D^{3}Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding

Dave Zhenyu Chen[1] Qirui Wu[2] Matthias Nießner[1] Angel X. Chang[2]
[1]Technical University of Munich [2]Simon Fraser University
https://daveredrum.github.io/D3Net/

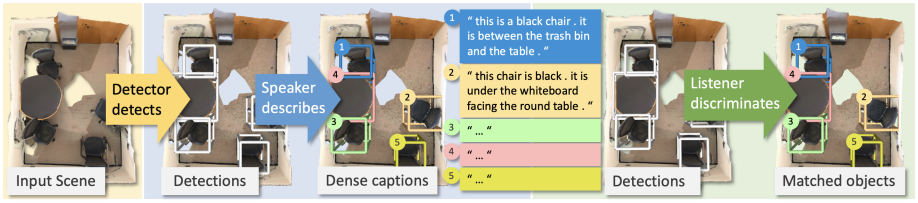

Fig. 1: We introduce D^{3}Net, an end-to-end neural speaker-listener architecture that can **d**etect, **d**escribe and **d**iscriminate. D^{3}Net also enables semi-supervised training on ScanNet data with partially annotated descriptions.

Abstract. Recent work on dense captioning and visual grounding in 3D have achieved impressive results. Despite developments in both areas, the limited amount of available 3D vision-language data causes overfitting issues for 3D visual grounding and 3D dense captioning methods. Also, how to discriminatively describe objects in complex 3D environments is not fully studied yet. To address these challenges, we present D^{3}Net, an end-to-end neural speaker-listener architecture that can **d**etect, **d**escribe and **d**iscriminate. Our D^{3}Net unifies dense captioning and visual grounding in 3D in a self-critical manner. This self-critical property of D^{3}Net encourages generation of discriminative object captions and enables semi-supervised training on scan data with partially annotated descriptions. Our method outperforms SOTA methods in both tasks on the ScanRefer dataset, surpassing the SOTA 3D dense captioning method by a significant margin.

1 Introduction

Recently, there has been increasing interest in bridging 3D visual scene understanding [41, 18, 19, 5, 11, 22, 46] and natural language processing [48, 13, 4, 34, 55]. The task of 3D visual grounding [6, 59, 60] localizes 3D objects described by natural language queries. 3D dense captioning proposed by Chen et al. [7] is

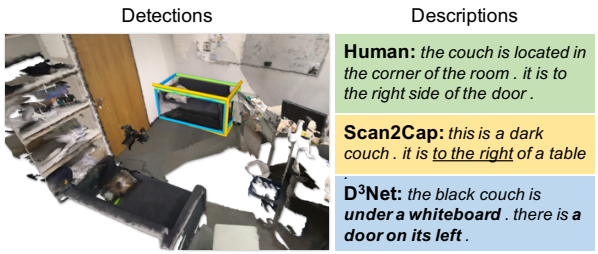

Fig. 2: Prior work [7] struggle to produce discriminative object captions. Also, captions often appear to be template-based. In contrast, our D^{3}Net generates discriminative object captions.

the reverse task where we generate descriptions for 3D objects in RGB-D scans. Both tasks enable applications such as assistive robots and natural language control in AR/VR systems.

However, existing work on 3D visual grounding [6, 1, 59, 23, 60] and dense captioning [7, 58] treats the two problems as separate, with *detect-then-discriminate* or *detect-then-describe* being the common strategies for tackling the two tasks. Separating the two complementary tasks hinders holistic 3D scene understanding where the ultimate goal is to create models that can infer: 1) what are the objects; 2) how to describe each object; 3) what object is being referred to through natural language. The disadvantages of having separated strategies are twofold. First, the detect-then-describe strategy often struggles to describe target objects in a discriminative way. In Fig. 2, the generated descriptions from Scan2Cap [7] fail to uniquely describe the target objects, especially in scenes with several similar objects. Second, existing 3D visual grounding methods [6, 60] in the detect-then-discriminate strategy suffer from severe overfitting issue, partly due to the small amount of 3D vision-language data [6, 1] which is limited compared to counterpart 2D datasets such as MSCOCO [32].

To address these issues, we propose an end-to-end self-critical solution, D^{3}Net, to enable discriminability in dense caption generation and utilize the generated captions improve localization. Relevant work in image captioning [36, 33] tackles similar issues where the generated captions are indiscriminative and repetitive by explicitly reinforcing discriminative caption generation with an image retrieval loss. Inspired by this scheme, we introduce a speaker-listener strategy, where the captioning module "speaks" about the 3D objects, while the localization module "listens" and finds the targets. Our proposed speaker-listener architecture can **d**etect, **d**escribe and **d**iscriminate, as illustrated in Fig. 1. The key idea is to reinforce the speaker to generate discriminative descriptions so that the listener can better localize the described targets given those descriptions.

This approach brings another benefit. Since the speaker-listener architecture self-critically generates and discriminates descriptions, we can train on scenes without any object descriptions. We see further improvements in 3D dense captioning and 3D visual grounding performance when using this additional data

alongside annotated scenes. This can allow for semi-supervised training on RGB-D scans beyond the ScanNet dataset. To summarize, our contributions are:

- We introduce a unified speaker-listener architecture to generate discriminative object descriptions in RGB-D scans. Our architecture allows for a semi-supervised training scheme that can alleviate data shortage in the 3D vision-language field.
- We study how the different components impact performance and find that having a strong detector is essential, and that by jointly optimizing the detector, speaker, and listener we can improve detection as well as 3D dense captioning and visual grounding.
- We show that our method outperforms the state-of-the-art for both 3D dense captioning and 3D visual grounding method by a significant margin.

