D^{3}Net: A Speaker-Listener Architecture for Semi-supervised Dense Captioning and Visual Grounding in RGB-D Scans

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/

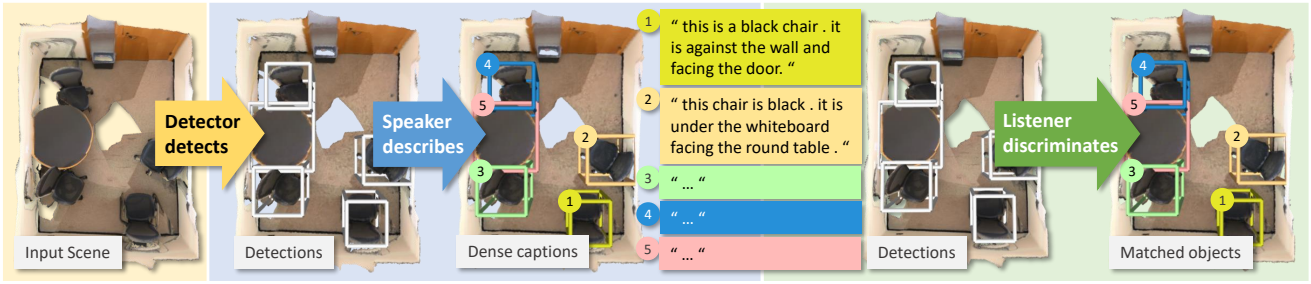


Figure 1. We introduce D^{3}Net, an end-to-end neural speaker-listener architecture that can **d**etect, **d**escribe and **d**iscriminate. D^{3}Net also enables semi-supervised training on Scan-Net data with partially annotated descriptions.

Abstract

*Recent studies 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 also introduces discriminability during object caption generation and enables semi-supervised training on ScanNet 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 (**23.56% CiDEr@0.5IoU** improvement).*

1. Introduction

Recently, there has been increasing interest in bridging 3D visual scene understanding [5, 13, 19, 20, 24, 44, 49] and natural language processing [4, 14, 36, 52, 59]. The task of 3D visual grounding [6, 62, 64] localizes 3D objects described by natural language queries. 3D dense captioning proposed by Chen et al. [7] is 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.

To enable such research, Chen et al. [7] and Achlioptas et al. [2] contributed datasets consisting of around 51k and 41k descriptions for objects in ScanNet [13] scenes, respectively. However, these datasets are limited in scale compared to 2D image datasets. For instance, MSCOCO [34] consists of more than 400k free-form descriptions for around 100k images. The limited amount of data results in overfitting in existing 3D visual grounding methods [6, 64]. It can also hinder the application of more complex vision-language architectures such as UniT [21] and UniLM [65] on 3D data. Moreover, qualitative results from prior work on 3D dense captioning [7] indicate that the captioning model struggles to describe target objects in a discriminative way (see Fig. 2) with generated sentences often appearing to follow pre-defined templates and lacking in diversity.

To address these issues, we propose an end-to-end self-critical solution, D^{3}Net, to enable discriminability in dense caption generation. Relevant work in image captioning [35, 39] 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

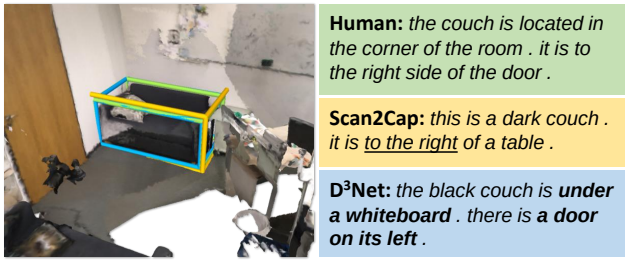


Figure 2. Prior work [7] struggle to produce discriminative object captions. Also, captions often appear to be template-based. In contrast, our D^{3}Net generates discriminative object captions.

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 allows for semi-supervised training on RGB-D scans beyond the ScanNet dataset. To summarize, our contributions are:

- We introduce a novel speaker-listener architecture to generate discriminative object descriptions in RGB-D scans.
- We propose a semi-supervised training scheme to alleviate data shortage in the 3D vision-language field.
- We show that our method simultaneously outperforms the state-of-the-art 3D dense captioning and 3D visual grounding method by **23.56% CiDEr@0.5IoU** and **3.20% Acc@0.5IoU**, respectively.

