3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds

Daigang Cai[1], Lichen Zhao[1], Jing Zhang[†1], Lu Sheng[1], Dong Xu[2]
[1]College of Software, Beihang University, China, [2]The University of Sydney, Australia
{caidaigang, zlc1114, zhang_jing, lsheng}@buaa.edu.cn, dong.xu@sydney.edu.au

Abstract

Observing that the 3D captioning task and the 3D grounding task contain both shared and complementary information in nature, in this work, we propose a unified framework to jointly solve these two distinct but closely related tasks in a synergistic fashion, which consists of both shared task-agnostic modules and lightweight task-specific modules. On one hand, the shared task-agnostic modules aim to learn precise locations of objects, fine-grained attribute features to characterize different objects, and complex relations between objects, which benefit both captioning and visual grounding. On the other hand, by casting each of the two tasks as the proxy task of another one, the lightweight task-specific modules solve the captioning task and the grounding task respectively. Extensive experiments and ablation study on three 3D vision and language datasets demonstrate that our joint training framework achieves significant performance gains for each individual task and finally improves the state-of-the-art performance for both captioning and grounding tasks.

1. Introduction

There is increasing research interest in the intersection field between 3D visual understanding and natural language processing, such as 3D dense captioning [9] and 3D visual grounding [1, 7, 21, 50]. These two tasks push the advance of the intersection field along different directions (*i.e.*, from vision to language versus from language to vision), and encouraging progress has been achieved by separately solving each task. It still remains an open issue on whether it is possible to develop a unified framework to jointly solve the two closely related tasks in a synergistic fashion.

We observe that the two 3D vision-language tasks contain both shared and complementary information in nature, and it is possible to enhance the performance of both tasks if we treat one task as a proxy task of the other. On one hand, each of the two tasks can be decomposed into several sub-tasks, and some of these sub-tasks share the common objectives and network structures. For example, as shown in the previous vision-language works [1,7,9,21,44,47,50] on RGB-D scans, both 3D dense captioning and 3D visual grounding require: 1) a 3D object detector to detect the salient object proposals in a 3D scene, 2) a relation modeling module to model complex 3D relations among these detected objects, and 3) a multi-modal learning module to learn fused information from both visual features and textual features to generate sentences or produce bounding boxes based on each input sentence. On the other hand, the opposite procedures are also used to separately solve the two problems, namely, the captioning task is to generate a meaningful textual description from the detected boxes (*i.e.*, from vision to language), while the grounding task is to locate the desired box by understanding a given textual description (*i.e.*, from language to vision).

Moreover, the 3D point clouds generated from RGB-D scans often contain rich and complex relations among different objects, while the corresponding RGB data provides more fine-grained attribute information, such as color, texture, and materials. Thus, the RGB-D scans intrinsically contain rich and abundant attribute and relation information for enhancing both 3D captioning and 3D grounding tasks. However, we empirically observe that the 3D dense captioning task is more object-oriented, which tends to learn more attribute information of the target objects (*i.e.*, the objects of interest) in a scene and only the primary relationship between the target object and its surrounding objects. In contrast, the 3D visual grounding task is more relation-oriented, which focuses more on the relations between objects and distinguishes different objects (especially the objects from the same class) based on their relations. Thus, it is desirable to develop a joint framework to unify both 3D dense captioning and 3D visual grounding tasks and take advantage of each other for improving the performance of both tasks.

To this end, in this work, we propose a joint framework by unifying the distinct but closely related 3D vision-

† Corresponding author: Jing Zhang.

language tasks of 3D dense captioning and 3D visual grounding. Specifically, the proposed framework consists of three main modules: (1) a 3D object detector, (2) an attribute and relation-aware feature enhancement module, and (3) a task-specific grounding or captioning head. Specifically, the 3D object detector and the feature enhancement module are task-agnostic, which are designed for collaboratively supporting both captioning and grounding tasks. The two modules output the object proposals as the initial localization results of the potential objects in a scene, as well as the improved features within the proposals by integrating both attribute information from each object proposal and the complex relations between multiple proposals. With the strong task-agnostic modules, the task-specific captioning head and grounding head are designed as lightweight networks for dealing with each task, which consist of a lightweight transformer-based module together with simple preprocessing modules (*i.e*., the Query/Key/Value generation modules) and lightweight postprocessing modules (*i.e*., the word prediction or bounding box selection module). In this way, the 3D captioning and 3D visual grounding tasks can be cast as the proxy task of each other. In other words, the more object-oriented captioning task can provide more attribute information to potentially improve the grounding performance, while the more relation-oriented grounding task can help improve the captioning results by enhancing the captioning task with more relation information. Moreover, our joint framework also inspires the insights of the design of each individual captioning network and grounding network.

The contribution of this work is two-fold: (1) By analyzing both 3D dense captioning and 3D visual grounding tasks, we propose a unified framework to jointly solve the two distinct but closely related tasks by using our simple and strong network structure, which consists of a task-agnostic module with a 3D object detector and an attribute and relation-aware feature enhancement module, and two lightweight task-specific modules (*i.e*., a captioning head and a grounding head). (2) Extensive experiments conducted on three benchmark datasets ScanRefer [7], Scan2Cap [9], and Nr3D dataset [1] demonstrate our joint framework achieves the state-of-the-art results for both 3D dense captioning and 3D visual grounding tasks.

