Fig. 8(c), our model (baseline+L+G) successfully describes both the local and global spatial relationship of the chair, which is on the left side of the table and on the left side of the room. However, the baseline model can not achieve the right spatial relationship in Fig. 8(a)-(c) respectively.

Visualization for ablation of MAC. Fig.9 shows the visualization results of our model with MAC and the baseline without MAC. As seen from Fig.9(a) and (b), ours (baseline+A2) successfully recognize that the objects are monitor and bookshelf respectively, while the baseline mistakenly regarded them as lamp and door caused by the inter-modality discrepancy between vision and language.

Visualization for low quality scene. We test our model on both manually edited low quality point cloud and original one with poor quality. As shown in Fig. 10, we made holes (Fig. 10(b)) and added noise (Fig. 10(c)) in a point cloud from ScannetV2 by using Meshlab software. Comparing Fig. 10(a) and Fig. 10(b), we found that after we removed the back of the chair, the predicted results of our model do not change. Comparing Fig. 10(a) and Fig. 10(c), when we translated the back of the chair along X and Y axes, although some changes have taken place in the description, the meaning is the same as the original one. The results demonstrates the robustness of our method in dealing with scenes with holes and noises. We also tested on



GT: it is a wooden desk chair . the chair is the last on the left side of the table .

Baseline+L: the chair is the leftmost one on the northern side of the table . the chair has a curved backside and four legs .

Baseline: this is a black chair . it is to the left of another chair .

(a) ablation for local relationship

GT: the cabinet is in the middle of the room . the cabinet is a white box with a blue rectangular top .

Baseline+G: the cabinet is in the center of the room . the cabinet is a white rectangular prism .

Baseline: this is a brown cabinet . it is below a white counter top .

(b) ablation for global relationship

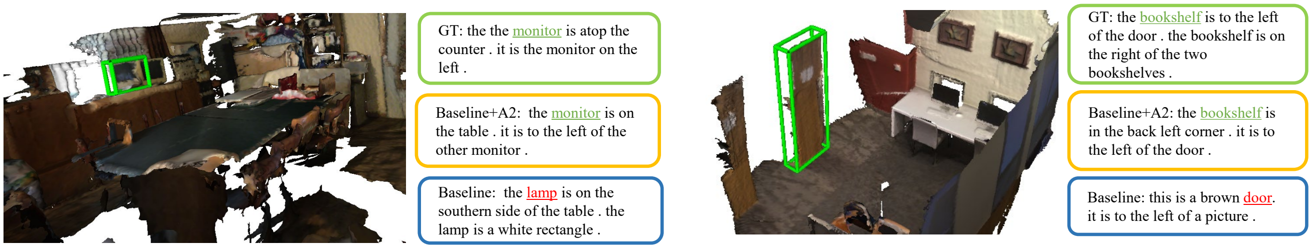
GT: the highlighted object is the chair on the left in this view of the room . it is a light brown . there is a table in the corner to the right of the chair.

Baseline+L+G: the chair is in the back left corner of the room . it is to the left of the table .

Baseline: this is a chair . it sets at a table .

(c) ablation for both local and global relationship

Fig. 8. Visualization of ablation study on spatial relationship.



GT: the the monitor is atop the counter . it is the monitor on the left .

Baseline+A2: the monitor is on the table . it is to the left of the other monitor .

Baseline: the lamp is on the southern side of the table . the lamp is a white rectangle .

(a)

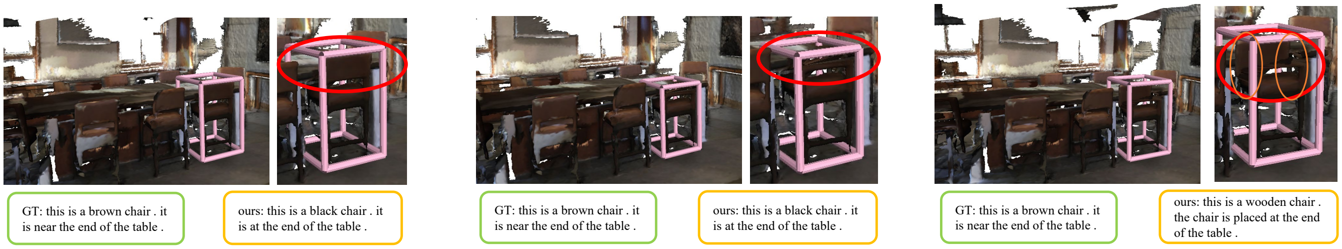
GT: the bookshelf is to the left of the door . the bookshelf is on the right of the two bookshelves .