2 Related Work

Vision and language in 3D. Recently, there has been growing interest in grounding language to 3D data [8, 2, 6, 1, 52, 44, 47]. Chen et al. [6] and Achlioptas et al. [1] introduce two complementary datasets consisting of descriptions of real-world 3D objects from ScanNet [11] reconstructions, named ScanRefer and ReferIt3D, respectively. ScanRefer proposes the joint task of detecting and localizing objects in a 3D scan based on a textual description, while ReferIt3D is focused on distinguishing 3D objects from the same semantic class given ground-truth bounding boxes. Yuan et al. [59] localize objects by decomposing input queries into fine-grained aspects, and use PointGroup [25] as their visual backbone. However, the frozen detection backbone is not fine-tuned together with the localization module. Zhao et al. [60] propose a transformer-based architecture with a VoteNet [41] backbone to handle multimodal contexts during localization. Despite the improved matching module, their work still suffers from poor quality detections due to the weak 3D detector. We show that fine-tuning an improved 3D detector is essential to getting good predictions and good localization performance. Chen et al. [7] introduce the task of densely detecting and captioning objects in RGB-D scans. Recently, Yuan et al. [58] aggregate the 2D features to point cloud to generate faithful object descriptions. Although their methods can effectively detect objects and generate captions w.r.t. their attributes, the quality of the bounding boxes and the discriminability of the captions are inadequate. Our method explicitly handles the discriminability of the generated captions through a self-critical speaker-listener architecture, resulting in the state-of-the-art performance in both 3D dense captioning and 3D visual grounding tasks.

Generating captions in images. Image captioning has attracted a great deal of interest [50, 53, 14, 28, 35, 3, 26, 43, 45]. Recent work [36, 33] suggest that traditional encoder-decoder-based image captioning methods suffer from the discriminability issues. Luo et al. [36] propose an additional image retrieval branch to reinforce discriminative caption generation. Liu et al. [33] propose a reinforcement learning method to train not only on annotated web images, but also

images without any paired captions. In contrast to generating captions for the entire image, in the dense captioning task we densely generate captions for each detected object in the input image [27, 54, 30]. Although such methods are effective for generating captions in 2D images, directly applying such training techniques on 3D dense captioning can lead to unsatisfactory results, since the captions involve 3D geometric relationships. In contrast, we work directly on 3D scene input dealing with object attributes as well as 3D spatial relationships.

Grounding referential expressions in images. There has been tremendous progress in the task of grounding referential expressions in images, also known as visual grounding [29, 40, 38, 21, 56, 20]. Given an image and a natural language text query as input, the target object is either localized by a bounding box [21, 56], or a segmentation mask [20]. These methods have achieved great success in the image domain. However, they are not designed to deal with 3D geometry inputs and handle complex 3D spatial relationships. Our proposed method directly decomposes the 3D input data with a sparse convolutional detection backbone, which produces accurate object proposals as well as semantically rich features.

Speaker-listener models for grounding. The speaker-listener model is a popular architecture for pragmatic language understanding, where a line of research explores how the context and communicative goals affect the linguistics [10, 16]. Recent work use neural speaker-listener architectures to tackle referring expression generation [38, 57, 37], vision-language navigation [15], and shape differentiation [2]. Mao et al. [38] construct a CNN-LSTM architecture optimized by a softmax loss to directly discriminate the generated referential expressions. There is no separate neural listener module compared with our method. Luo and Shakhnarovich [37] and Yu et al. [57] introduce a LSTM-based neural listener in the speaker-listener pipeline, but generating the referential expression is not directly supervised via the listener model, but rather trained via a proxy objective. In contrast, our method directly optimizes the Transformer-based neural listener for the visual grounding task by discriminating the generated object captions without any proxy training objective. Similarly, Achlioptas et al. [2] includes a pretrained and frozen listener in the training objective, while ours enables joint end-to-end optimization for both the speaker and listener via policy gradient algorithm. We experimentally show our method to be effective for semi-supervised learning in the two 3D vision-language tasks.

3 Method

D^{3}Net has three components: a 3D object detector, the speaker (captioning) module, and the listener (localization) module. Fig. 3 shows the overall architecture and training flow. The point clouds are fed into the detector to predict object proposals. The speaker takes object proposals as input to produce captions. To increase caption discriminability, we match these captions with object proposals via the listener. Caption quality is measured by the CIDEr [49] scores and the listener loss, which are back-propagated via REINFORCE [51] as re-