2. Related Work

Vision and language in 3D. Recently, there has been growing interest in grounding language to 3D data [1, 2, 6, 8, 47, 50, 56]. Chen et al. [6] and Achlioptas et al. [2] introduce two complementary datasets consisting of descriptions for real-world 3D objects from ScanNet [13] 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. [62] localize objects by decomposing input queries into fine-grained aspects, and used PointGroup [27] as their visual backbone. However, they used pre-computed instance predictions, so the detection backbone is not fine-tuned together with the localization module. Zhao et al. [64] propose a transformer-based architecture with a VoteNet [44] 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 object in RGB-D scans. Although their method can effectively generate object captions w.r.t. their attributes, the discriminability of the generated captions is 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 [3, 15, 28, 30, 37, 46, 48, 54, 57]. Recent work [35, 39] suggest that traditional encoder-decoder-based image captioning methods suffer from the discriminability issues. Luo et al. [39] propose an additional image retrieval branch to reinforce discriminative caption generation. Liu et al. [35] 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 [29, 32, 58]. 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 [22, 23, 31, 40, 43, 61]. Given an image and a natural language text query as input, the target object is either localized by a bounding box [23, 61], or a segmentation mask [22]. 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 [11]. Prior to the era of deep learning, seminal works study the communicative approach for generating object descriptions [17]. More recent work use neural speaker-

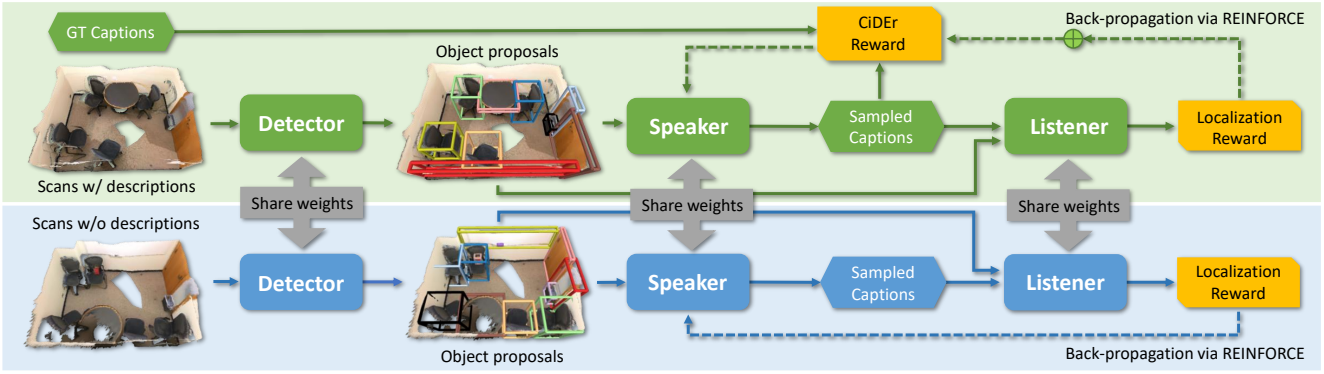


Figure 3. D^{3}Net architecture. We input point clouds into the detector to predict object proposals. Then, those proposals are fed into the speaker to generate object captions. To discriminate those descriptions, the listener matches the generated captions with object proposals. The captioning and localization results are back-propagated via REINFORCE [55] as rewards through the dashed lines. D^{3}Net also enables end-to-end training on point clouds with no GT object descriptions, as shown in the blue block.

listener architectures to tackle referring expression generation [38, 40, 60], vision-language navigation [16], and shape differentiation [1]. Mao et al. [40] construct a CNN-LSTM architecture optimized by a softmax loss to directly discriminate the generated referential expressions. In their work, the listener model uses a uniform prior over the candidate objects combined a model of the speaker to evaluate how likely the speaker is to be describing each candidate. There is hence no separate neural module that functions as a listener. Luo and Shakhnarovich [38] and Yu et al. [60] 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. Our architecture equipped with a Transformer-based neural listener directly reinforces the generation process via discriminating the generated object captions without any proxy training objective.

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 [53] scores and the listener loss, which are back-propagated via REINFORCE [55] as rewards to the speaker. Our architecture can handle scenes without ground-truth (GT) object descriptions by reinforcing the speaker through the listener loss only.