Baseline+A2: the bookshelf is in the back left corner . it is to the left of the door .

Baseline: this is a brown door. it is to the left of a picture .

(b)

Fig. 9. Visualization of ablation study on MAC(A2).



GT: this is a brown chair . it is near the end of the table .

ours: this is a black chair . it is at the end of the table .

(a) result for original point cloud

GT: this is a brown chair . it is near the end of the table .

ours: this is a black chair . it is at the end of the table .

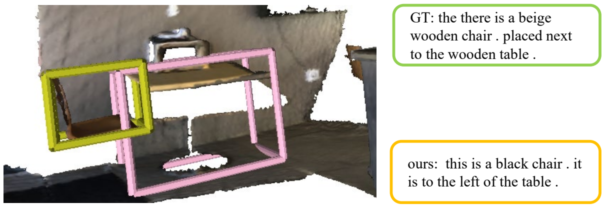
(b) result for removing part of the back of the chair.

GT: this is a brown chair . it is near the end of the table .

ours: this is a wooden chair . the chair is placed at the end of the table .

(c) result for moving part of the back of the chair.

Fig. 10. Predicted results of manually edited low quality point cloud.



GT: the there is a beige wooden chair . placed next to the wooden table .

ours: this is a black chair . it is to the left of the table .

Fig. 11. Predicted results of original low quality point cloud.

a point cloud data with poor quality in the original dataset, in which the four legs of the chair and the legs of the table were not scanned, as shown in Fig. 11. Our model can still successfully recognize the chair and the table, and describe their relationship. Note that these point clouds in Fig. 10 and Fig. 11 are not used in the training process.

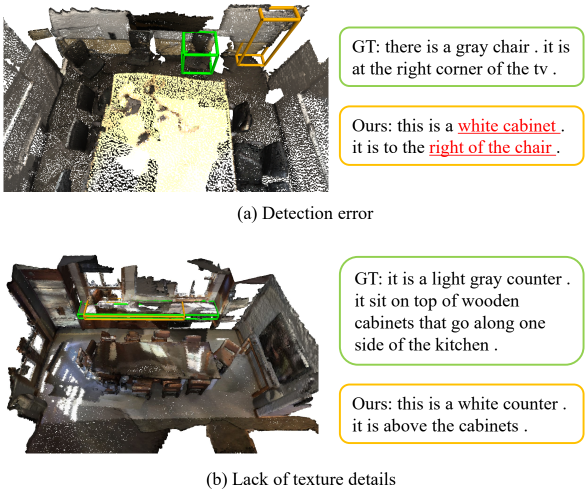
Failure case. The inaccuracy of the object detection backbone can lead to inaccurate descriptions. As shown in Fig.12(a), we expect to describe the chair, but the detection backbone VoteNet [48] detects the object as cabinet. Furthermore, the description of the spatial relationship is also inaccurate, because the detected bounding box position is inaccurate.

Sometimes our generated descriptions lack texture and appearance details. As shown in Fig.12(b), the generated sentence is relatively simple compared with the ground truth. Meanwhile, the generated statement only describes the target object in detail and only describes the spatial relationship for the related objects.

5 Conclusion and Future Work

In this paper, we propose a complete 3D network (REMAN) for 3D dense captioning. REMAN obtains accurate descriptions of multi-scale relationships by introducing a thorough definition of spatial relationships and considering the impact of view transformation on the relationship extraction. We efficiently bridge the semantic gap problem between different modalities by leveraging the prior information in the word embedding and propose two modality alignment strategies, outperforming the state-of-the-art methods on multiple datasets. Our modality alignment method is lightweight and may be used for a wide range of cross-modality tasks.

Although REMAN can generate descriptive statements about various spatial relationships, the descriptions of the



GT: there is a gray chair . it is at the right corner of the tv .

Ours: this is a white cabinet . it is to the right of the chair .

(a) Detection error

GT: it is a light gray counter . it sit on top of wooden cabinets that go along one side of the kitchen .

Ours: this is a white counter . it is above the cabinets .

(b) Lack of texture details

Fig. 12. Visualization of some failure cases.

appearance and attributes are relatively lacking. Furthermore, the method has not explored the language structure and logic in generating the descriptions. In this regard, future works can be carried out from the following aspects: (1) grammatical structure relationships can be introduced to help generate complex and reasonable descriptions; (2) we may enhance the ability to capture the appearance and attributes of the features of objects in the scene and further enhance the richness and diversity of description sentences; (3) we may optimize the model structure to improve the calculation efficiency since the bounding boxes generated by the 3D object detection backbone greatly influence the network's performance.