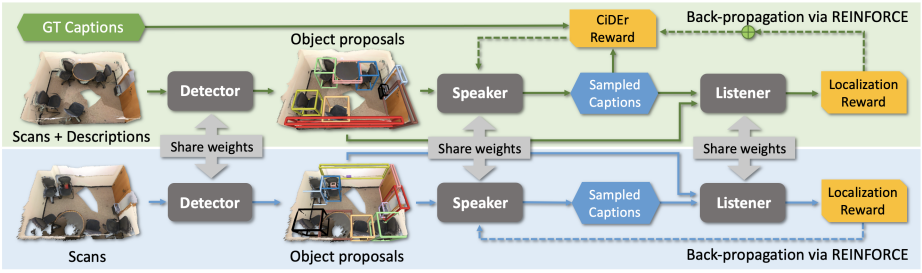

Fig. 3: D^{3}Net architecture. We input point clouds into the *detector* to predict object proposals. Then, those proposals are fed into the speaker to generate captions that *describes* each object. To *discriminate* the object described by each caption, the listener matches the generated captions with object proposals. The captioning and localization results are back-propagated via REINFORCE [51] as rewards through the dashed lines. D^{3}Net also enables end-to-end training on point clouds with no GT object descriptions (bottom blue block).

wards to the speaker. Our architecture can handle scenes without ground-truth (GT) object descriptions by reinforcing the speaker with the listener loss only.

3.1 Modules

Detector. We use PointGroup [25] as our detector module. PointGroup is a relatively simple model for 3D instance segmentation that achieves competitive performance on the ScanNet benchmark. We use ENet to augment the point clouds with multi-view features, following Dai and Nießner [12]. PointGroup uses a U-Net architecture with a SparseConvNet backbone to encode point features, cluster the points, and uses ScoreNet, another U-Net structure, to score each cluster. We take the cluster features after ScoreNet as the encoded object features. We refer readers to the original paper [25] for more details. The object bounding boxes are determined by taking the minimum and maximum points in the point clusters, and are produced as final outputs of our detector module.

Speaker. We base our speaker on the dense captioning method introduced by Chen et al. [7]. Our speaker module has two submodules: 1) a relational graph module, which is responsible for learning object-to-object spatial location relationships; 2) a context-aware attention captioning module, which attentively generates descriptive tokens with respect to the object attributes as well as the object-to-object spatial relationships.

Listener. For the listener, we follow the architecture introduced by Chen et al. [6] but replace the multi-modal fusion module with the transformer-based multi-modal fusion module of Zhao et al. [60]. Our listener module has two submodules: 1) a language encoding module with a GRU cell; 2) a transformer-based multi-modal fusion module similar to Zhao et al. [60], which attends to elements in the input query descriptions and the detected object proposals. As in Chen et al.

[6], we also incorporate a language object classifier to discriminate the semantics of the target objects in the input query descriptions.

3.2 Training Objective

The three modules are designed to be trained in an end-to-end fashion (see Figure 3). In this section, we describe the loss for each module, and how they are combined for the overall loss.

Detection loss. We use the instance segmentation loss introduced in PointGroup [25] to train the 3D backbone. The detection loss is composed of four parts: $L_{\text{det}} = L_{\text{sem}} + L_{\text{o\_reg}} + L_{\text{o\_dir}} + L_{\text{c\_score}}$. L_{sem} is a cross-entropy loss supervising semantic label prediction for each point. $L_{\text{o\_reg}}$ is a L_1 regression loss constraining the learned point offsets belonging to the same cluster. $L_{\text{o\_dir}}$ constrains the direction of predicted offset vectors, defined as the means of minus cosine similarities. It helps regress precise offsets, particularly for boundary points of large-size objects, since these points are relatively far from the instance centroids. $L_{\text{c\_score}}$ is another binary cross-entropy loss supervising the predicted objectness scores.

Listener loss. The listener loss is composed of a localization loss L_{loc} and a language-based object classification loss L_{lobjcls}. To obtain the localization loss L_{loc}, we first require a target bounding box. We use the detected bounding box with the highest IoU with the GT bounding box as the target bounding box. Then, a cross-entropy loss L_{loc} is applied to supervise the matching score prediction. In the end-to-end training scenario, the detected bounding boxes associated with the generated descriptions from the speaker are treated as the target bounding boxes. The language object classification loss is a cross-entropy loss L_{lobjcls} to supervise the classification based on the input description. The target classes are consistent with the ScanNet 18 classes, excluding structural objects such as "floor" and "wall".

Speaker loss using MLE training objective. The speaker loss is a standard captioning loss from maximum likelihood estimation (MLE). During training, provided with a pair of GT bounding box and the associated GT description, we optimize the description associated with the predicted bounding box which has the highest IoU score with the current GT bounding box. We first treat the description generation task as a sequence prediction task, factorized as: $L_{\text{spk-XE}}(\theta) = -\sum_{t=1}^{T} \log p(\hat{c_t}|\hat{c_1}, ..., \hat{c_{t-1}}; I, \theta)$, where $\hat{c_t}$ denotes the generated token at step t; I and θ represent the visual signal and model parameter, respectively. The token $\hat{c_t}$ is sampled from the probability distribution over the pre-defined vocabulary. The generation process is performed by greedy decoding or beam search in an autoregressive manner, and we use the argmax function to sample each token.