2. Related Work

2D Vision and Language tasks. Deep learning technologies have been extensively studied in various 2D vision and language tasks, such as visual grounding [15, 26, 35, 45]. image captioning & dense captioning [2, 11, 17, 18, 42], visual question answering [2, 4, 43] and text-to-image generation [25]. These impactful research problems advance the intersection research field between computer vision and natural language processing. With the rapid development of deep learning, researchers introduced several collaborative methods (*e.g*., speaker-listener models [3, 46]) to solve various 2D vision and language tasks jointly. However, these models focus on 2D image-based tasks, while our method focuses on RGB-D-based tasks, where different types of data to be handled in our work require different network design strategies. Specifically, we propose a carefully designed task-agnostic feature enhancement module and the lightweight task-specific captioning and grounding heads, which all build upon the transformer architecture. Recently, several joint frameworks [8,23,24,27,41,48] focus on learning more generalizable image-text representations through a **cumbersome** model (*e.g*., VilBERT [23]) by using abundant and diverse 2D vision and language datasets. By contrast, based on in-depth analysis of the intrinsic properties of RGB-D scans and the characteristics of both 3D captioning and grounding tasks, our carefully designed joint learning framework with **lightweight** modules can effectively solve both tasks in a synergistic fashion without relying on a huge amount of paired training data.

3D Dense Captioning and Visual Grounding. Deep learning in 3D data has attracted a great deal of interest [10, 13, 20, 22, 32–34, 39, 40, 49, 51]. Recently, some dense captioning and visual grounding tasks tailored to 3D data are proposed. For example, some researches [9] proposed the 3D dense captioning methods and achieved impressive results by explicitly modeling the relation between different objects [9]. However, the dense captioning task is more object-oriented, which often focuses on the precise attribute descriptions based on the object appearance and thus the complex 3D geometrical relations among different objects might be ignored (even though they are intrinsically contained in the 3D data). As a result, the generated captions may be monotony.

Except for 3D dense captioning, visual grounding on 3D point clouds [1, 7, 14, 16, 44, 47, 50] has also attracted increasing research interest. Chen *et al*. [7] introduced the ScanRefer dataset for localizing objects by using natural language descriptions. Most recent 3D visual grounding methods [7, 16, 47] are composed of two stages. In the first stage, a 3D object detector or a panoptic segmentation model is applied to generate the target object proposals from the input scenes. In the second stage, a referring module is used to match the most relevant regions from the selected object proposals and the query sentences. These methods mainly focus on how to model the complex relations based on the object detection results, and pay less attention to the appearance features that characterize different objects, especially the objects within the same class. In other words, the current grounding methods are more relation-oriented.

Our joint framework takes advantage of the overlooked attribute information in the grounding task through the help of the more object-oriented captioning task, and employs

the relatively less explored relation information in the captioning task to increase the variety of generated sentences with the help of the more relation-oriented grounding task.

3. Methodology

In this section, we describe the technical details of our framework. As shown in Fig. 1(a), our framework consists of three modules: 1) the object detection module, 2) the attribute and relation-aware feature enhancement module, and 3) the task-specific captioning head and grounding head. The object detection module and feature enhancement module are task-agnostic and shared by both tasks. The captioning and grounding heads are task-specific with the lightweight transformer-based network structures for the captioning and grounding tasks, respectively. Specifically, the point clouds are encoded by the VoteNet [31] object detection module with an improved bounding box modeling method to more precisely locate the salient objects and produce the initial object proposals. Then the proposal features are enhanced through a task-agnostic attribute and relation-aware feature enhancement module to generate the enhanced object proposals. The enhanced object proposals are then fed into the captioning head and grounding heading for the dense captioning task and the visual grounding task, respectively, and generate the final result for each task.

3.1. Detection Module

The input of the detection module is the point cloud $\boldsymbol{P} \in \mathbb{R}^{N\times(3+K)}$, which represents the whole 3D scene by N 3D coordinates together with K-dimensional auxiliary features. Here, we adopt the same 132-dimensional auxiliary features as in [7, 9], which include the pretrained 128-dimensional multi-view appearance features [7], 3-dimensional normals, and 1-dimensional height of each point above the ground.

We use VoteNet [31] as our detection module. Since the success of both captioning and grounding tasks relies on precise localization of initial object proposals together with discriminative features, we borrow the idea from the anchor-free FCOS method [36] to generate the initial object proposals by predicting the distance between the voting point and each side of the object proposal.

3.2. Attribute and Relation-aware Feature Enhancement Module

The initial object proposal features produced by the detection module are discriminative with respect to different object classes, thanks to the detection-related loss. However, they are unaware of the fine-grained object attributes (*e.g.*, object positions, colors, and materials), especially for the within-class objects, and the complex relations among different objects, which are the key to the success of both 3D captioning and 3D grounding tasks. Hence, we further propose an attribute and relation-aware feature enhancement module to strengthen the features for each proposal and better model the relations between proposals. Motivated by the Transformer encoder structure [37], we model the proposal feature enhancement module as two multi-head self-attention layers with additional attribute encoding module and relation encoding module, where the attribute or relation encoding module is composed of several fully connected layers.

