MORE: Multi-Order RElation Mining for Dense Captioning in 3D Scenes

Yang Jiao[1,2⋆], Shaoxiang Chen[3⋆], Zequn Jie[3], Jingjing Chen[1,2†], Lin Ma[3†], and Yu-Gang Jiang[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3] Meituan Inc.

Abstract. 3D dense captioning is a recently-proposed novel task, where point clouds contain more geometric information than the 2D counterpart. However, it is also more challenging due to the higher complexity and wider variety of inter-object relations contained in point clouds. Existing methods only treat such relations as by-products of object feature learning in graphs without specifically encoding them, which leads to sub-optimal results. In this paper, aiming at improving 3D dense captioning via capturing and utilizing the complex relations in the 3D scene, we propose MORE, a Multi-Order RElation mining model, to support generating more descriptive and comprehensive captions. Technically, our MORE encodes object relations in a progressive manner since complex relations can be deduced from a limited number of basic ones. We first devise a novel Spatial Layout Graph Convolution (SLGC), which semantically encodes several first-order relations as edges of a graph constructed over 3D object proposals. Next, from the resulting graph, we further extract multiple triplets which encapsulate basic first-order relations as the basic unit, and construct several Object-centric Triplet Attention Graphs (OTAG) to infer multi-order relations for every target object. The updated node features from OTAG are aggregated and fed into the caption decoder to provide abundant relational cues, so that captions including diverse relations with context objects can be generated. Extensive experiments on the Scan2Cap dataset prove the effectiveness of our proposed MORE and its components, and we also outperform the current state-of-the-art method. Our code is available at https://github.com/SxJyJay/MORE.

Keywords: Point Cloud, Graph, Caption Generation

1 Introduction

Dense captioning, which aims at comprehending the visual scene through jointly localizing and describing multiple objects, has been extensively studied in the 2D

⋆ Equal contribution.

† Corresponding authors.

computer vision community [10, 9, 13, 20, 39]. However, 2D data such as images and videos inherently lack the ability of accurately capturing the physical extent of objects and their locations in the scene. Recently, the 3D dense captioning task has been proposed by Chen et al. [11], where pure point clouds are adopted as the visual representation to perform object localization and captioning on. By connecting 3D scenes with natural language, 3D dense captioning has widespread application prospects in the field of human-machine interaction in augmented reality [21, 43], autonomous agents [32, 42], etc.

The abundant geometric information contained in 3D point clouds can support describing object relations and the holistic scene (scene layouts) in a diversified manner, since the 3D point clouds are less limited by the occlusion and better capture object size and relative position. For example, as shown in Fig.1, the circled chair can be described with diversiform relations like *"on the right side"* and *"second ... from ..."*. So in 3D dense captioning, mining the complex inter-object relations is of vital importance for generating comprehensive captions. But directly adapting relation modeling techniques in 2D images [26, 27, 8, 38, 33, 17, 19, 18] to 3D point clouds can lead to poor performances as discussed in the previous work [11]. Scan2Cap, as the first 3D dense captioning work, develops a graph-based encoder to provide relation feature for the caption decoder and achieves promising results. However, it treats inter-object relations as by-products derived from node recognition in a graph, overlooking the gap between the visual world (represented by point clouds) and semantic concepts. Hence, complex relations, such as *"between"*, *"surrounded by"*, *"rightmost"*, etc, can not be properly mapped into the semantic space, which leads to unitary descriptions being generated. As illustrated in Fig.1, Scan2Cap tends to describe the target object by capturing simple relations (*"on the left"*).

To address the above problem, we propose MORE, a Multi-Order RElation mining model which can support generating more descriptive captions for objects in 3D point clouds. The core motivation behind MORE lies in the confidence that the multi-order relations can be deduced from a limited number of basic first-order relations. For example, given a scene where there are three objects in a row, dubbed A, B, and C, and *the relations between AB and BC are both "on the left of"*, then the conclusion that *C is the rightmost one* can be made. Hence, the main goal of our MORE is to first explicitly extract and encode basic first-order spatial relations among objects and then try to further infer multi-order relations for generating comprehensive captions.

Concretely, as shown in the Fig.1 (green background part), MORE consists of two components: Spatial Layout Graph Convolution (SLGC) and Object-centric Triplet Attention Graphs (OTAG). First, to capture the concepts of basic first-order relation, SLGC adaptively introduces spatial semantics into the edges of an object graph. Afterward, inside the OTAG, triplets in the form of $<node_1,edge,node_2>$ are extracted from the previous graph as the basic descriptors of first-order inter-object relations, then several such object-centric triplet graphs are constructed where the triplets targeting the same object serve as nodes within the same triplet graph. On top of these graphs, the attention-based

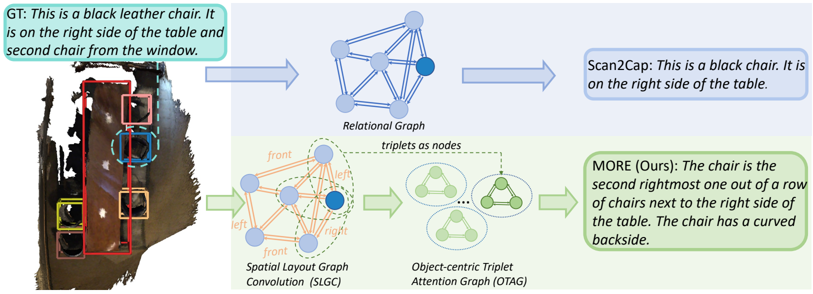
GT: *This is a black leather chair. It is on the right side of the table and second chair from the window.*
Relational Graph
Scan2Cap: *This is a black chair. It is on the right side of the table.*
triplets as nodes
front
left
left
right
front
...
Spatial Layout Graph Convolution (SLGC)
Object-centric Triplet Attention Graph (OTAG)
MORE (Ours): *The chair is the second rightmost one out of a row of chairs next to the right side of the table. The chair has a curved backside.*