Joint loss using REINFORCE training objective. We use REINFORCE to train the detector-speaker-listener jointly. We first describe the enhanced speaker-loss, $L_{\text{spk-R}}$ that is trained using reinforcement learning to produce discriminative captions. We then describe the overall loss used in end-to-end training. Following prior work [36, 33, 42, 17, 57, 43], generating descriptions is

treated as a reinforcement learning task. In the setting of reinforcement learning, the speaker module is treated as the "agent", while the previously generated words and the input visual signal I are the "environment". At step t, generating word $\hat{c}_t$ by the speaker module is deemed as the "action" taken with the policy p_θ, which is defined by the speaker module parameters θ. Specifically, with the generated description $\hat{C} = \{c_1, ..., c_T\}$, the objective is to maximize the reward function $R(\hat{C}, I)$. We apply the "REINFORCE with baseline" algorithm following Rennie et al. [43] to reduce the variance of this loss function, where a baseline reward $R(C^*, I)$ of the description C^* independent of $\hat{C}$ is introduced. We apply beam search to sample descriptions and choose the greedily decoded descriptions as the baseline. The simplified policy gradient is:

$$L_{\text{spk-R}}(\theta) \approx -(R(\hat{C}, I) - R(C^*, I)) \sum_{t=1}^{T} \log p(\hat{c}_t | I, \theta) \tag{1}$$

Rewards. As the word-level sampling through the argmax function is non-differentiable, the subsequent listener loss cannot be directly back-propagated through the speaker module. A workaround is to use the gumbel softmax re-parametrization trick [24]. Following the training scheme of Liu et al. [33] and Luo et al. [36], the listener loss can be inserted into the REINFORCE reward function to increase the discriminability of generated referential descriptions. Specifically, given the localization loss L_{loc} and the language object classification loss L_{lobjcls}, the reward function $R(\hat{C})$ is the weighted sum of the CIDEr score of the sampled description and the listener-related losses:

$$R(\hat{C}, I) = R^{\text{CIDEr}}(\hat{C}, I) - \alpha[L_{\text{loc}}(\hat{C}) + \beta L_{\text{lobjcls}}(\hat{C})] \tag{2}$$

where α and β are the weights balancing the CIDEr reward and the listener rewards. We empirically set them to 0.1 and 1 in our experiments, respectively. To stabilize the training, the reward related to the baseline description $R(C^*)$ should be formulated analogously. Note that there should be no gradient calculation and back-propagation for the baseline C^*. For scenes with no GT descriptions provided, the CIDEr reward is cancelled in the reward function, which in this case becomes $R(\hat{C}, I) = -\alpha[L_{\text{loc}}(\hat{C}) + \beta L_{\text{lobjcls}}(\hat{C})]$.

Relative orientation loss. Following Chen et al. [7], we adopt the relative orientation loss on the message passing module as a proxy loss. The object-to-object relative orientations ranging from 0° to 180° are discretized into 6 classes. We apply a simple cross-entropy loss L_{ori} to supervise the relative orientation predictions.

Overall loss. We combine loss terms in our end-to-end joint training objective as: $L = L_{\text{det}} + L_{\text{spk-R}} + 0.3L_{\text{ori}}$.

3.3 Training

We use a stage-wise training strategy for stable training. We first pretrain the detector backbone on all training scans in ScanNet via the detector loss L_{det}.

We then train the dense captioning pipeline with the pretrained detector and a newly initialized speaker end-to-end via the detector loss and the speaker MLE loss $L_{\text{spk-XE}}$. After the speaker MLE loss converges, we train the visual grounding pipeline with the fine-tuned frozen detector and the listener via the listener loss L_{loc}. Finally, we fine-tune the entire speaker-listener architecture with the overall loss L.

3.4 Inference

During inference, we use the detector and the speaker to do 3D dense captioning and the listener to do visual grounding. The detector first produces object proposals, and the speaker generates a description for each object proposal. We take the minimum and maximum coordinates in the predicted object instance masks to construct the bounding boxes. For the object proposals that are assigned to the same ground truth, we keep only the one with the highest IoU with the GT bounding box. When evaluating the detector itself, the non-maximum suppression is applied.

4 Experiments

4.1 Dataset

We use the ScanRefer [6] dataset consisting of around 51k descriptions for over 11k objects in 800 ScanNet [11] scans. The descriptions include information about the appearance of the objects, as well as the object-to-object spatial relationships. We follow the official split from the ScanRefer benchmark for training and validation. We report our visual grounding results on the validation split and benchmark results on the hidden test set[1]. Our dense captioning results are on the validation split due to the lack of the test grounding truth. We also conduct experiments on the ReferIt3D dataset [1] (please see the supplemental).

4.2 Semi-supervised Training with Extra Data

As the scans in ScanRefer dataset are only a subset of scans in ScanNet, we extend the training set by including all re-scans of the same scenes for semi-supervised training. Unlike the scans in ScanRefer, these re-scans do not have per object descriptions. We can control how much extra data to use by randomly sampling (with replacement) from the set of re-scans. We experiment with augmenting our data with 0.1 to 1 times the amount of annotated data as extra data. During training, we randomly select detected objects in the sampled extra scans for subsequent dense captioning and visual grounding. For the complete 'extra' scenario, we use a comparable amount (1x) of extra data as the annotated data in ScanRefer.

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

Table 1: Quantitative results on 3D dense captioning and object detection. As in Chen et al. [7], we average the conventional captioning evaluation metrics with the percentage of the predicted bounding boxes whose IoU with the GTs are higher than 0.5. Our speaker model outperforms the baseline Scan2Cap without training via REINFORCE, while training with CIDEr reward further boosts the dense captioning performance. We also showcase the effectiveness of training with additional scans with no description annotations. Our speaker-listener architecture trained with 1x extra data achieves the best performance.