The attribute encoding module. To aggregate the attribute features and the initial object features, we encode the auxiliary bounding box attribute related features (*i.e.*, a 155-dimensional feature via a concatenation operation on the 27-dimensional box center and corner coordinates, and the 128-dimensional multi-view RGB features that potentially contain the attribute information such as colors and materials) into a 128-dimensional attribute embedding by using a fully connected layer. The attribute embedding has the same dimension as the initial object proposal features. It can then be added to the initial proposal features to enhance the initial object features with more attribute information.

The relation encoding module. Motivated by [50], we also encode the pairwise distances between any two object proposals to capture the complex object relations. Different from [50], we encode not only the (inverse) relative Euclidean distances (*i.e.*, $\boldsymbol{Dist} \in \mathbb{R}^{M\times M\times 1}$) but also three pairwise distances between any two centers of the initial object proposals along x, y, z direction (*i.e.*, $\boldsymbol{D_x}$, $\boldsymbol{D_y}$, $\boldsymbol{D_z}$ $\in \mathbb{R}^{M\times M\times 1}$) to better capture object relations along different directions, where M is the number of initial object proposals. All four spatial proximity matrices ($\boldsymbol{D_x}$, $\boldsymbol{D_y}$, $\boldsymbol{D_z}$, and $\boldsymbol{Dist}$) are then aggregated along the channel dimension and fed into fully connected layers to produce the relation embeddings with the channel dimension H matches the number of attention heads (*i.e.*, $H = 4$ in our implementation) in the multi-head attention module. Each relation embedding (with the size of $M \times M \times 1$) is then added with the similarity matrix (*i.e.*, the so-called attention map) generated from each head of the multi-head self-attention module.

Note the task-agnostic 3D object detector and the feature enhancement module can produce more accurate localization results and improved object features for both captioning and grounding tasks, and thus we can use more lightweight task-specific captioning head and grounding head in our framework which are simpler than the state-of-the-art methods [9, 50]. For both task-specific heads, we adopt similar lightweight 1-layer multi-head cross-attention-based network structures together with simple preprocessing modules (*i.e.*, Query/Key/Value generation as shown in Fig. 2) and postprocessing modules (*i.e.*, word prediction or BBox selection).

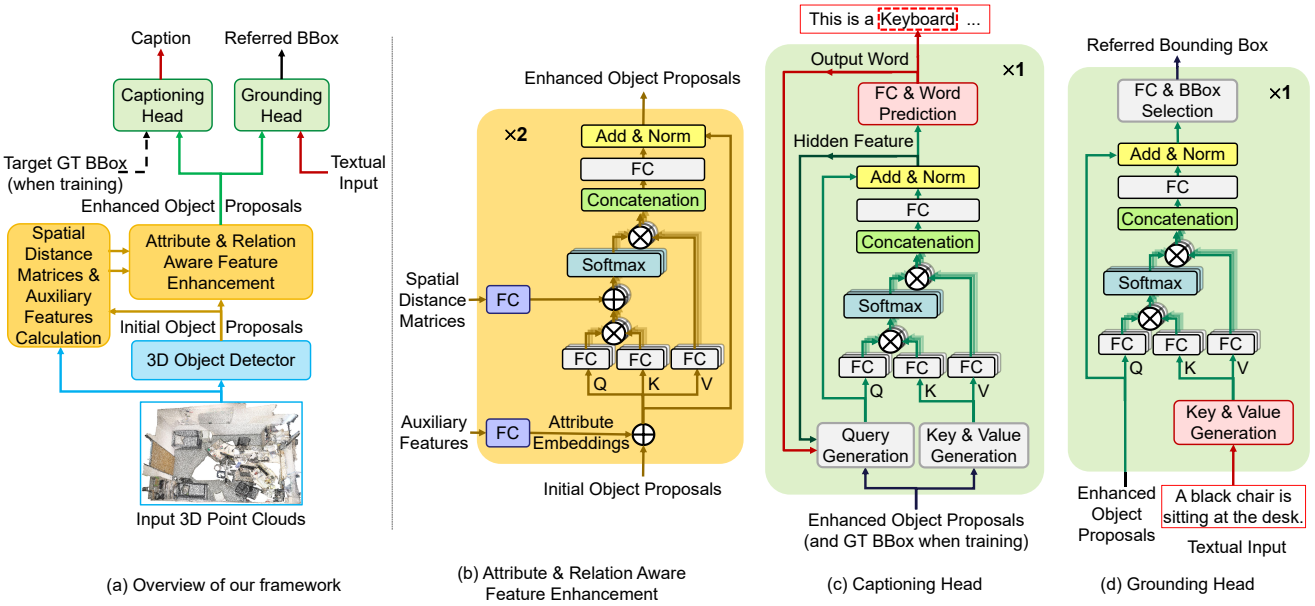

Figure 1. (a) The overview of our framework. (b) The attribute and relation aware feature enhancement module. (c) The captioning head within our framework (d) The grounding head within our framework. "FC" means the fully connected layer.