Fig. 1. The comparison of our proposed Multi-Order RElation mining model (MORE) with the previous method (i.e., Scan2Cap [11]). The core components of our MORE and the Scan2Cap are distinguished with green and blue background, respectively. Scan2Cap treats the inter-object relations as by-products derived from node feature learning in the relational graph, thus the diverse spatial relations are under-explored. While in our MORE, we model such relations via a Spatial Layout Graph Convolution (SLGC) and Object-centric Triplet Attention Graphs (OTAG) to progressively encode more complex spatial relations. For simplicity, we only show the caption of one specific object (in the dashed circle) from different models. It is clear that our method can describe more complex relations.

multi-order relation reasoning is performed within each of them. The advantage of using such triplets is that the basic relations are more explicitly preserved, and we can also flexibly extend a triplet to cover larger contexts. Finally, our caption decoder receives the target object feature and performs node aggregation on OTAG as the context for sentence generation. As illustrated by our captioning results in Fig.1, the captions generated by MORE are more descriptive and comprehensive compared with the baseline method.

In summary, our contributions are threefold: (1) We propose a Multi-Order RElation mining model (MORE) for 3D dense captioning, which can generate more descriptive and comprehensive captions for each object. (2) Within the MORE, a Spatial Layout Convolution (SLGC) and Object-centric Triplet Attention Graphs (OTAG) are proposed and coupled together, where the former semantically encodes basic first-order spatial relations among objects in a 3D scene, and the latter infers the multi-order relations via attention-based graph reasoning. (3) Extensive experimental results prove that our MORE achieves superior performances than existing 3D dense captioning methods on prevalent benchmarks.

2 Related Work

Image and Video Captioning. Generating captions for 2D visual data (i.e., images and videos) has recently attracted significant research interest [10, 9, 13, 20, 39, 45, 47, 23, 37]. It is acknowledged that exploring the inter-object relationship benefits the caption generation, and such an idea has been widely investigated [45, 47, 23, 37, 44, 20]. Yang et al. [44] directly adopt the global image feature as context. To further introduce extra linguistic prior upon various object relations, scene graph detectors are utilized by some image captioning methods [45, 47] to parse the give image and assign textual tags to the relations. However, such detectors require expensive annotations to train. As an alternative, the part-of-speech tags are utilized as the prior to explicitly encode relations between objects in [20]. Although they are effective in 2D image and video caption generation, the spatial structures are much more complicated in the 3D scene, hence they can not be directly transferred to 3D dense captioning task.

3D Dense Captioning and Visual Grounding. Recently, investigating the 3D point cloud data and natural language has become a trending research topic [5, 1, 48, 16, 14, 11]. Among the pioneer works, Chen et al. [5] and Achloptas et al. [1] first proposed two datasets, referred to as ScanRefer and ReferIt3D, respectively, which both contain descriptions for real-world 3D objects in ScanNet [12]. On top of them, 3D visual grounding [5, 1, 48, 16, 14] and 3D dense captioning [11], as two dual tasks, are concurrently investigated, where the former focuses on localizing 3D objects described by natural language queries and the latter aims at generating descriptions for 3D objects in RGB-D scans. Exploring object relations is essential for both tasks, and many relevant attempts have been made in recent works [48, 16, 14, 15, 50, 46, 11].

In the 3D visual grounding task, earlier works, namely TGNN [16] and InstanceRefer [48], construct a directed instance graph with instance features as vertices and relative instance coordinates as edges. Later, in order to capture the object-object and object-scene co-occurrence, FFD [14] develops a multilevel proposal relation graph module with a geometric structure feature of each bounding box encoded in each graph node. Aiming at a unified intra- and intermodalities modeling scheme, Transfer3D [15], 3DVG [50], and SAT [46] adopt a standard Transformer architecture [34] for promoting inter-object relations. However, these methods overlooked the importance of encoding the multi-order visual object relations, which might be because that the key to improve grounding performance is learning the explicit correspondence between vision and language. And existing methods still struggle at establishing such correspondence due to the lack of semantics in the point clouds data [46].

In the 3D dense captioning task, Scan2Cap [11] first proposes a relational graph implemented with a static version of EdgeConv [40] to enhance object relation representation. However, the inter-object relations are only treated as by-products of graph node recognition, thus leads to sub-optimal results. Theoretically, 3D visual relation detectors [36, 3, 41] can mitigate such the problem, however, the current results delivered by them are unsatisfactory when faced with highly unrestricted scene compositions [49, 25]. Therefore, in this paper, we aim to develop a relation mining method which can properly encode both the basic first-order and complex multi-order relations contained in the 3D scene, so as to benefit more comprehensive caption generation.

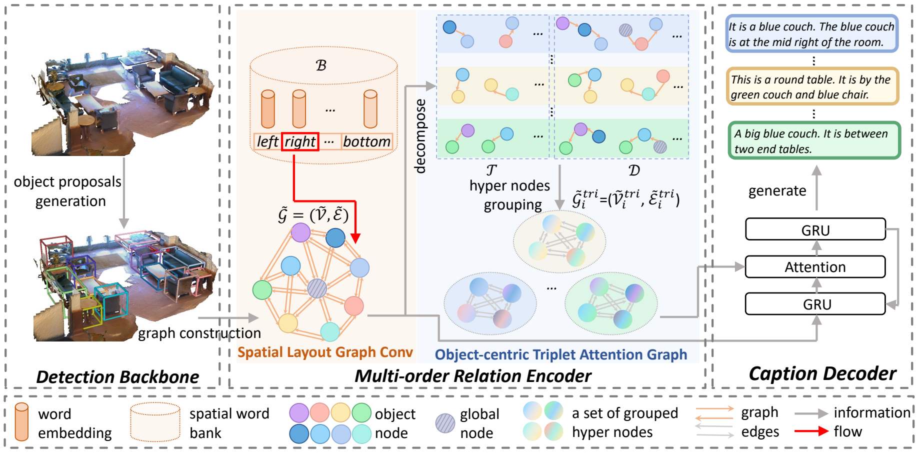
object proposals generation
graph construction
Detection Backbone
$\mathcal{B}$
left
right
...
bottom
$\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$
Spatial Layout Graph Conv
decompose
$\mathcal{T}$
$\mathcal{D}$
hyper nodes grouping
$\tilde{\mathcal{G}}_i^{tri}=(\tilde{\mathcal{V}}_i^{tri}, \tilde{\mathcal{E}}_i^{tri})$
Object-centric Triplet Attention Graph
Multi-order Relation Encoder
It is a blue couch. The blue couch is at the mid right of the room.
This is a round table. It is by the green couch and blue chair.
A big blue couch. It is between two end tables.
generate
GRU
Attention
GRU
Caption Decoder
word embedding
spatial word bank
object node
global node
a set of grouped hyper nodes
graph edges
information flow