	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5
Scan2Cap [7]	39.08	23.32	21.97	44.78	32.21
X-Trans2Cap [58]	43.87	25.05	22.46	44.97	35.31
Ours (MLE)	46.07	30.29	24.35	51.67	50.93
Ours (CIDEr)	57.88	32.64	24.86	52.26	51.01
Ours (CIDEr+fixed loc.)	58.93	33.36	25.12	52.62	51.04
Ours (CIDEr+loc.)	61.30	34.50	25.25	52.80	52.07
Ours (CIDEr+loc.+lobjcls.)	61.50	35.05	25.48	53.31	52.58
Ours (w/ 0.1x extra data)	61.91	35.03	25.38	53.25	52.64
Ours (w/ 0.5x extra data)	62.36	35.54	25.43	53.67	53.17
Ours (w/ 1x extra data)	**62.64**	**35.68**	**25.72**	**53.90**	**53.95**

4.3 Implementation Details

We implement the PointGroup backbone using the Minkowski Engine [9] (see supplement). For the backbone, we train using Adam [31] with a learning rate of 2e-3, on the ScanNet train split with batch size 4 for 140k iterations, until convergence. For data augmentation, we follow Jiang et al. [25], randomly applying jitter, mirroring about the YZ-plane, and rotation about the Z axis (up-axis) to each point cloud scene. We then use the Adam optimizer with learning rate 1e-3 to train the detector and the listener on the ScanRefer dataset with batch size 4 for 60k iterations, until convergence. Each scan is paired with 8 descriptions (i.e. 4 scans and 32 descriptions per batch iteration). Then, we combine the trained detector with the newly initialized speaker on the ScanRefer dataset for the 3D dense captioning task, where the weights of the detector are frozen. We again use Adam with learning rate 1e-3, with the training process converging within 14k iterations. All our experiments are conducted on a RTX 3090, and all neural modules are implemented using PyTorch [39].

4.4 Quantitative Results

3D dense captioning and detection Tab. 1 compares our 3D dense captioning and object detection results against the baseline methods Scan2Cap [7] and X-Trans2Cap [58]. Leveraging the improved PointGroup based detector, our speaker model trained with the conventional MLE objective (Ours (MLE)) outperforms Scan2Cap and X-Trans2Cap by a large margin in all metrics. As expected, training with the CIDEr reward (Ours (CIDEr)) significantly improves the CIDEr score. We note that other captioning metrics are also improved, but

Table 2: Quantitative results on 3D visual grounding. We adapt the evaluation setting as in Chen et al. [6]. "Unique" means there is only one object belongs to a specific class in the scene, while "multiple" represents the cases where more than one object from a specific class can be found in the scene. Clearly, our base visual grounding network outperforms all baselines even before being put into the speaker-listener architecture. After the speaker-listener fine-tuning, our method achieves the state-of-the-art performance on the ScanRefer validation set and the public benchmark. Note that 3DVG-Trans+ is an unpublished extension of 3DVG-Trans [60] which appears only on the public benchmark.

	Val Acc@0.5IoU			Test Acc@0.5IoU		
	Unique	Multiple	Overall	Unique	Multiple	Overall
ScanRefer [6]	53.51	21.11	27.40	43.53	20.97	26.03
TGNN [23]	56.80	23.18	29.70	58.90	25.30	32.80
InstanceRefer [59]	66.83	24.77	32.93	66.69	26.88	35.80
3DVG-Trans [60]	60.64	28.42	34.67	55.15	29.33	35.12
3DVG-Trans+ [60]	-	-	-	57.87	**31.02**	37.04
Ours (w/o fine-tuning)	70.35	27.11	35.58	65.79	27.26	35.90
Ours	**72.04**	**30.05**	**37.87**	**68.43**	30.74	**39.19**

the detection mAP@0.5 remains similar. Training with object localization reward (Ours (CIDEr+loc.)) improves both captioning and detection further due to the improved discriminability during description generation. Note that if we use a frozen pretrained listener (Ours (CIDEr+fixed loc.)), the improvement is not as significant as when we allow the listener weights to be fine-tuned (Ours (CIDEr+loc.)). Our full model with the full listener reward incorporates an additional language object classification loss (Ours (CIDEr+loc.+lobjcls.)) and further improves the performance for both tasks.

Does additional data help? As our method allow for training the listener with scans without language data, we investigate the effectiveness of training with additional ScanNet data that have not been annotated with descriptions. We vary the amount of extra scan data (without descriptions) from 0.1x to 1x of fully annotated data and train our full model with CIDEr and full listener reward (loc.+lobjcls.). Our results (last three rows of Tab. 1), show that our semi-supervised training strategy can leverage the extra data to improve both dense captioning and object detection.