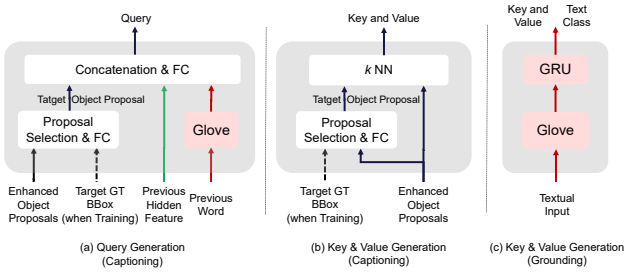

Figure 2. The Query, Key & Value generation processes for both captioning head and grounding head. For the captioning head, we firstly choose the object of interest to produce the target object proposal. We concatenate the target proposal feature, the tokenized word feature from the previous word and the hidden feature recurrently output by the multi-head cross-attention module, and use fully connected layers to generate the Query. We select K nearest neighbors of the target object proposal as the Key and Value. For the grounding head, the textual input is firstly tokenized and fed into a GRU cell to produce the the Key and Value of the multi-head cross-attention module. The Query for the grounding task is the enhanced object proposal features (see Fig. 1(d)).

3.3. Captioning Head

The 3D dense captioning task is to generate descriptions for each detected bounding box from the input point cloud, which is more object-oriented. Thus, the objectness (for accurately locating each object), the attribute information (for reasonably describing the attributes of objects), and the primary context (for further describing the key relations between each object with other objects) of all the objects in a scene are of great importance. Since the object detector and the feature enhancement module can provide rich object class information, attribute features, and global context features, we simply design our captioning head with a 1-layer multi-head cross-attention network structure for effective message passing between the enhanced features from the target object proposal and all other initial object proposals, which will focus more on the primary context features.

For generating the query (Q) input of the multi-head cross-attention module, we firstly select the target object proposal and then encode the corresponding object features with a fully connected layer. During the training stage, we select the object proposal with the highest IoU score with the ground-truth bounding box as the query object. In the testing stage, we use all object proposals in the scene (after the Non-Maximum Suppression (NMS) process) in a one-by-one fashion as the query object. For the target object proposal, we follow most of the captioning methods [9] to use a recurrent network structure to progressively generate each word of the caption. Then, we recurrently aggregate the hidden feature output by the multi-head cross-attention module and the tokenized word feature of the previous word (which is the ground-truth word in the training stage, and the newly predicted word in the testing stage) with the current query object features. The fused features form the final generated query input.

In the recursive query generation process, to alleviate the exposure bias [6] in the sequence generation task be-

tween the training stage (which uses the ground-truth word) and the testing stage (which uses the previously predicted word), we randomly use the autoregressive strategy during training. In details, we randomly replace 10% of the ground-truth word tokens with the predicted word tokens as the input word feature during the training process.

In the key (K) and value (V) generation module, we use the k-NN strategy to select the top k object proposals that are located closest to the target proposal based on their center distance in the 3D coordinate space, which filters out the less related objects in the scene. The selected object proposals are used as the key and value for the multi-head cross-attention module. In our experiment, k is empirically set as 20. This strategy is specially designed for the captioning task, because it mainly cares about the most obvious (or primary) relations between the target object and its surrounding objects and the rest of the relation information might be less important to the captioning task.

Finally, the multi-head cross-attention module is followed by a fully connected layer and a simple word prediction module to predict each word of the caption in a one-by-one fashion.

3.4. Grounding Head

For the 3D visual grounding task, the inputs include the 3D point clouds of a scene and the text-form language descriptions of one of the objects in the scene, and the task is to locate the object of interest based on the language description. Since the task-agnostic 3D object detector and the feature enhancement module already capture the object attributes and the complex relations among objects in a scene, the grounding head mainly focuses on matching between the given language descriptions and the detected object proposals. The grounding head in our method is more lightweight by simply using a 1-layer multi-head cross-attention module instead of multiple stacked cross-attention modules as used in [50] and [14].

The key (K) and value (V) inputs are generated based on the input language descriptions. Specifically, we use the similar language encoder as in ScanRefer [7]. The input language is firstly encoded by using a pretrained Glove [30] module, and then input to a GRU cell. The output word feature of the GRU cell forms the key (K) and value (V) inputs. Moreover, a global language feature is also generated from the GRU cell to predict the subject category of each sentence. The object proposals are used as the query (Q) input. By using the multi-head cross-attention mechanism between the language descriptions (K & V) and the object proposals (Q), the relationship between the sentence and the detected proposals is well captured.

To fully explore the contextual relations among the given textual description, we follow [50] to use two data augmentation strategies for both modalities (*e.g.*, randomly erase some words or change the order of the input text for the text input, and randomly copy some object proposals from other scene as the negative samples for enhancing object proposals), please refer to [50] for more details about the two data augmentation strategies.

Finally, a grounding classifier is used to generate the confidence score of each object proposal, and the proposal with the highest prediction score is considered as the final grounding result.

3.5. Training details

The loss function of our framework is a combination of the detection loss $L_{\text{detection}}$, the grounding loss $L_{\text{grounding}}$ and the captioning loss $L_{\text{captioning}}$.