3.1. Modules

Detector. We use PointGroup [27] 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, then clusters 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 the readers to the original paper [27] for more details. The object bounding boxes are determined by taking the minimum and maximum points in the point clusters, and are produced as the 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. [64]. 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. [64], 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 [27] 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 predefined 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 [18, 35, 39, 45, 46, 60], 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. [46] 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:

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

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 the gumbel softmax reparametrization trick [26]. Following the training scheme of Liu et al. [35] and Luo et al. [39], 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:

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

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

Since it can be challenging to train the entire network end-to-end from scratch, we use a stage-wise training strategy. We first pretrain the detector backbone on all training scans in ScanNet via the detector loss. We then train the dense captioning pipeline with the pretrained detector and a newly initialized speaker in an end-to-end manner via the detector loss and the speaker MLE loss. 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. Finally, we fine-tune the whole speaker-listener architecture with the previously stated overall loss.

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. To speed up the evaluation process for visual grounding and captioning, we do not apply non-maximum suppression. Instead, 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. Unlike Chen et al. [6], we remove the non-maximum suppression module, as the network has already learned the best matching bounding box despite possible duplication in object proposals. When evaluating the detector itself, the non-maximum suppression is still applied.

4. Experiments

4.1. Dataset

We use the ScanRefer [6] dataset consisting of around 51k descriptions for over 11k objects in 800 ScanNet [13] 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 [2] (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

	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
Ours (MLE)	46.07	30.29	24.35	51.67	50.93
Ours (CiDEr)	60.93	34.36	25.12	52.26	51.44
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**

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.

4.3. Implementation Details

We implement the PointGroup backbone using the Minkowski Engine [10] (see supplement). For the backbone, we train using Adam [33] 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. [27], 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 3090Ti, and all neural modules are implemented using PyTorch [42].

4.4. Quantitative Results

3D dense captioning Tab. 1 compares our 3D dense captioning and object detection results against the baseline method Scan2Cap [7]. Leveraging the improved detector,

	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 [25]	56.80	23.18	29.70	58.90	25.30	32.80
InstanceRefer [62]	66.83	24.77	32.93	66.69	26.88	35.80
3DVG-Trans [64]	60.64	28.42	34.67	55.15	29.33	35.12
3DVG-Trans+ [64]	-	-	-	57.87	**31.02**	37.04
Ours (w/o fine-tuning)	72.04	27.11	35.58	65.79	27.26	35.90
Ours	**70.35**	**30.05**	**37.87**	**68.43**	30.74	**39.19**

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 [64] which appears only on the public benchmark.

	detection	Unique Acc@0.5IoU	Multiple Acc@0.5IoU	Overall Acc@0.5IoU
Scan2Cap [7]	VN [44]	80.52	29.95	39.08
Ours (w/ CiDEr & lis.)	PG [27]	81.16	30.22	41.62
Ours (w/ CiDEr & lis. & extra)	PG [27]	**81.27**	**30.33**	**41.73**
Ours (w/ CiDEr & lis.)	GT	**90.29**	38.53	48.07
Ours (w/ CiDEr & lis. & extra)	GT	89.76	**40.66**	**49.71**

Table 3. We automatically evaluate the discriminability of the generated object descriptions. A pretrained neural listener similar to Zhao et al. [64] 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.

our speaker model trained with the conventional MLE objective (marked "Ours (MLE)") outperforms Scan2Cap by a significant margin in all metrics. Our results are further improved via training with the CiDEr reward (marked "Ours (CiDEr)") while having the similar detection mAP@0.5. Training with object localization reward (marked "Ours (CiDEr+loc.)") improves the dense captioning results due to the reinforced discriminability during description generation. We also show that utilizing language object classification loss as an additional reward (marked "Ours (CiDEr+loc.+lobjcls.)") is effective. To showcase the effectiveness of training with listener as well as additional ScanNet data without descriptions, we present our results in the last three rows, where all models are trained with CiDEr and listener reward. By using additional ScanNet data during training. i.e. from 0.1x to 1x, all results in terms of dense captioning and object detection increase. This trend indicates the benefit of the semi-supervised training setup.

3D visual grounding Tab. 2 compares our results against prior 3D visual grounding methods ScanRefer [6], TGNN [25], InstanceRefer [62] and 3DVG-Transformer [64], and 3DVG-Trans+, an unpublished extension. Our method trained only with the detection loss and the listener loss (marked "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 [64] still outperforms our base listener when discriminating objects from the same class ("Multiple" case). Our end-to-end speaker-listener (last row) outperforms all previous method including 3DVG-Trans, despite a small performance drop in the "Unique" subset indicating potential overfitting in the detection backbone.