3D visual grounding Tab. 2 compares our results against prior 3D visual grounding methods ScanRefer [6], TGNN [23], InstanceRefer [59] and 3DVG-Transformer [60], and 3DVG-Trans+, an unpublished extension. Our method trained only with the detection loss and the listener loss ("Ours w/o fine-tuning"), i.e. without the speaker-listener setting, outperforms all the previous methods in the "Unique" and "Overall" scenarios. We find the improved fusion module together with the improved detector is sufficient to outperform 3DVG-Trans. Due to the improved detector, our method can distinguish objects in the "Unique" case, where the semantic labels play an important role. Meanwhile, 3DVG-Trans [60] still outperforms our base listener when discriminating objects

Fig. 4: Qualitative results in 3D dense captioning task from Scan2Cap [7] and our method. We underline the inaccurate words and mark the spatially discriminative phrases in bold. Our method qualitatively outperforms Scan2Cap in producing better object bounding boxes and more discriminative descriptions.

from the same class ("Multiple" case). Our end-to-end speaker-listener (last row) outperforms all previous method including 3DVG-Trans.

4.5 Qualitative Analysis

3D dense captioning Fig. 4 compares our results with object captions from Scan2Cap [7]. Descriptions generated by Scan2Cap cannot uniquely identify the target object in the input scenes (see the yellow block on the bottom right). Also, Scan2Cap produces inaccurate object bounding boxes, which affects the quality of object captions (see the yellow block on the top left). Compared to captions from Scan2Cap, our method produces more discriminative object captions that specifies more spatial relations (see bolded phrases in the blue blocks).
3D visual grounding Fig. 5 compares our results with 3DVG-Transformer [60]. Though 3DVG-Transformer is able to pick the correct object, it suffers from poor object detections and is constrained by the performance of the VoteNet-based detection backbone (see the first column). Our method is capable of selecting the queried objects while also predicting more accurate object bounding boxes.

4.6 Analysis and Ablation Studies

Does better detection backbone help? From Tab. 1, we see that using a better detector can significant improve performance. We further examine the ef-

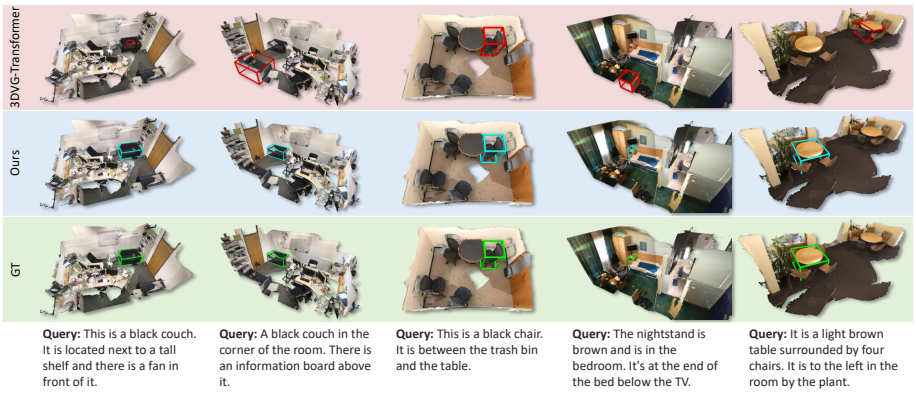

Fig. 5: 3D visual grounding results using 3DVG-Transformer [60] and our method. 3DVG-Transformer fails to accurately predict object bounding boxes, while our method produces accurate bounding boxes and correctly distinguishes target objects from distractors.

fect of using different detection backbones (VoteNet and PointGroup) compared to GT bounding boxes in Tab. 3. For each detection backbone, we use four variants of our method: the models trained without the joint speaker-listener architecture, and the speaker-listener architecture trained with CIDEr reward, listener reward and extra ScanNet data. The results with GT boxes show the effectiveness of our speaker-listener architecture, when detections are perfect. The large improvement from VoteNet [41] to PointGroup [25] show the importance of a better detection backbone. The gap between GT and VoteNet/PointGroup shows there is room for further improvement.

Are the generated descriptions more discriminative? To check whether the speaker-listener architecture generates more discriminative descriptions, we conduct an automatic evaluation via a reverse task. In this task, we feed the generated descriptions and GT bounding boxes into a pretrained neural listener model similar to Zhao et al. [60]. The predicted visual grounding results are evaluated in the same way as in our 3D visual grounding experiments. Higher grounding accuracy indicates better discrimination, especially in the "Multiple" case. Results (Tab. 4) show that our speaker-listener architecture generates more discriminative descriptions compared to Scan2Cap [7]. The discrimination is further improved when training with extra ScanNet data. To disentangle the affect of imperfectly predicted bounding boxes, we also train and evaluate our method with GT boxes (see last two rows in Tab. 4). We see that our semi-supervised speaker-listener architecture generates more discriminative descriptions.

Does the listener help with captioning? The third to the sixth rows in Tab. 1 measure the benefit of training the speaker together with the listener (Ours (CIDEr+loc.) and Ours (CIDEr+loc.+lobjcls.)) rather than training the speaker alone (Ours (CIDEr)). Training with the listener improves all captioning metrics. Also, training jointly with an unfrozen listener (Ours (CIDEr+loc.)

Table 3: Quantitative results on object detection, dense captioning and visual grounding in RGB-D scans. We train our method using different detection backbones as well as the ground truth bounding boxes. Our method trained with CIDEr and listener reward as well as the additional data outperforms the pretrained speaker and listener models.