The object detection loss is similar to that used in Qi *et al*. [31] for the ScanNet dataset [12], where $L_{\text{detection}} = 10L_{\text{vote-reg}} + L_{\text{objn-cls}} + L_{\text{sem-cls}} + 200L_{\text{boundary-reg}}$, except that we replace the bounding box classification loss $L_{\text{box-cls}}$ and the regression loss $L_{\text{box-reg}}$ in [7, 31] with the boundary regression loss $L_{\text{boundary-reg}}$ [36]. For the visual grounding task, we apply the similar loss function as used in ScanRefer [7], which is a combination of the localization loss L_{loc} for visual grounding and an auxiliary language-to-object classification loss L_{cls} to enhance the subject classification of the input sentence, and $L_{\text{grounding}} = L_{\text{loc}} + L_{\text{cls}}$. For the dense captioning task, we input the ground-truth words (or the predicted words with a probability of 10%) sequentially and $L_{\text{captioning}}$ is the average cross-entropy loss over all generated words. The final loss is a linear combination of these loss terms, *i.e.*, $L = L_{\text{detection}} + 0.3L_{\text{grounding}} + 0.2L_{\text{captioning}}$, where the trade-off parameters are empirically set for balancing different loss terms.

4. Experiments

4.1. Datasets and implementation details

Visual Grounding Dataset: We use the ScanRefer [7] dataset to evaluate our method for the visual grounding task. The ScanRefer dataset contains $51,583$ textual descriptions about $11,046$ objects from 800 scenes. The overall accuracy and the accuracies on both"unique" and "multiple" subsets are reported. We label each grounding data as "unique" if it only contains a single object from its class in the scene, otherwise it will be labeled as "multiple". For this dataset, we use Acc@0.25IoU and Acc@0.5IoU as our evaluation metrics. We also compare our method with the baseline methods on both the validation set and the online test set available at the ScanRefer's benchmark website[1].

Visual Captioning Datasets: Scan2Cap [9] is a dense captioning dataset for 3D scenes. The descriptions that are longer than 30 tokens in the ScanRefer dataset are truncated and two special tokens `[SOS]` and `[EOS]` are added to

[1] http://kaldir.vc.in.tum.de/scanrefer_benchmark

Table 1. Comparison of the visual grounding results from different methods on the ScanRefer [7] dataset. We report the percentage of the correctly predicted bounding boxes whose IoU scores with the ground-truth boxes are larger than 0.25 and 0.5, respectively. The results on both "unique" and "multiple" subsets are also reported. [*]: Note the InstanceRefer [47] method filters the predicted 3D proposals based on the object class prediction results such that this method only selects the target object proposal from the proposals in the same class, which simplifies the 3D visual grounding problem. This strategy is not adopted in our work.

			Unique		Multiple		Overall	
	Detector	Data	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Validation set								
ScanRefer [7]	VoteNet	3D Only	67.64	46.19	32.06	21.26	38.97	26.10
InstanceRefer [47]*	PointGroup	3D Only	77.13	**66.40**	28.83	22.92	38.20	31.35
Non-SAT [44]	VoteNet	3D Only	68.48	47.38	31.81	21.34	38.92	26.40
3DVG-Transformer [50]	VoteNet	3D Only	77.16	58.47	38.38	28.70	45.90	34.47
Ours	VoteNet	3D Only	**78.75**	61.30	**40.13**	**30.08**	**47.62**	**36.14**
ScanRefer [7]	VoteNet	2D + 3D	76.33	53.51	32.73	21.11	41.19	27.40
TGNN [16]	3D-UNet	2D + 3D	68.61	56.80	29.84	23.18	37.37	29.70
SAT [44]	VoteNet	2D + 3D	73.21	50.83	37.64	25.16	44.54	30.14
InstanceRefer [47]*	PointGroup	2D + 3D	75.72	**64.66**	29.41	22.99	38.40	31.08
3DVG-Transformer [50]	VoteNet	2D + 3D	81.93	60.64	39.30	28.42	47.57	34.67
Ours	VoteNet	2D + 3D	**83.47**	64.34	**41.39**	**30.82**	**49.56**	**37.33**
Online Benchmark								
ScanRefer [7]	VoteNet	2D + 3D	68.59	43.53	34.88	20.97	42.44	26.03
TGNN [16]	3D-UNet	2D + 3D	68.34	58.94	33.12	25.26	41.02	32.81
InstanceRefer [47]*	PointGroup	2D + 3D	**77.82**	**66.69**	34.57	26.88	44.27	35.80
3DVG-Transformer [50]	VoteNet	2D + 3D	75.76	55.15	42.24	29.33	49.76	35.12
Ours	VoteNet	2D + 3D	76.75	60.59	**43.89**	**31.17**	**51.26**	**37.76**

Table 2. Comparison of the 3D dense captioning results from different methods on the Scan2Cap [9] validation set. We average the scores from the conventional captioning metrics based on the predicted bounding boxes whose IoU scores with the ground-truth boxes are larger than 0.25 and 0.5, respectively.