Fig. 2. The overall framework of our proposed method, which consists of three parts: the detection backbone, the multi-order relation encoder, and the caption decoder. Given a 3D scene represented by point clouds, the detection backbone extracts a set of object proposals. Then, based on the objects, first-order and multi-order spatial relations are progressively encoded through a novel Spatial Layout Graph Convolution (SLGC) and several Object-centric Triplet Attention Graphs (OTAG), respectively. Finally, the OTAG's output, which encapsulates rich spatial relational cues, are served as the context to aid comprehensive caption generation. To keep the figure concise, we omit part of the object nodes and their corresponding captions, as well as the attention calculation of each triplet graph.

3 Multi-order Relation Mining Network

The overall framework of our Multi-Order RElation mining (MORE) network consists of three main components: an object detection backbone, a multi-order relation encoder, and a caption decoder. As shown in Fig.2, given a point cloud as input, the detection backbone extracts several 3D object proposals with bounding boxes (Sec.3.1). Then, as the core component of our framework, the multi-order relation encoder first takes object proposals as input and explicitly encodes the first-order spatial relations via our proposed Spatial Layout Graph Convolution (SLGC). During this process, SLGC maintains a bank of spatial words and dynamically selects word embeddings from it as graph edges, so as to narrow the gap between point cloud and relational concepts. The resulting graph is decomposed into triplets and recomposed as several Triplet Object-centric Attention Graphs (OTAG) for multi-order spatial relation reasoning (Sec.3.2). Finally, the decoder takes updated node features from OTAG as inputs to incorporate contextual cues with an attention module and generates language descriptions for the corresponding objects (Sec.3.3).

3.1 Detection Backbone

Given a 3D scene input represented by a point cloud $\mathcal{P} = \{(p_i, f_i)\}_{i=1}^{N_P}$ (N_P is the number of points), where $p_i \in \mathbb{R}^3$ and $f_i \in \mathbb{R}^3$ are the coordinates (x-y-z) and the color (r-g-b) of the i-th point, respectively, we adopt the PointNet++ [31] backbone and the voting module in VoteNet [30] to extract a set of object proposals. We denote these object proposals as $\mathcal{O} = \{(x_i, c_i)\}_{i=1}^{N_O}$ (N_O is the number of valid object proposals), where $x_i \in \mathbb{R}^{128}$ is the i-th object proposal's visual feature and $c_i = (b^x, b^y, b^z, b^h, b^w, b^l)$ is the location indicator of the corresponding bounding box. b^x, b^y, b^z are the coordinates of the box center and b^h, b^w, b^l are height, width, length of the box, respectively.

3.2 Multi-order Relation Encoding

The key to generating accurate and diverse captions for 3D real world is to capture or infer the complicated spatial relations, however, this is overlooked by the previous method [11]. As the core component of our method, the multi-order relation encoder takes object proposals $\mathcal{O}$ as input, and is responsible for explicitly encoding the basic first-order spatial relations by Spatial Layout Graph Convolution (SLGC) and further inferring possible multi-order relations via Object-centric Triplet Attention Graphs (OTAG). In the rest of this section, we will elaborate on each of them.

Spatial Layout Graph Convolution. Following [11], we wish to connect each object with its K nearest neighboring instances, and thereby construct an object graph to represent the spatial layout of the whole scene, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Note that in addition to the object proposals, we also add a global node into $\mathcal{V}$ to model the interaction between each object and the whole scene. Formally, $\mathcal{V} = \{v_i\}_{i=1}^{N_O+1}$, where $v_j = x_j$ ($j = 1, \ldots, N_O$) and $v_{N_O+1} = g$. g is the feature of global node calculated by averaging all object proposals' features.

As for the edges, conventional methods [11, 14] treat the graph edges $\mathcal{E}$ as the by-products derived from node representation learning, thus the edges do not have clear semantics. So we construct our spatial layout graph by delicately encoding basic first-order spatial relational concepts that are transferred from the linguistic embedding space.

First, we maintain a spatial word bank $\mathcal{B}$ to provide semantic concepts of basic first-order spatial relations:

$$\mathcal{B} = \{\text{``}left\text{''}, \text{``}right\text{''}, \text{``}front\text{''}, \text{``}behind\text{''}, \text{``}besides\text{''}, \text{``}top\text{''}, \text{``}bottom\text{''}\}. \quad (1)$$

For illustration purpose, we list the spatial words in $\mathcal{B}$, while we use their corresponding GloVE [29] word embeddings in our model. These relations can be divided into horizontal and vertical subsets, $\mathcal{B}_h$ and $\mathcal{B}_v$, respectively[4].