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). Our method produces more discriminative object captions in comparison with the ambiguous ones from Scan2Cap, using more relative spatial relationships (see the phrases in bold in the blue blocks).

3D visual grounding Fig. 5 compares our results with 3DVG-Transformer [64]. Though 3DVG-Transformer is able to pick the correct object, it still 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

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. [64]. 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. Tab. 3 shows that our speaker-listener architecture generates more discriminative descriptions compared to Scan2Cap [7]. The discrimination is

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	**89.86**	36.85	47.14
Ours (CiDEr+lis.+extra)	GT	100.00	**96.31**	**48.20**	30.80	66.10	89.76	**40.66**	**48.17**
Ours (MLE)	VoteNet	36.13	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	**58.40**	20.73	28.03
Ours (CiDEr+lis.+extra)	VoteNet	**38.82**	**48.38**	**26.09**	**22.15**	44.74	57.90	**21.66**	**29.25**
Ours (MLE)	PointGroup	47.19	46.07	30.29	24.35	51.67	**72.04**	27.11	35.58
Ours (CiDEr)	PointGroup	52.44	61.30	34.36	25.12	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**	70.35	**30.05**	**37.87**

Table 4. 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.

Figure 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.

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. 3). We see that our semi-supervised speaker-listener architecture generates more discriminative descriptions.

Does the listener help with captioning? The third to the fifth rows in Tab. 1 measure the benefit of training the speaker 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. Additionally, as the detector is not

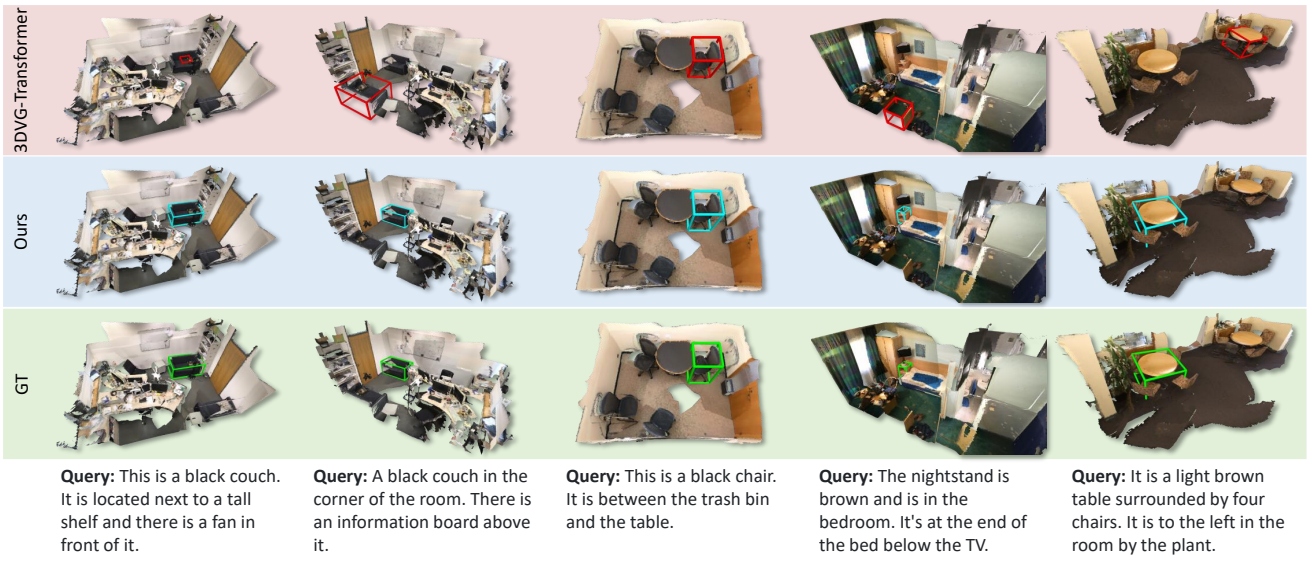


Figure 5. 3D visual grounding results using 3DVG-Transformer [64] and our method. 3DVG-Transformer fails to accurately predict object bounding boxes, while our method produces accurate bounding boxes and correctly distinguishes target objects from distractors.

	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)**

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.

only fine-tuned with the speaker but also with the listener, the additional supervision from the listener helps with the detection performance as well. A manual analysis (Tab. 5) also 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.

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.