References

[1] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European Conference on Computer Vision*. Springer, 2020, pp. 202–221.

[2] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 422–440.

[3] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2cap: Context-aware dense captioning in rgb-d scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3193–3203.

[4] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, Z. Li, and S. Cui, "X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning," *arXiv preprint arXiv:2203.00843*, 2022.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[6] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[7] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 2003, pp. 150–157.

[8] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[11] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[12] P. Kinghorn, L. Zhang, and L. Shao, "A region-based image caption generator with refined descriptions," *Neurocomputing*, vol. 272, pp. 416–424, 2018.

[13] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4125–4134.

[14] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1242–1250.

[15] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[16] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 313–10 322.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.

[22] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[23] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[25] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2019.

[26] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[27] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[28] S. Ye, D. Chen, S. Han, and J. Liao, "3d question answering," *arXiv preprint arXiv:2112.08359*, 2021.

[29] X. Yan, Z. Yuan, Y. Du, Y. Liao, Y. Guo, Z. Li, and S. Cui, "Clevr3d: Compositional language and elementary visual reasoning for question answering in 3d real-world scenes," *arXiv preprint arXiv:2112.11691*, 2021.

[30] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," *arXiv preprint arXiv:2112.10482*, 2021.

[31] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese, "Text2shape: Generating shapes from natural language by learning joint embeddings," in *Asian conference on computer vision*. Springer, 2018, pp. 100–116.

[32] Z. Han, C. Chen, Y.-S. Liu, and M. Zwicker, "Shapecaptioner: Generative caption network for 3d shapes by learning a mapping from parts detected in multiple views to sentences," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1018–1027.

[33] P.-H. Huang, H.-H. Lee, H.-T. Chen, and T.-L. Liu, "Text-guided graph neural networks for referring 3d instance segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1610–1618.

[34] Z. Yuan, X. Yan, Y. Liao, R. Zhang, S. Wang, Z. Li, and S. Cui, "Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1791–1800.

[35] D. He, Y. Zhao, J. Luo, T. Hui, S. Huang, A. Zhang, and S. Liu, "Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2344–2352.

[36] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[37] Y. Kant, D. Batra, P. Anderson, A. Schwing, D. Parikh, J. Lu, and H. Agrawal, "Spatially aware multimodal transformers for textvqa," in *European Conference on Computer Vision*. Springer, 2020, pp. 715–732.

[38] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2286–2293.

[39] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2615–2624.

[40] J. Wang, J. Tang, M. Yang, X. Bai, and J. Luo, "Improving ocr-based image captioning by incorporating geometrical relationship," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1306–1315.

[41] Z. Zhang, D. Xu, W. Ouyang, and L. Zhou, "Dense video captioning using graph-based sentence summarization," *IEEE Transactions on Multimedia*, vol. 23, pp. 1799–1810, 2021.

[42] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[43] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval-an empirical odyssey," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6439–6448.

[44] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[45] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1047–1055.

[46] W. Li, Y. Wang, Y. Su, X. Li, A. Liu, and Y. Zhang, "Multi-scale fine-grained alignments for image and sentence matching," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[47] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, and G. Xu, "Entity-oriented multi-modal alignment and fusion network for fake news detection," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[48] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

[49] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 206–215.

[50] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[51] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[52] 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.

[53] 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.

[54] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

Aihua Mao is a professor with the School of Computer Science and Engineering, South China University of Technology (SCUT), China. He received the PhD degree from the Hong Kong Polytechnic University in 2009, the M.Sc degree from Sun Yat-Sen University in 2005 and the B.Eng degree from Hunan University in 2002. His research interests include 3D vision and computer graphics.

Zhi Yang received the B.S. degree in software engineering from Fuzhou University, Fuzhou,China in 2020. He is currently pursuing the M.S. degree in computer science with South China University of Technology, Guangzhou, China. His current research interest is multimodal learning.

Wanxin Chen received the B.S. degree in international economy and trading from Northeastern University in 2018. She obtained her M.S. degree in computer science with South China University of Technology, Guangzhou, China. Her current research interests include machine learning, 3D point cloud.

Ran Yi is an assistant professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. She received the BEng degree and the PhD degree from Tsinghua University, China, in 2016 and 2021. Her research interests include computer vision, computer graphics and computational geometry.

Yong-Jin Liu is a professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include computer graphics and computer-aided design.