Next, we need to select a corresponding item from the word bank as the edge to describe the spatial relationship of the two connected nodes, however, this is nontrivial in the 3D world, as the varying viewpoints may bring ambiguity

[4] "top" and "bottom" are the vertical relations we focus on.

to the horizontal spatial relations. To this end, we select the horizontal spatial word in an adaptive manner, and the less ambiguous vertical spatial word based on rules. To compute the horizontal spatial relation between objects i and j, we first calculate the relative coordinates $c_{j,i}$ and object feature $v_{j,i}$, and combine them to predict a distribution $\alpha_{j,i} \in \mathbb{R}^{N_h}$ of the horizontal relational words:

$$\begin{aligned} &v_{j,i} = W_v[v_i; v_j - v_i], \quad c_{j,i} = W_c[b^x_j - b^x_i; b^y_j - b^y_i], \\ &\alpha_{j,i} = \text{softmax}(W_\alpha \tanh(v_{j,i} + c_{j,i})), \end{aligned} \tag{2}$$

where $[;]$ represents the concatenation operation, and $W_v \in \mathbb{R}^{256\times128}$, $W_c \in \mathbb{R}^{2\times128}$, and $W_\alpha \in \mathbb{R}^{128\times N_h}$ are weight matrices. Afterward, the horizontal spatial relational feature $r^h \in \mathbb{R}^{300}$ can be calculated by a weighted combination of the corresponding items of the spatial word bank:

$$r^h = \alpha_{j,i} \cdot \mathcal{B}_h, \tag{3}$$

where $\cdot$ is the dot product operation. In this way,

As for the vertical spatial relations, namely “top” and “bottom”, we design a metric to directly infer vertical relation between a given pair of objects i and j. We first generate the box corner coordinates $b^c_i, b^c_j \in \mathbb{R}^8$ according to their initial box coordinates c_i and c_j. Then, we calculate the relative vertical distance between each pair of corner points between objects i and j, and obtain the pairwise distances $d^v_{j,i} \in \mathbb{R}^{64}$. The vertical spatial relations can then be inferred by taking the sign of the maximum or the minimum element of $d^v_{j,i}$. And we can formulate the process of obtaining vertical spatial relational feature r^v as:

$$r^v = \mathbb{I}(\min(d^v_{j,i}) > 0) \times \mathcal{B}^{top}_v + \mathbb{I}(\max(d^v_{j,i}) < 0) \times \mathcal{B}^{bottom}_v \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 when the condition inside the bracket is satisfied, otherwise 0, and $\mathcal{B}^{top}_v$ and $\mathcal{B}^{bottom}_v$ are the corresponding items in the word bank, respectively. Finally, the first-order spatial relation between object i and j, also known as the directed graph edge $e_{j,i}$, can be obtained through combining two relational features r^h and r^v:

$$e_{j,i} = W_e[r^v; r^h], \tag{5}$$

where $W_e \in \mathbb{R}^{600\times128}$ is the projection matrix. As for the global node g, we manually assign its center coordinates as the location of the scene’s center for the horizontal relational feature calculation. And the vertical relational feature is an all-zero vector due to that no bounding box is available for the whole scene.

At this point, we have constructed the initial spatial layout graph $\mathcal{G}$ and the basic first-order spatial relations between objects have been incorporated into the graph edges via the semantically rich word embeddings. Then, we perform message passing upon the spatial layout graph to enhance the node features by letting them aggregate information from neighboring nodes:

$$\begin{aligned} &\beta_{j,i} = (W_1 v_i)^{\mathrm{T}}(W_2 v_j + W_3 e_{j,i}), \quad \hat{\beta}_{j,i} = \frac{\exp(\beta_{j,i})}{\sum_{k\in\mathcal{N}(i)} \exp(\beta_{k,i})}, \\ &v'_i = W_4 v_i + \sum_{j\in\mathcal{N}(i)} \hat{\beta}_{j,i} \odot (W_5 v_j + W_6 e_{j,i}), \end{aligned} \tag{6}$$

where the $\odot$ represents the element-wise multiplication with broadcast operation. Note that when computing aggregation weights, we also take the semantic edges into consideration. The above Spatial Layout Graph Convolution (SLGC) can be stacked multiple times and we denote the final resulting graph as $\widetilde{\mathcal{G}} = (\widetilde{\mathcal{V}}, \widetilde{\mathcal{E}})$, whose node and edge features can be denoted as $\widetilde{\mathcal{V}} = \{\widetilde{v}_i\}_{i=1}^{N_O+1}$ and $\widetilde{\mathcal{E}} = \{\widetilde{e}_{i,j}\}_{i,j=1}^{N_O+1}$.

Object-centric Triplet Attention Graph. Theoretically, by conducting multiple rounds of message passing with stacked SLGC, the model can become aware of the multi-order relations, however, we find in our experiments (Table.4) that this can not achieve the expected effects since the captioning performance does not improve with more layers. We postulate such a phenomenon might be caused by the over-smoothing problem [7, 51], and the experimental results (will be analyzed in detail) in Table.4 support our assumption to some extent. Thus we decide to not solely rely on the SLGC, but further specifically design an object-centric graph that preserves more information of the target objects and the relations.

Concretely, based on the first-order spatial relation aware graph $\widetilde{\mathcal{G}}$ that we obtained from SLGC, we directly extract a set of <*node*$_1$*,edge,node*$_2$> triplets from it, and the *node*$_2$ is the target object that we wish to generate caption for. Formally, these triplets can be represented as $\mathcal{T} = \{< \widetilde{v}_j, \widetilde{e}_{j,i}, \widetilde{v}_i >\}_{j \in \mathcal{N}(i), i=1:N_O}$. Note that we do not extract triplets targeting the global node, since currently we are focusing on generating captions for the objects but not the whole scene. Such triplets are a combined representation of a target object and its surrounding contexts, as well as their relations, and the target object is emphasized in each triplet graph so that its information will not be overwhelmed in the following operations. To enlarge the context, we can further extend these triplets via constructing a new set of quintuplets which can be represented as $\mathcal{Q} = \{(\widetilde{v}_k, \widetilde{e}_{k,j}, \widetilde{v}_j, \widetilde{e}_{j,i}, \widetilde{v}_i)\}_{k \in \mathcal{N}(j), j \in \mathcal{N}(i), i=1:N_O}$. In practice, we find these extended quintuplets to be helpful.