Does additional data help? As our method allows us to expand the training data beyond the ScanNet subset annotated with descriptions, we also examine whether training with additional data helps to improve captioning. Additional training scans improve our speaker-listener in all captioning metrics (Tab. 1, last three rows). This increasing trend indicates the benefit of training with additional ScanNet data, enabled by our speaker-listener architecture.

Does better detection backbone help? Tab. 4 shows results with different detection backbones as well as with GT bounding boxes. 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 improvements from VoteNet [44] to PointGroup [27] show the benefit of a better detection backbone. The gap between GT and VoteNet/PointGroup shows there is room for further improvement.

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. Our proposed architecture explicitly reinforces the speaker to generate discriminative descriptions so that the listener can better localize the described targets given those descriptions. 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. Overall, we hope that our work will encourage more future research in 3D vision and language field.

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 by the Canada CIFAR AI Chair program and an NSERC Discovery Grant.

References

[1] Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. ShapeGlot: Learning language for shape differentiation. In *Proceedings of the IEEE international conference on computer vision*, 2019. 2, 3

[2] 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 *European Conference on Computer Vision*, pages 422–440. Springer, 2020. 1, 2, 5, 12, 13, 14

[3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *Proceedings of the International Conference on 3D Vision*, pages 667–676. IEEE, 2017. 1

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 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*, pages 202–221. Springer, 2020. 1, 2, 3, 5, 6, 12, 13, 15

[7] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 1, 2, 3, 4, 5, 6, 7, 8, 12, 13

[8] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*, pages 100–116. Springer, 2018. 2

[9] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3D instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15467–15476, October 2021. 14

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 5, 13

[11] Peter Cole and Jerry L Morgan. Syntax and semantics. volume 3: Speech acts. *Tijdschrift Voor Filosofie*, 39(3), 1977. 2

[12] Angela Dai and Matthias Nießner. 3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 3, 14

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 2, 5, 12

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[15] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[16] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, 2018. 3

[17] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419, 2010. 2

[18] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European conference on computer vision*, pages 3–19. Springer, 2016. 4

[19] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 1

[20] Ji Hou, Angela Dai, and Matthias Nießner. RevealNet: Seeing behind objects in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 1

[21] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021. 1

[22] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 2

[23] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 2

[24] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 92–101. IEEE, 2016. 1

[25] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3D instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021. 6

[26] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[27] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-set point grouping for 3D instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 2, 3, 4, 5, 6, 8, 13, 14

[28] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018. 2

[29] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 2

[30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[32] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2019. 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 14

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[35] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 338–354, 2018. 1, 2, 4

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[37] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 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*, pages 375–383, 2017. 2

[38] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111, 2017. 3

[39] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018. 1, 2, 4

[40] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 3

[41] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 14

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[44] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep Hough voting for 3D object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1, 2, 6, 8, 14

[45] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 4

[46] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 2, 4

[47] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-language model for 3D visual grounding. In *Proceedings of the Conference on Robot Learning*, 2021. 2

[48] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758. Springer, 2020. 2

[49] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[50] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3D objects. In *Proceedings of the Conference on Robot Learning*, 2021. 2

[51] traveller59. spconv. https://github.com/traveller59/spconv, 2021. 13

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[53] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3

[54] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[55] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992. 3

[56] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. 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. 2

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[58] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017. 2

[59] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 1

[60] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290, 2017. 3, 4

[61] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2

[62] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. 1, 2, 6

[63] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3DNet: 3D object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 14

[64] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 1, 2, 3, 6, 8, 12, 14

[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 1

Supplementary Material

In this supplementary material, we provide results on the ReferIt3D dataset in Sec. A. To showcase the effectiveness of our speaker-listener architecture, we provide additional results on extra ScanNet [13] data in Sec. B. We also include details about our PointGroup implementation as well as the detection and segmentation results in Sec. C and Sec. D, respectively.

A. Experiments on ReferIt3D

A.1. Quantitative Results

We conduct additional experiments on the ReferIt3D Nr3D dataset [2]. It contains about 33k free-form object descriptions annotated by human experts for training and 8k for validation. We report our results on the validation split since there is no test set.