	Detector	Data	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5
Scan2Cap [9]	VoteNet	3D Only	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57
Ours	VoteNet	3D Only	**60.86**	**39.67**	**27.45**	**59.02**	**47.68**	**31.53**	**24.28**	**51.08**
VoteNetRetr [31]	VoteNet	2D + 3D	15.12	18.09	19.93	38.99	10.18	13.38	17.14	33.22
Scan2Cap [9]	VoteNet	2D + 3D	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.48
Ours	VoteNet	2D + 3D	**64.70**	**40.17**	**27.66**	**59.23**	**49.48**	**31.03**	**24.22**	**50.80**

indicate the start and end of the description, and thus the textual descriptions for ScanRefer and Scan2Cap datasets are different.

As a sub-dataset of ReferIt3D [1], Nr3D is also built based on ScanNet with additional textual descriptions, and it contains 41, 503 samples collected by ReferItGame. We use the same metric as used for performance evaluation on the Scan2Cap dataset.

Specifically, the metric for performance evaluation on these two 3D captioning datasets combines the standard image captioning metrics under different IoU scores between the predicted bounding boxes and the target bounding boxes. The combined metric is defined as $m@kIoU = \frac{1}{P}\sum_{i=0}^{P} m_i u_i$, where $u_i \in \{0, 1\}$ is set to 1 if the detection IoU score for the i-th bounding box is greater than k, and 0 otherwise. We use m_i to represent the captioning metrics such as CiDEr [38], BLEU [28], METEOR [5] and ROUGE-L [19], which are respectively abbreviated as `C`, `B-4`, `M` and `R` in the following tables. P is the number of ground-truth or detected object bounding boxes.

Implementation Details. We follow [50] to use 8 sentences for each scene from both datasets when training our framework. Our experiment is carried out on the machine with a single NVIDIA 11GB 2080Ti GPU and it tasks 200 epochs to train our framework on both ScanRefer [7] and Scan2Cap [9] datasets with a batch size of 10 in each iteration (*i.e.*, there are 80 sentences from 10 point clouds). We apply the cosine learning rate decay strategy with the AdamW optimizer and a weight decay factor of 1e-5 to train our method. We empirically set the initial learning rate as 2e-3 for the detector, and 5e-4 for other modules of our framework (*i.e.*, the feature enhancement module and two task-specific heads). In addition, the captioning task with the cross-entropy loss is prone to overfitting, so we only add the captioning loss during the last 50 epochs.

4.2. Comparison with the state-of-the-art methods

Following the works ScanRefer [7] and Scan2Cap [9], we report the results under both "3D Only" and "2D + 3D" settings according to whether the auxiliary features are used. Under the "3D Only" setting, we use "xyz + RGB + normals" as the auxiliary features. Under the "2D + 3D" setting, the auxiliary features contain "xyz + multiviews + normals", where "multiviews" means multiview image features from a pretrained ENet [29], and "normals" means the normal vectors from point clouds.

In Table 1 and Table 2, we compare the dense captioning and visual grounding results of our framework with several state-of-the-art methods on both ScanRefer [7] and Scan2Cap [9] datasets. Specifically, on the ScanRefer dataset, we compare our method with the 3D instance segmentation-based methods TGNN [16] and InstanceRefer [47] and the 3D detection-based methods including ScanRefer [7] and 3DVG-Transformer [50]. On the Scan2Cap dataset, we compare our method with the state-of-the-art 3D detection-based method Scan2Cap [9] and VoteNetRetr [31].

From Table 1, we observe that our method outperforms the baseline methods for the visual grounding task. Note that we use a simpler network structure when compared with the state-of-the-art method 3DVG-Transformer [50], so the results validate that our joint learning framework can benefit the grounding task with only a lightweight grounding head. Specifically, in terms of Acc@0.25 and Acc@0.5 metrics, our method achieves around 1.9% and 2.6% improvements in the "overall" case when compared with 3DVG-Transformer [50] on the validation set under the "2D+3D" setting. When compared with other detection-based methods, our method achieves more improvement on the "Unique" subset, possibly because the attribute information of the objects plays a more important role in the "Unique" subset when there is no confusing objects from the same category in the scene. The results also verify that the object-oriented captioning task enhances the grounding performance by providing more attribute information. Note that the baseline methods InstanceRefer [47] and TGNN [16] use the extra instance segmentation masks for generating 3D proposals, while the InstanceRefer [47] method further filters the instances based on the semantic prediction results, namely, it only retains the instances from the same predicted class for generating the visual grounding results. Possibly due to these two aspects, the InstanceRefer [47] method achieve good results in the "Unique" subset. In contrast to [16, 47], our work only relies on the detection results, and it still outperforms both methods in both "Multiple" and "Overall" cases.

When compared with the baseline method "Scan2Cap", from the results in Table 2, we observe that our joint learning framework using a simple feature enhancement module and a lightweight captioning head achieves significant performance improvement for the captioning task. Under the "2D+3D" setting, our method achieves remarkable performance improvement of 10.4%, 7.71% and 6.32% in terms of C@0.5IoU, B-4@0.5IoU and R@0.5IoU, respectively. For this task, the improvement comes from both network structure design (*e.g.*, the attribute and relation aware feature enhancement module, and the lightweight captioning head) and the joint training strategy. The contribution of each module will be discussed in the ablation study below.