We then take the triplets/quintuplet ($\mathcal{T}/\mathcal{Q}$) as hyper-nodes, and construct object-centric graphs upon them. The triplets/quintuplets targeting the same object node are regarded as related hyper-nodes and we will further model their relations as a object-centric triplet attention graph. We denote the recomposed triplet graph that is centered at object i as $\mathcal{G}_i^{tri} = (\mathcal{V}_i^{tri}, \mathcal{E}_i^{tri})$ $(i = 1, \ldots, N_O)$. Since the hyper-nodes from $\mathcal{T}$ and $\mathcal{Q}$ have different formats, we first unify their channel dimensions by mapping them into the same lower-dimensional space:

$$v_{j,i} = W_t[\widetilde{v_j}; \widetilde{e}_{j,i}; \widetilde{v_i}], \quad v_{k,j,i} = W_q[\widetilde{v_k}; \widetilde{e}_{k,j}; \widetilde{v_j}; \widetilde{e}_{j,i}; \widetilde{v_i}], \tag{7}$$

where $W_t \in \mathbb{R}^{384 \times 128}$ and $W_q \in \mathbb{R}^{640 \times 128}$ are projection matrices. Then, the hyper-nodes can be represented as $\mathcal{V}_i^{tri} = \{v_{j,i}\}_{j \in \mathcal{N}(i)} \cup \{v_{k,j,i}\}_{k \in \mathcal{N}(j), j \in \mathcal{N}(i)}$. As for the edges $\mathcal{E}_i^{tri}$, we connect each pair of nodes in $\mathcal{V}_i^{tri}$ by computing the edge weights via attention:

$$e_{\{j,i\},\{k,i\}} = \frac{\exp(\sigma(W_7 v_{j,i})\sigma(W_8 v_{k,i}))}{\mathcal{Z}_{j,i}}, \tag{8}$$

where $\{j,i\}$ and $\{k,i\}$ represents two hyper-nodes of triplets (or quintuplets), W_7, W_8 are projection matrices, and σ is an activation function (e.g., ReLU). We denote the normalization denominator as $\mathcal{Z}_{j,i}$ to avoid clutter. We then aggregate neighbor information through

$$v'_{j,i} = \sum_{k\in\mathcal{N}(i)} e_{\{j,i\},\{k,i\}} W_9 v_{k,i}. \tag{9}$$

Here we only include triplet-based hyper-nodes for clarity, and quintuplets can be similarly incorporated. By performing the above message passing among hyper-nodes, the model can infer multi-order relations while significantly preserving the target object's information. The final object-centric triplet graph after pair-wise node attention calculation (i.e., Eq.(10) and Eq.(11)) is denoted as $\widetilde{\mathcal{G}}_i^{tri} = (\widetilde{\mathcal{V}}_i^{tri}, \widetilde{\mathcal{E}}_i^{tri})$ $(i = 1, \ldots, N_O)$.

3.3 Caption Generation

It is a common practice to make the caption decoder responsible for incorporating contextual cues generated by the encoder [2, 11]. We design our decoder based on a two-layer GRU network, and equip it with an attention module in between the two layers to aggregate OTAG's object-centric nodes that are aware of multi-order spatial relations.

In particular, the caption is iteratively generated in a word-by-word manner. At the t-th time step, the first GRU takes the concatenation of the GloVE embedding of the word $w^{(t-1)}$, the previous hidden state $h_2^{(t-1)}$ of the second GRU, as well as the initial visual feature x_i of the target object proposal as inputs, and update the its hidden state $h_1^{(t)}$ as:

$$h_1^{(t)} = \mathrm{GRU}_1([w^{(t-1)}; h_2^{(t-1)}; x_i]; h_1^{(t-1)}), \tag{10}$$

Next, depending on the updated hidden state $h_1^{(t)}$ of the first GRU, we adaptively compute aggregation weights for $\widetilde{\mathcal{G}}_i^{tri}$'s node features, which contains sptatial relation cues between the target object and its neighbors, and these node features are then aggregated by a weighted sum:

$$\begin{gathered}\gamma_i = W_\gamma(\tanh(W_{10} h_1^{(t)} + W_{11}\widetilde{v}_i^{tri})), \quad \hat{\gamma}_i = \frac{\exp(\gamma_i)}{\sum_{j\in\mathcal{N}(i)}\exp(\gamma_j)}, \\ \hat{v}^{(t)} = \sum_{i=1}^{N_O} \hat{\gamma}_i \odot \widetilde{v}_i^{tri},\end{gathered} \tag{11}$$

where $\hat{v}^{(t)}$ is the resulting contextual feature vector with spatial relations embedded in it. Afterward, the contextual vector $\hat{v}^{(t)}$, together with the hidden state $h_1^{(t)}$ of the first GRU, are fed into the second GRU to obtain its updated hidden state $h_2^{(t)}$:

$$h_2^{(t)} = \mathrm{GRU}_2([\hat{v}^{(t)}; h_1^{(t)}]; h_2^{(t-1)}), \tag{12}$$

Finally, $h_2^{(t)}$ is leveraged to predict the current word $w^{(t)}$ through a linear classifier. The details about the captioning loss and model training process can be found in the supplementary materials.

4 Experiments

4.1 Experimental Setup

Dataset. Following previous work [11, 6], we use the ScanRefer [5] dataset, which consists of 51,583 descriptions for 11,046 objects in 800 ScanNet [12] scenes. The descriptions include the appearance of the objects (e.g. "this is a black tv"), and the spatial relations between the target object and surrounding objects (e.g. "the cabinet is next to the desk"). The dataset is split into train/val sets with 36,665 and 9,508 samples respectively following the ScanRefer [5] benchmark. Scenes in the train and val sets are disjoint with each other. Since the test set has not been officially released, all experimental results and analysis in the following sections are conducted on top of the val set.

Evaluation Metrics. To jointly evaluate the quality of detected bounding boxes and generated captions, a combined metric $m@k\text{IoU} = \frac{1}{N}\sum_{i=1}^{N} m_i u_i$ defined in [11] is adopted, where u_i is set to 1 if the IoU score for the i^{th} box is greater than k, otherwise 0, and m can be one of the caption metrics, such as CiDEr [35], BLEU-4 [28], METEOR [4], and ROUGE [24], abbreviated as C, B-4, M, R in the following part, respectively. Meanwhile, mean average precision (mAP) at specified IoU threshold is utilized to evaluate the object detection performance.