Method	Detection	mAP@0.5	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	Unique Acc@0.5IoU	Multiple Acc@0.5IoU	Overall Acc@0.5IoU
Ours (MLE)	GT	100.00	71.41	42.95	29.67	64.93	88.45	36.46	46.03
Ours (CIDEr)	GT	100.00	94.80	47.92	30.80	**66.34**	-	-	-
Ours (CIDEr+lis.)	GT	100.00	95.62	47.65	**30.93**	66.31	89.76	36.85	47.14
Ours (CIDEr+lis.+extra)	GT	100.00	**96.31**	**48.20**	30.80	66.10	**89.86**	**40.66**	**48.17**
Ours (MLE)	VoteNet	32.21	39.08	23.32	21.97	**44.78**	56.41	21.11	27.95
Ours (CIDEr)	VoteNet	37.66	46.88	25.96	22.10	44.69	-	-	-
Ours (CIDEr+lis.)	VoteNet	38.03	47.32	24.76	21.66	43.62	57.90	20.73	28.03
Ours (CIDEr+lis.+extra)	VoteNet	**38.82**	**48.38**	**26.09**	**22.15**	44.74	**58.40**	**21.66**	**29.25**
Ours (MLE)	PointGroup	47.19	46.07	30.29	24.35	51.67	70.35	27.11	35.58
Ours (CIDEr)	PointGroup	52.44	57.88	32.64	24.86	52.26	-	-	-
Ours (CIDEr+lis.)	PointGroup	52.58	61.50	35.05	25.48	53.31	71.04	27.40	35.62
Ours (CIDEr+lis.+extra)	PointGroup	**53.95**	**62.64**	**35.68**	**25.72**	**53.90**	**72.04**	**30.05**	**37.87**

Table 4: We automatically evaluate the discriminability of the generated object descriptions. A pretrained neural listener similar to Zhao et al. [60] is fed with the GT object features and the descriptions generated by Scan2Cap [7] as well as our method. Higher grounding accuracy indicates better discriminability, especially in the "multiple" case. To alleviate noisy detections, the evaluation results on the descriptions generated from the GT object features are also presented. Our method generates more discriminative descriptions compared to Scan2Cap.

	detection	Unique Acc@0.5IoU	Multiple Acc@0.5IoU	Overall Acc@0.5IoU
Scan2Cap [7]	VN [41]	80.52	29.95	39.08
Ours (w/ CIDEr & lis.)	PG [25]	81.16	30.22	41.62
Ours (w/ CIDEr & lis. & extra)	PG [25]	**81.27**	**30.33**	**41.73**
Ours (w/ CIDEr & lis.)	GT	89.76	38.53	48.07
Ours (w/ CIDEr & lis. & extra)	GT	**90.29**	**40.66**	**49.71**

leads to a better performance when compared with the variant with a pretrained and frozen listener (Ours (CIDEr+fixed loc.), which is similar to Achlioptas et al. [2]. Additionally, as the detector is not only fine-tuned with the speaker but also with the listener, the additional supervision from the listener helps with the detection performance as well.

To analyze the quality of the generated object captions, we asked 5 students to perform a fine-grained manual analysis of the captions. Each student was presented with a batch of 100 randomly selected object captions with associated objects highlighted in the 3D scene. The student are then asked to indicate if the respective aspects were included and correctly described. The manual analysis results in Tab. 5 shows that our method generates more accurate descriptions compared to Scan2Cap. In particular, training with the listener and extra ScanNet data produces more accurate spatial relations in the descriptions. The results

Table 5: Manual analysis of captions generated by Scan2Cap [7] and variants of our method. We measure accuracy in three different aspects: object categories, appearance attributes and spatial relations. Our method generates more accurate descriptions in all aspects, especially for describing spatial relations.

	Acc (Category)	Acc (Attribute)	Acc (Relation)
Scan2Cap [7]	84.10	64.21	69.00
Ours (MLE)	88.00 (+3.84)	74.73 (+10.53)	69.00 (+0.00)
Ours (CIDEr)	88.89 (+4.73)	75.00 (+10.79)	68.00 (-1.00)
Ours (CIDEr+lis.)	90.91 (+6.75)	77.38 (+13.17)	75.00 (+6.00)
Ours (CIDEr+lis.+extra)	92.93 **(+8.77)**	80.95 **(+16.74)**	78.57 **(+9.57)**

of fine-grained manual analysis complements the automatic captioning evaluation metric. While metrics such as CIDEr captures the overall similarity of the generated sentences against the references, the accuracies in Tab. 5 measures the correctness of the decomposed visual attributes.

Does the speaker help with grounding? Tab. 2 compares grounding results between a pretrained listener (Ours w/o fine-tuning) and a fine-tuned speaker-listener model (Ours). Although the grounding performance drops in the "Unique" subset, the improvements in "Multiple" suggests better discriminability in tougher and ambiguous scenarios.

5 Conclusion

We present D^{3}Net, an end-to-end speaker-listener architecture that can **d**etect, **d**escribe and **d**iscriminate. Specifically, the speaker iteratively generates descriptive tokens given the object proposals detected by the detector, while the listener discriminates the object proposals in the scene with the generated captions. The self-discriminative property of D^{3}Net also enables semi-supervised training on ScanNet data without the annotated descriptions. Our method outperforms the previous SOTA methods in both tasks on ScanRefer, surpassing the previous SOTA 3D dense captioning method by a significant margin. Our architecture can serve as an initial step towards leveraging unannotated 3D data for language and 3D vision. Overall, we hope that our work will encourage more future research in 3D vision and language.