4.3. Ablation Study

Effectiveness of the feature enhancement module and the joint training strategy. To evaluate the effectiveness of the proposed task-agnostic feature enhancement module as well as the joint training strategy, we conduct the ablation study and report the corresponding results in Table 3. Without using the joint training strategy, the alternative method "w/o Grounding Head" (*resp.*, "w/o Captioning Head") means we train the two separate networks consisting of two task-agnostic modules and the captioning head (*resp.*, grounding head) for the 3D dense captioning task (*resp.*, the visual grounding task). "w/o Feature Enhancement" means we remove the "attribute & relation aware feature enhancement" module in our joint learning framework. For both dense captioning and the visual grounding tasks, our complete 3DJCG method based on the default training data (*i.e.*, from both Scan2Cap and ScanRefer datasets) outperforms those alternative methods, which indicate both strategies contribute to the final performance improvement to certain degree.

Does performance improvement come from more training data? Our joint training framework uses both the captioning and grounding training data, in which the only difference is the textual descriptions (*i.e.*, the descriptions used for the grounding task are relatively longer or with more complex relations, while the dense captions are shorter textual descriptions focusing more on the object class and the corresponding attributes). Hence, we conduct the experiments to verify whether the performance improvement is due to the utilization of more training data (*i.e.*, more textual descriptions from both tasks).

In Table 3 (a) and (b), 3DJCG ("Captioning Data Only" (*resp.*, 3DJCG ("Grounding Data Only")) indicates that we only use the 3D captioning dataset Scan2Cap [9] (*resp.*, the 3D visual grounding dataset ScanRefer [7]) when training our joint learning framework including both captioning and grounding heads and the two task-agnostic modules. Note both Scan2Cap and ScanRefer datasets can be readily used as the training data for these two tasks. By default, we use both datasets as the default training data when training our joint learning framework.

The results show that our 3DJCG framework using "Cap-

Table 3. Comparison of the visual grounding results under the "2D+3D" setting and the dense captioning results based on the correctly predicted bounding boxes whose IoU scores with the ground-truth boxes are larger than 0.5. In the "Network Modules" column, for better presentation, we label our detector, the feature enhancement module, the captioning head and the grounding head as "DE", "FE", "CH" and "GH", respectively.

(a) The 3D dense captioning results on the dataset Scan2Cap [9]

	Training Dataset(s)		Network Modules				Dense Captioning Results			
	Scan2Cap	ScanRefer	DE	FE	CH	GH	B-4@0.5	C@0.5	R@0.5	M@0.5
3DJCG (w/o Grounding Head) / 3DJCG-C	✓		✓	✓	✓		26.24	45.04	46.69	23.27
3DJCG (w/o Feature Enhancement)	✓	✓	✓		✓	✓	29.08	47.67	49.58	23.78
3DJCG (Captioning Data Only)	✓		✓	✓	✓	✓	30.40	47.29	50.29	23.91
3DJCG (Default Training Data)	✓	✓	✓	✓	✓	✓	**31.03**	**49.48**	**50.80**	**24.22**

(b) The 3D visual grounding results on the dataset ScanRefer [7]

	Training Dataset(s)		Network Modules				Visual Grounding Results		
	Scan2Cap	ScanRefer	DE	FE	CH	GH	Unique@0.5	Multiple@0.5	Overall@0.5
3DJCG (w/o Captioning Head) / 3DJCG-G		✓	✓	✓		✓	62.60	30.48	36.72
3DJCG (w/o Feature Enhancement)	✓	✓	✓		✓	✓	63.20	28.36	35.12
3DJCG (Grounding Data Only)		✓	✓	✓	✓	✓	**64.50**	30.29	36.93
3DJCG (Default Training Data)	✓	✓	✓	✓	✓	✓	64.34	**30.82**	**37.33**

Table 4. The dense captioning results of different methods and different training strategies on the Nr3D dataset from ReferIt3D [1].

	B-4@0.5	C@0.5	R@0.5	M@0.5
Scan2Cap [9]	17.24	27.47	49.06	21.80
3DJCG-C (From Scratch)	20.45	33.03	51.73	23.05
3DJCG-C* (Finetune)	22.82	38.06	52.99	23.77

tioning Data Only" (*resp.*, "Grounding Data Only") generally improves the performance for the captioning task (*resp.*, the grounding task) when compared to the alternative method 3DJCG ("w/o Grounding Head") (*resp.*, 3DJCG ("w/o Captioning Head")), especially for the dense captioning task. The results validate that the performance gains come from both strategies (*i.e.*, our network design and utilization of the additional training data). Moreover, the improved results from our joint learning framework under "Captioning Data Only" and "Grounding Data Only" settings also verify that our joint framework can also inspire the network design of each individual task.

Experiments on the Nr3D [1] dataset. We also take the dense captioning task on the Nr3D dataset as an example to evaluate our proposed framework when training from scratch or using the fine-tuning strategy. "3DJCG-C (From Scratch)" indicates that we train our 3DJCG-C network from scratch without using any pre-training strategies. "3DJCG-C* (Finetune)" indicates we fine-tune the pretrained model based on the Nr3D dataset. Note the pretrained model is learnt based on both ScanRefer and Scan2Cap datasets, and we also remove "Grounding Head" before performing the finetune process. We also list the results of the baseline method Scan2Cap trained from scratch based on the Nr3D dataset. As shown in Table 4, our method "3DJCG-C (From Scratch)" outperforms the baseline method "Scan2Cap [9]", which further verifies the effectiveness of our newly designed network structure. We also observe that our "3DJCG-C* (Finetune)" method further improves "3DJCG-C (From Scratch)", which demonstrates that the results of our framework could also be boosted by using the fine-tuning strategy.