Implementation Details. In our experiment, we randomly sample 40,000 points from every ScanNet scenes. The maximum number of object proposals K is set as 256. Unless otherwise specified, we utilize the color (r-g-b) of each point as the input visual feature to conduct experiments. The number of stacked SLGC layers are set as 1 and 2 when the point color and multi-view feature are adopted, respectively. We train our model with the Adam optimizer [22], and set the learning rate to $1e^{-4}$, weight decay to $1e^{-5}$, and batch size to 12. Following [11], we adopt the same data augmentation strategy and truncate the descriptions longer than 30.

4.2 Comparison with State-of-the-art

We compare our method with the state-of-the-art approach Scan2Cap [11] on the ScanRefer [5] benchmark as shown in Table.1, where the VoteNet [30] is adopted as the detector and the entire network is trained end-to-end. Note that the results that we reproduced with the officially released code of Scan2Cap [11] have discrepancy with the results reported in the paper, and we report our reproduced results in Table.1, denoted as "Scan2Cap*". The methods in the first three rows are simple baselines provided in the previous work [11]. The results demonstrate that our method achieves consistent improvements across most of the evaluation metrics, especially on CiDEr. When using *"r-g-b"* color and multiview features as

Table 1. Comparison with state-of-the-art methods on the ScanRefer dataset with VoteNet [30] as the detector of all methods. "Scan2Cap*" represents the results we reproduced with the officially released code. We use subscript "rgb" and "mul" to denote using *"r-g-b"* color and multi-view feature as additional point features.

Method	C@0.25IoU	B-4@0.25IoU	M@0.25IoU	R@0.25IoU	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5
2D-3D Proj. [11]	18.29	10.27	16.67	33.63	8.31	2.31	12.54	25.93	10.50
3D-2D Proj. [11]	19.73	17.86	19.83	40.68	11.47	8.56	15.73	31.65	31.83
VoteNetRetr [30]	15.12	18.09	19.93	38.99	10.18	13.38	17.14	33.22	31.83
Scan2Cap$_{mul}$ [11]	56.82	34.18	26.29	55.27	39.08	***23.32***	***21.97***	***44.78***	32.21
Scan2Cap*$_{mul}$ [11]	53.88	32.71	25.64	53.87	38.11	22.63	21.60	44.06	31.47
MORE$_{mul}$	***62.91***	***36.25***	***26.75***	***56.33***	***40.94***	22.93	21.66	44.42	***33.75***
Scan2Cap$_{rgb}$ [11][5]	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57	29.13
Scan2Cap*$_{rgb}$ [11]	51.05	32.99	25.59	53.82	35.11	22.26	21.44	43.70	28.86
MORE$_{rgb}$	**58.89**	**35.41**	**26.36**	**55.41**	**38.98**	**23.01**	**21.65**	**44.33**	**31.93**

additional point features, compared with Scan2Cap, our method obtains 5.16% and 6.09% improvements on C@0.25IoU, as well as 3.78% and 1.86% improvements on C@0.5IoU in respectively. We emphasize the CIDEr metric because compared with other evaluation metrics, CiDEr shows higher agreement with consensus as assessed by humans [35]. Overall, the significant improvements on C@0.25IoU and C@0.5IoU can prove the superiority of our method.

Meanwhile, considering that the overall performances of the dense captioning can be affected by the detector, we adopt ground-truth bounding boxes as input to specifically compare the captioning performance as shown in Table.2. Methods in the first three lines are simple baselines provided in [11]. Since now the performance is no longer affected by the detection results, we omit the mAP and captioning results with the 0.25 IoU threshold. We observe from the table that our method surpasses the baseline with a large margin across all evaluation metrics. Besides, the improvements over Scan2Cap are more evident than in Table.1 where VoteNet is used as the detector. This indicates that when the context objects are more reliable, our method can benefit more and perform relation encoding with a higher quality, thereby promoting the captioning performances.

4.3 Comprehensive Analysis

Ablation Studies. We conduct ablation studies to verify the effectiveness of our proposed SLGC and OTAG designs, and the results are shown in Table.3. In the baseline method (the first row), we remove OTAG and the edge feature $e_{j,i}$ in Eq.(6) when performing message passing and keep other settings of SLGC the same. As shown in the second row, including SLGC improves C@0.5IoU by 1.79% comparing to the baseline, and other metrics are also improved.

This demonstrates that explicitly encoding the basic first-order spatial relational concepts when updating node features is helpful to improving the captioning accuracy. In the third row, OTAG alone brings 4.28% C@0.5IoU improvement over the baseline, which indicates that these recomposed object-centric graphs could effectively enhance our model's ability to capture multi-order relations,

[5] Results of this setting can be found in the supplementary materials of [11].

Table 2. Comparison with state-of-the-art methods on the ScanRefer dataset using ground-truth bounding boxes for all methods. Since the experimental results using ground-truth bounding boxes with *"r-g-b"* as additional point features are not reported by Scan2Cap, we only report our reproduced results with their officially released code.

	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU
OracleRetr2D [11]	20.51	20.17	23.76	50.98
Oracle2Cap2D [11]	58.44	37.05	28.59	61.35
OracleRetr3D [11]	33.03	23.36	25.80	52.99
Scan2Cap$_{mul}$ [11]	67.95	41.49	29.23	63.66
Scan2Cap*$_{mul}$ [11]	65.51	39.62	29.23	62.87
MORE$_{mul}$	**70.39**	**42.34**	**29.55**	**64.31**
Scan2Cap*$_{rgb}$ [11]	64.19	38.90	28.96	62.38
MORE$_{rgb}$	**67.15**	**43.52**	**29.55**	**65.09**

Table 3. Ablation studies of the individual components, including the SLGC for first-order relation encoding, and the OTAG for multi-order relation modeling.