Acknowledgements

This work is funded by Google (AugmentedPerception), the ERC Starting Grant Scan2CAD (804724), and a Google Faculty Award. We would also like to thank the support of the TUM-IAS Rudolf Mößbauer and Hans Fischer Fellowships (Focus Group Visual Computing), as well as the the German Research Foundation (DFG) under the Grant *Making Machine Learning on Static and Dynamic 3D Data Practical*. This work is also supported in part by the Canada CIFAR AI Chair program and an NSERC Discovery Grant.

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, 3, 8
2. Achlioptas, P., Fan, J., Hawkins, R., Goodman, N., Guibas, L.J.: ShapeGlot: Learning language for shape differentiation. In: Proceedings of the IEEE international conference on computer vision (2019) 3, 4, 13
3. 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
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020) 1
5. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: Proceedings of the International Conference on 3D Vision, pp. 667–676, IEEE (2017) 1
6. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pp. 202–221, Springer (2020) 1, 2, 3, 5, 6, 8, 10
7. Chen, D.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) 1, 2, 3, 5, 7, 9, 11, 12, 13, 14
8. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In: Asian Conference on Computer Vision, pp. 100–116, Springer (2018) 3
9. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3075–3084 (2019) 9
10. Cole, P., Morgan, J.L.: Syntax and semantics. volume 3: Speech acts. Tijdschrift Voor Filosofie **39**(3) (1977) 4
11. 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) 1, 3, 8
12. Dai, A., Nießner, M.: 3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 452–468 (2018) 5
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 1
14. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634 (2015) 3

15. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: Advances in Neural Information Processing Systems (2018) 4
16. Golland, D., Liang, P., Klein, D.: A game-theoretic approach to generating spatial descriptions. In: Proceedings of the 2010 conference on empirical methods in natural language processing, pp. 410–419 (2010) 4
17. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European conference on computer vision, pp. 3–19, Springer (2016) 6
18. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4421–4430 (2019) 1
19. Hou, J., Dai, A., Nießner, M.: RevealNet: Seeing behind objects in RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2098–2107 (2020) 1
20. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: European Conference on Computer Vision, pp. 108–124, Springer (2016) 4
21. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4555–4564 (2016) 4
22. Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K.: Scenenn: A scene meshes dataset with annotations. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 92–101, IEEE (2016) 1
23. 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 (2021) 2, 10
24. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016) 7
25. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: PointGroup: Dual-set point grouping for 3D instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4867–4876 (2020) 3, 5, 6, 9, 12, 13
26. Jiang, W., Ma, L., Jiang, Y.G., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 499–515 (2018) 3
27. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4565–4574 (2016) 4
28. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128–3137 (2015) 3
29. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 787–798 (2014) 4
30. 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) 4

31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, pp. 740–755, Springer (2014) 2
33. Liu, X., Li, H., Shao, J., Chen, D., Wang, X.: Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 338–354 (2018) 2, 3, 6, 7
34. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019) 1
35. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 375–383 (2017) 3
36. Luo, R., Price, B., Cohen, S., Shakhnarovich, G.: Discriminability objective for training descriptive captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6964–6974 (2018) 2, 3, 6, 7
37. Luo, R., Shakhnarovich, G.: Comprehension-guided referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7102–7111 (2017) 4
38. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 11–20 (2016) 4
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035, Curran Associates, Inc. (2019) 9
40. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp. 2641–2649 (2015) 4
41. 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) 1, 3, 12, 13
42. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732 (2015) 6
43. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7008–7024 (2017) 3, 6, 7
44. Roh, J., Desingh, K., Farhadi, A., Fox, D.: LanguageRefer: Spatial-language model for 3D visual grounding. In: Proceedings of the Conference on Robot Learning (2021) 3
45. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European Conference on Computer Vision, pp. 742–758, Springer (2020) 3
46. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 567–576 (2015) 1

47. Thomason, J., Shridhar, M., Bisk, Y., Paxton, C., Zettlemoyer, L.: Language grounding with 3D objects. In: Proceedings of the Conference on Robot Learning (2021) 3
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, pp. 5998–6008 (2017) 1
49. 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) 4
50. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164 (2015) 3
51. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**(3), 229–256 (1992) 4, 5
52. Wu, X., Averbuch-Elor, H., Sun, J., Snavely, N.: Towers of babel: Combining images, language, and 3D geometry for learning multimodal vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 3
53. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048–2057 (2015) 3
54. 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) 4
55. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019) 1
56. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: MattNet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1307–1315 (2018) 4
57. Yu, L., Tan, H., Bansal, M., Berg, T.L.: A joint speaker-listener-reinforcer model for referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7282–7290 (2017) 4, 6
58. Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Li, Z., Cui, S.: X-Trans2Cap: Cross-modal knowledge transfer using transformer for 3D dense captioning (2022), arXiv:2203.00843 2, 3, 9
59. 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) 1, 2, 3, 10
60. 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) 1, 2, 3, 5, 10, 11, 12, 13