A.1.1 3D dense captioning

We compare our 3D dense captioning and object detection results against the baseline Scan2Cap [7] in Tab. 6. Our method trained with the speaker MLE loss (marked "Ours (MLE)") outperforms Scan2Cap by a big margin, leveraging the improved object detection backbone. After training with the CiDEr reward (marked "Ours (CiDEr)"), our dense captioning results are further boosted. Training with the listener loss as the additional reward (marked "Ours (CiDEr+lis.)") further improves our results due to the explicit reinforcement of the discriminability of generated object descriptions. Here, our object detection mAP is also improved due to the end-to-end joint fine-tuning of our speaker-listener architecture. We showcase the effectiveness of training with extra ScanNet data in the last row in Tab. 6, where 3D dense captioning and object detection results are improved simultaneously.

	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5
Scan2Cap [7]	22.38	13.87	20.44	47.96	33.21
Ours (MLE)	33.85	20.70	23.13	53.38	49.71
Ours (CiDEr)	36.79	21.12	23.91	53.83	50.89
Ours (CiDEr+lis.)	37.35	21.40	24.10	54.14	51.58
Ours (CiDEr+lis.+extra)	**38.42**	**22.22**	**24.74**	**54.37**	**52.69**

Table 6. Quantitative results on 3D dense captioning and object detection on ReferIt3D Nr3D dataset [2]. 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 method outperforms the baseline Scan2Cap [7] by a significant margin. We showcase the effectiveness of our speaker-listener architecture trained with partially annotated ScanNet data, where it achieves the best performance in all metrics.

	Acc@0.5IoU		
	Unique	Multiple	Overall
ScanRefer [6]	-	12.17	12.17
3DVG-Trans [64]	-	14.22	14.22
Ours (w/o fine-tuning)	-	19.64	19.64
Ours (w/ fine-tuning)	-	24.41	24.41
Ours (w/ fine-tuning + extra)	-	**25.23**	**25.23**

Table 7. Quantitative results on 3D visual grounding on ReferIt3D Nr3D dataset [2]. We adapt the evaluation setting as in Chen et al. [6] to be consistent with the main paper. We report results on "Multiple" and "Overall", as there is no case in ReferIt3D that is "Unique". Our base visual grounding network outperforms the baseline methods. Results are further improved after the joint fine-tuning with the speaker-listener architecture. Speaker-listener fine-tuning and semi-supervised training with partially annotated ScanNet data provide the best overall results.

A.1.2 3D visual grounding

We compare our 3D visual grounding results against the baseline ScanRefer [6] and 3DVG-Transformer [64] in Tab. 7. As the descriptions in ReferIt3D dataset all refer to objects in the scene where multiple similar objects with the same class label are present, there is no such case that can be allocated to "Unique" subset where only one object with a specific class label can be found in the scene. Therefore, we allocate our results to "Multiple" and "Overall". Our method trained with the detector loss and the listener loss (marked "Ours(w/o fine-tuning)") clearly outperforms the baseline methods. Our results (marked "Ours(w/ fine-tuning)") are significantly improved after fine-tuning jointly with the speaker. Our best results are obtained after jointly training with speaker-listener architecture on partially annotated ScanNet data, as demonstrated in the last row in Tab. 7.

A.2. Qualitative Analysis

A.2.1 3D dense captioning

We compare our results with object captions from Scan2Cap [7] in Fig. 6. Object captions generated by Scan2Cap include more inaccurate spatial relationships. Also, those object captions cannot be used to uniquely localize the associated object. In contrast, our method produces more accurate and discriminative object captions with more spatial relationship information.

A.2.2 3D visual grounding

Fig. 7 compares our results with 3DVG-Transformer [64] on ReferIt3D Nr3D dataset [2]. 3DVG-Transformer clearly suffers from overfitting issue, as it tends to predict that same object bounding box given different queries as inputs (see