5. Conclusion and Future Work

Observing the shared and complementary properties of two different but closely related tasks 3D dense captioning and 3D visual grounding, we propose a unified framework to jointly solve the two tasks in a synergistic manner. In our framework, the task-agnostic modules are responsible for the precise object localization, the enhancement of the geometry and the fine-grained attribute features, and fully exploration of the complex geometrical relations between objects in a 3D scene, while the task-specific lightweight captioning head and grounding head solve the two tasks, respectively. The experimental results validate the effectiveness of the proposed framework for both tasks. While the joint framework improves the performance of both tasks, the performance improvement for the visual grounding task is not as significant as that for the dense captioning task. In our future work, we will develop more advanced joint training framework to further improve the 3D visual grounding performance.

Acknowledgement This work was supported by the National Key Research and Development Project of China (No. 2018AAA0101900), and the National Natural Science Foundation of China (No.61906012, No.62006012, No.62132001).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pages 422–440, 2020. 1, 2, 6, 8

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 2

[3] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*, 2016. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433, 2015. 2

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72, 2005. 6

[6] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *NeurIPS*, 28, 2015. 4

[7] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *ECCV*, pages 202–221, 2020. 1, 2, 3, 5, 6, 7, 8

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV*, pages 104–120, 2020. 2

[9] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, pages 3193–3203, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[10] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3d object detection in point clouds. In *CVPR*, pages 8963–8972, 2021. 2

[11] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020. 2

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 5

[13] Jinyang Guo, Jiaheng Liu, and Dong Xu. JointPruning: Pruning networks along multiple dimensions for efficient point cloud processing. *IEEE TCSVT*, 2021. 2

[14] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *ACM MM*, pages 2344–2352, 2021. 2, 5

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016. 2

[16] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, pages 1610–1618, 2021. 2, 6, 7

[17] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016. 2

[18] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *CVPR*, pages 6271–6280, 2019. 2

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, pages 74–81, 2004. 6

[20] Guanze Liu, Yu Rong, and Lu Sheng. Votehmr: Occlusion-aware voting network for robust 3d human mesh recovery from partial point clouds. In *ACM MM*, pages 955–964, 2021. 2

[21] Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in RGBD images. In *CVPR*, pages 6032–6041, 2021. 1

[22] Jiaheng Liu and Dong Xu. GeometryMotion-Net: A strong two-stream baseline for 3d action recognition. *IEEE TCSVT*, pages 4711–4721, 2021. 2

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, volume 32, 2019. 2

[24] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10437–10446, 2020. 2

[25] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 2

[26] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 2

[27] Duy-Kien Nguyen and Takayuki Okatani. Multi-task learning of hierarchical vision-language representation. In *CVPR*, pages 10492–10501, 2019. 2

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 6

[29] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. In *arXiv preprint arXiv:1606.02147*, 2016. 7

[30] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5

[31] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 3, 5, 6, 7

[32] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 2

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, volume 30, 2017. 2

[34] Zizheng Que, Guo Lu, and Dong Xu. VoxelContext-Net: An octree based framework for point cloud compression. In *CVPR*, pages 6042–6051, 2021. 2

[35] Rui Su, Qian Yu, and Dong Xu. STVGBert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, pages 14618–14627, 2021. 2

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 3, 5

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 3

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 6

[39] Feiyu Wang, Wen Li, and Dong Xu. Cross-dataset point cloud recognition using deep-shallow domain adaptation network. *IEEE TIP*, pages 7364–7377, 2021. 2

[40] Kaisiyuan Wang, Lu Sheng, Shuhang Gu, and Dong Xu. Sequential point cloud upsampling by exploiting multi-scale temporal dependency. *IEEE TCSVT*, pages 4686–4696, 2021. 2

[41] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, pages 9062–9069, 2019. 2

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 2

[43] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florêncio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. TAP: text-aware pre-training for text-vqa and text-caption. In *CVPR*, pages 8751–8761, 2021. 2

[44] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. SAT: 2d semantics assisted training for 3d visual grounding. *ICCV*, pages 1856–1866, 2021. 1, 2, 6

[45] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular attention network for referring expression comprehension. In *CVPR*, pages 4555–4564, 2018. 2

[46] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, pages 7282–7290, 2017. 2

[47] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. *ICCV*, pages 1791–1800, 2021. 1, 2, 6, 7

[48] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588, 2021. 2

[49] Weichen Zhang, Wen Li, and Dong Xu. SRDAN: Scale-aware and range-aware domain adaptation network for cross-dataset 3d object detection. In *CVPR*, pages 6769–6779, 2021. 2

[50] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021. 1, 2, 3, 5, 6, 7

[51] Lichen Zhao, Jinyang Guo, Dong Xu, and Lu Sheng. Transformer3D-Det: Improving 3d object detection by vote refinement. *IEEE TCSVT*, pages 4735–4746, 2021. 2