SLGC	OTAG	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5
		33.67	20.92	20.87	42.60	29.14
✓		35.46	21.31	21.01	43.10	29.18
	✓	37.95	21.55	21.17	43.49	29.53
✓	✓	**38.98**	**23.01**	**21.65**	**44.33**	**31.93**

thus significantly boosting the performance. Combining the SLGC and OTAG, as shown in the last row, can achieve the highest performance, especially for the C@0.5IoU (38.98%), which outperforms the baseline with a large margin of 5.31%. These experiments demonstrate that the key components of our model (SLGC and OTAG) are both beneficial to generating high quality dense captions, and that our multi-order encoding of the inter-object relations is effective.

Why not stacking SLGC for multi-order relation modeling? Theoretically, multi-order relations can be modeled by message passing on a graph for several rounds so that one node can reach out to further neighbors, but as our studies in Table.4 demonstrate, such an approach cannot attain ideal outcomes. In the first four rows, we stacked SLGC of 1, 2, 3, and 4 layers, respectively, and then feed the output nodes directly to the the caption generation decoder as in Scan2Cap [11]. We can see that although stacking two layers of SLGC can bring clear improvements over using only one SLGC layer However, adding more layers of SLGC hurts the model's performance. When the number of the SLGC layers increase to 3 and 4, the model's performance gradually degrades. Since stacking SLGC only changes the node features we fed to the caption decoder, we conjecture that the performance degradation might be due to the over-smoothing issue as described in [7, 51]. We will further analyze this issue and provide evidence in the subsequent paragraph. Next, instead of stacking multiple layers of SLGC, we stack our OTAG on top of 1-layer SLGC. The results are shown in the last row of Table.4, and by comparing it with the SLGC-only results, we can observe that SLGC×1+OTAG outperforms the best results of stacking SLGC (SLGC×2). This proves that constructing OTAG is a more suitable approach of modeling multi-order relations under our scenario.

Table 4. Comparison of different configurations of message passing for multi-order relation modeling in the graph structure.

	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5	MADGap
SLGC×1	35.46	21.31	21.01	43.10	29.18	17.46
SLGC×2	38.06	22.26	21.35	43.64	29.98	14.49
SLGC×3	38.27	21.99	21.16	43.43	30.43	13.17
SLGC×4	36.58	21.46	21.05	42.93	29.94	12.28
SLGC×1+OTAG	**38.98**	**23.01**	**21.65**	**44.33**	**31.93**	**21.70**

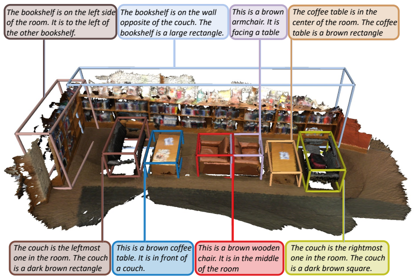
The bookshelf is on the left side of the room. It is to the left of the other bookshelf.
The bookshelf is on the wall opposite of the couch. The bookshelf is a large rectangle.
This is a brown armchair. It is facing a table
The coffee table is in the center of the room. The coffee table is a brown rectangle
The couch is the leftmost one in the room. The couch is a dark brown rectangle
This is a brown coffee table. It is in front of a couch.
This is a brown wooden chair. It is in the middle of the room
The couch is the rightmost one in the room. The couch is a dark brown square.

Fig. 3. The illustration of dense captions for objects within a whole scene predicted by our method. We distinguish the captions for each object with different colors.

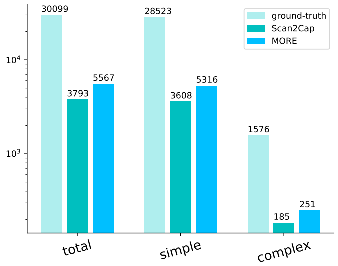

Fig. 4. Statistics of relational words in captions generated by different methods. “simple” and “complex” roughly represents first- and multi-order relations, respectively, and “total” is the sum of all relational words.

Why is OTAG more suitable? As shown in Eq.(10),(11),(12), the caption generation is conditioned on the target object feature x_i and its context feature $\hat{v}^{(t)}$, and the context feature is adaptively computed based on the decoding state and the encoder's graph nodes. Hence, obtaining distinctive node feature representation $\hat{v}^{(t)}$ from the multi-order relation encoder is the key to generate diverse captions. However, if the multi-order relation encoder is simply composed of multiple graph convolution layers, the node representation might suffer from the over-smoothing issue [7, 51] (i.e., features of the graph nodes from different classes would become indistinguishable when stacking multiple graph layers [7]), which hurts obtaining distinctive context features. Hence we conduct experiments to verify our OTAG is a more suitable graph layer that can learn more distinguishable node features to benefit the caption decoder.

In order to quantitatively evaluate the distinctiveness of graph nodes, we calculate the MADGap, a metric introduced in [7] to evaluate over-smoothness of graph nodes. The results are shown in the last column of Table.4. As can be observed, when the number of stacked graph layers increases from 1 to 4, the corresponding MADGap value decreases, which indicates that the node features gradually become more similar. Although 1-layer SLGC can achieve higher MADGap value, its overall performances of caption generation are inferior to the 2-layer SLGC, which is mainly due to that the spatial relations are not sufficiently encoded into the nodes with only one graph layer. When stacking more than 2 layers of SLGC, the MADGap and the overall captioning performances both decrease, demonstrating the over-smoothing is correlated with captioning quality degradation to some extent. Finally, combining 1-layer SLGC with our proposed OTAG, our method achieves much higher MADGap value than stacking any number of layers of SLGC, while also performs the best on caption generation. This indicates that the object-centric graph construction in OTAG can preserve distinctive information of the node features, make the aggregated context of in the caption decoder more diverse, and finally boost the captioning performances.