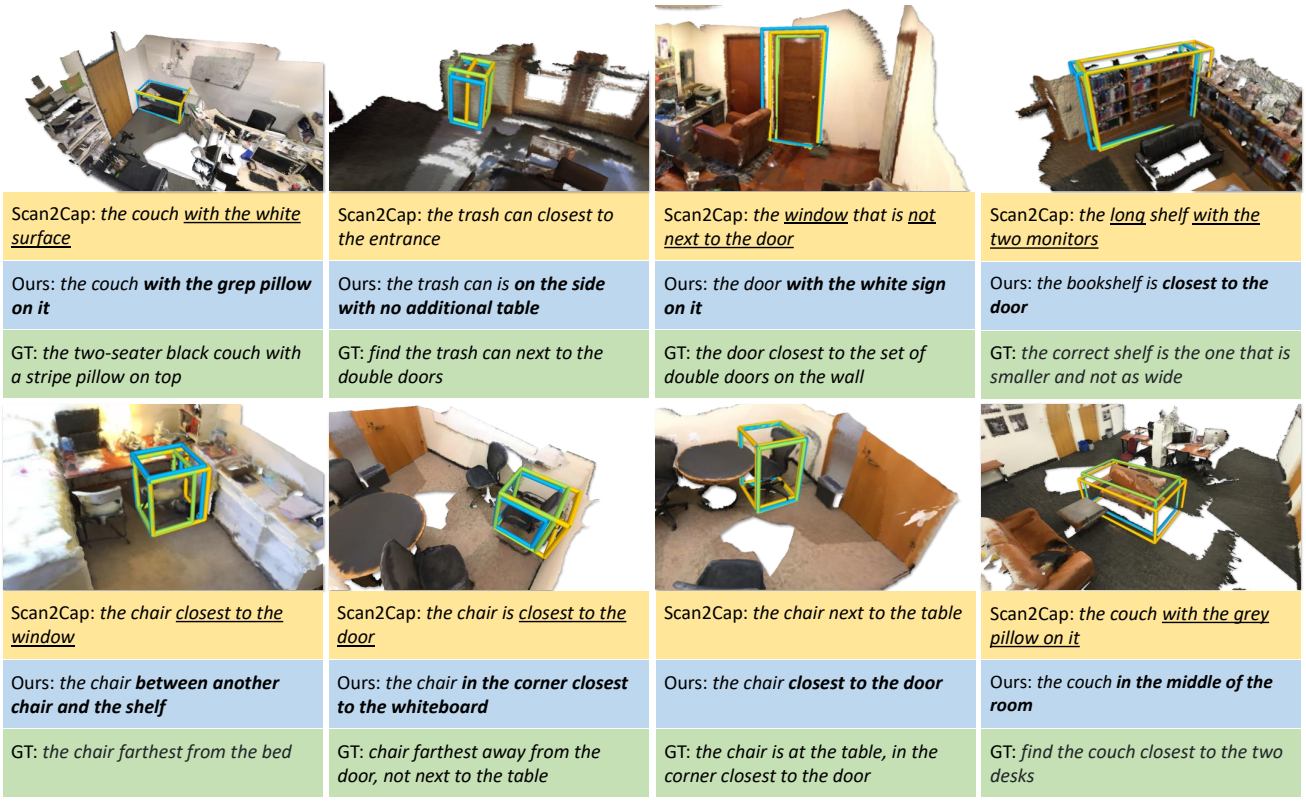


Figure 6. Qualitative results in 3D dense captioning task from Scan2Cap [7] and our method on ReferIt3D Nr3d dataset [2]. We underline the inaccurate words and mark the spatially discriminative phrases in bold.

Method	mAP@0.5	cab.	bed	chair	sofa	tab.	door	wind.	booksh.	pic.	cntr	desk	curt.	refrige.	s. curt.	toil.	sink	batht.	other
PG (*)	56.9	48.1	69.6	**87.7**	71.5	62.9	42.0	46.2	54.9	37.7	22.4	41.6	44.9	37.2	64.4	98.3	**61.1**	80.5	53.0
PG (Color)	56.6	47.5	64.1	83.8	**75.4**	63.7	42.7	45.7	49.6	**43.7**	17.5	42.9	**47.9**	35.0	65.6	**100.0**	60.7	**81.9**	51.5
PG (Multiview)	**62.8**	**58.3**	**83.4**	86.9	66.3	**68.6**	**47.3**	**52.4**	**64.9**	38.3	**23.0**	**56.9**	46.3	**64.3**	**83.0**	98.3	57.0	71.4	**63.1**

Table 8. Comparison of the performance of our implementation of PointGroup using the Minkowski Engine against the original PointGroup (PG(*)) for instance segmentation. We report the mAP for IoU threshold 0.5 on the ScanNet v2 validation set. Our re-implementation using color gives comparable performance as the original PointGroup implementation. Using multiview features, we are able to further improve the performance.

the third and fourth examples in the first row). Leveraging the speaker-listener architecture, our method can better distinguish object from the same class than 3DVG-Transformer.

B. Additional Results on Extra ScanNet Data

Fig. 8 showcase the intermediate dense captioning and visual grounding results for scans where no GT object captions are provided in the ScanRefer dataset [6]. Those intermediate object captions and the matched object bounding boxes are used during the semi-supervised training of our speaker-listener architecture. Our architecture produces plausible object captions with adequate and discriminative spatial relationships that inherently enables visual grounding.