Relational words statistics. To give a more intuitive analysis on the advantages of our method, we inspect the performance improvement gained by our method in terms of relation capturing. Fig.4 compares the relational words in the sentences generated by Scan2Cap and our MORE, as well as in ground-truth annotations. Specifically, we maintain a dictionary of all the relational words in the corpus. Then we classify these words into "simple" and "complex" according to whether multiple objects should be jointly considered to support predicting the relation. As can be seen, comparing to Scan2Cap, our MORE can capture more relations when describing an object, which demonstrates its effectiveness. The relational word dictionary, and the split of "simple" and "complex" words can be found in the supplementary materials.

4.4 Qualitative Results

We further show the qualitative result in Fig.3. Note that to avoid clutter caused by low-quality object proposals, we directly use the ground-truth bounding box of each object. We can observe that our method is able to capture diverse spatial relations among objects. For example, relations like "on the left side of", "in front of" and "opposite of" are all properly leveraged to describe objects. Besides, our method can also accurately describe multi-order relations, such as "leftmost" and "rightmost", which are used to describe the two couches at both ends of the scene. The results demonstrate that our MORE is effective in capturing and describing diverse spatial relations for multiple objects in 3D world.

5 Conclusions

In this paper, we improved 3D dense captioning by proposing a novel relation modeling method, named Multi-order RElation Mining Network (MORE). We progressively modeled spatial relations by encoding first-order and multi-order ones with our proposed Spatial Layout Graph Convolution (SLGC) and Object-centric Triplet Attention Graphs (OTAG), respectively. Extensive experimental results demonstrated the effectiveness and the advantage of MORE over the current state-of-the-art method.

Acknowledgements

This work was supported in part by Shanghai Pujiang Program (No. 20PJ1401900) and Shanghai Science and Technology Program (No. 21JC1400600).

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: European Conference on Computer Vision. pp. 422–440. Springer (2020)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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. pp. 6077–6086 (2018)
3. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5664–5673 (2019)
4. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
5. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European Conference on Computer Vision. pp. 202–221. Springer (2020)
6. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. arXiv preprint arXiv:2112.01551 (2021)
7. Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., Sun, X.: Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3438–3445 (2020)
8. Chen, J., Pan, L., Wei, Z., Wang, X., Ngo, C.W., Chua, T.S.: Zero-shot ingredient recognition by multi-relational graph convolutional network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10542–10550 (2020)
9. Chen, S., Jiang, W., Liu, W., Jiang, Y.G.: Learning modality interaction for temporal sentence localization and event captioning in videos. In: European Conference on Computer Vision. pp. 333–351. Springer (2020)
10. Chen, S., Jiang, Y.G.: Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8425–8435 (June 2021)
11. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3193–3203 (2021)
12. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
13. Deng, C., Chen, S., Chen, D., He, Y., Wu, Q.: Sketch, ground, and refine: Top-down dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 234–243 (2021)
14. Feng, M., Li, Z., Li, Q., Zhang, L., Zhang, X., Zhu, G., Zhang, H., Wang, Y., Mian, A.: Free-form description guided 3d visual graph network for object grounding in point cloud. arXiv preprint arXiv:2103.16381 (2021)

15. He, D., Zhao, Y., Luo, J., Hui, T., Huang, S., Zhang, A., Liu, S.: Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2344–2352 (2021)
16. Huang, P.H., Lee, H.H., Chen, H.T., Liu, T.L.: Text-guided graph neural networks for referring 3d instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1610–1618 (2021)
17. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. In: IJCAI. pp. 765–771 (2021)
18. Jiao, Y., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: Suspected object matters: Rethinking model's prediction for one-stage visual grounding. arXiv preprint arXiv:2203.05186 (2022)
19. Jiao, Y., Jie, Z., Luo, W., Chen, J., Jiang, Y.G., Wei, X., Ma, L.: Two-stage visual cues enhancement network for referring image segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1331–1340 (2021)
20. Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S.: Dense relational captioning: Triple-stream networks for relationship-based captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6271–6280 (2019)
21. Kim, K., Billinghurst, M., Bruder, G., Duh, H.B.L., Welch, G.F.: Revisiting trends in augmented reality research: A review of the 2nd decade of ismar (2008–2017). IEEE transactions on visualization and computer graphics **24**(11), 2947–2962 (2018)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Li, X., Jiang, S.: Know more say less: Image captioning based on scene graphs. IEEE Transactions on Multimedia **21**(8), 2117–2130 (2019)
24. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
25. Milewski, V., Moens, M.F., Calixto, I.: Are scene graphs good enough to improve image captioning? arXiv preprint arXiv:2009.12313 (2020)
26. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4594–4602 (2016)
27. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10971–10980 (2020)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
29. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
30. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413 (2017)
32. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9339–9347 (2019)

33. Song, X., Chen, J., Wu, Z., Jiang, Y.G.: Spatial-temporal graphs for cross-modal text2video retrieval. IEEE Transactions on Multimedia (2021)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
36. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020)
37. Wang, D., Beck, D., Cohn, T.: On the role of scene graphs in image captioning. In: Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN). pp. 29–34 (2019)
38. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: European Conference on Computer Vision. pp. 18–34 (2020)
39. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7190–7198 (2018)
40. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog) **38**(5), 1–12 (2019)
41. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7515–7525 (2021)
42. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079 (2018)
43. Xiong, J., Hsiang, E.L., He, Z., Zhan, T., Wu, S.T.: Augmented reality and virtual reality displays: emerging technologies and future perspectives. Light: Science & Applications **10**(1), 1–30 (2021)
44. Yang, L., Tang, K., Yang, J., Li, L.J.: Dense captioning with joint inference and visual context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2193–2202 (2017)
45. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10685–10694 (2019)
46. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. arXiv preprint arXiv:2105.11450 (2021)
47. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV). pp. 684–699 (2018)
48. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: 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. pp. 1791–1800 (2021)

49. Zhang, H., Niu, Y., Chang, S.F.: Grounding referring expressions in images by variational context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4158–4166 (2018)
50. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)
51. Zhou, K., Huang, X., Li, Y., Zha, D., Chen, R., Hu, X.: Towards deeper graph neural networks with differentiable group normalization. Advances in Neural Information Processing Systems **33**, 4917–4928 (2020)