C. PointGroup Implementation Details

The official implementation of PointGroup uses SpConv [51], a spatially sparse convolution library devoted to 3D data, to build its SparseConv-based U-Net architecture to encode point and cluster representation. We migrated the implementation of PointGroup from SpConv to MinkowskiEngine [10], another auto-differentiation library for sparse tensors, since it outperformed SpConv by providing faster computation operations on GPU, user-friendly documentations and consistent code maintenance at the time the project was initiated.

Following Jiang et al. [27], we use the same hyperpa-

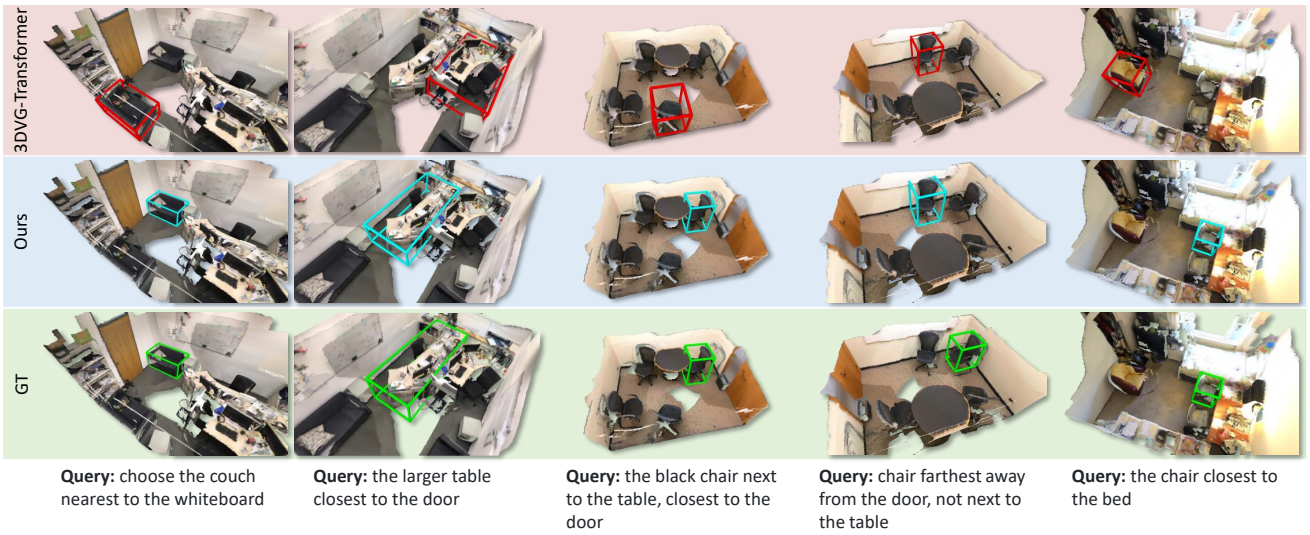


Figure 7. 3D visual grounding results using 3DVG-Transformer [64] and our method on ReferIt3D Nr3D dataset [2].

rameters for point voxelization and clustering. We set the maximum number of points per scene to $250,000$ by randomly adding small offsets to the point cloud and cropping out extra parts exceeding the predefined maximum scale of the scene if necessary. Limiting the number of points to $250,000$ allows us to fit the model on a RTX 3090. We augment each point cloud scene by jittering point coordinates slightly, mirroring about the YZ-plane, and rotation about the Z axis (up-axis) randomly from 0 to $360°$. We also apply elastic distortion, which was used by Jiang et al. [27], to the scaled points. We share the same SparseConv-based U-Net architecture as Jiang et al. [27] for both backbone and ScoreNet except that the input data may contain mutiview features and normals instead of RGB colors. For each voxel, we encode the color, normal and multiview features extracted using ENet [12], giving us a total input dimension of 134. To adapt PointGroup as an object detector, we obtain axis-aligned bounding boxes using predicted instance clusters by simply calculating their sizes and centers from points assigned to them. We set the thresholds of cluster scores as 0.09 and the minimum cluster point number as 100 to filter out bad cluster proposals. We train the PointGroup detector using Adam [33] with a learning rate of 2e-3, on the ScanNet train split with batch size 4 for 140k iterations until convergence.

D. Detection and Segmentation Results

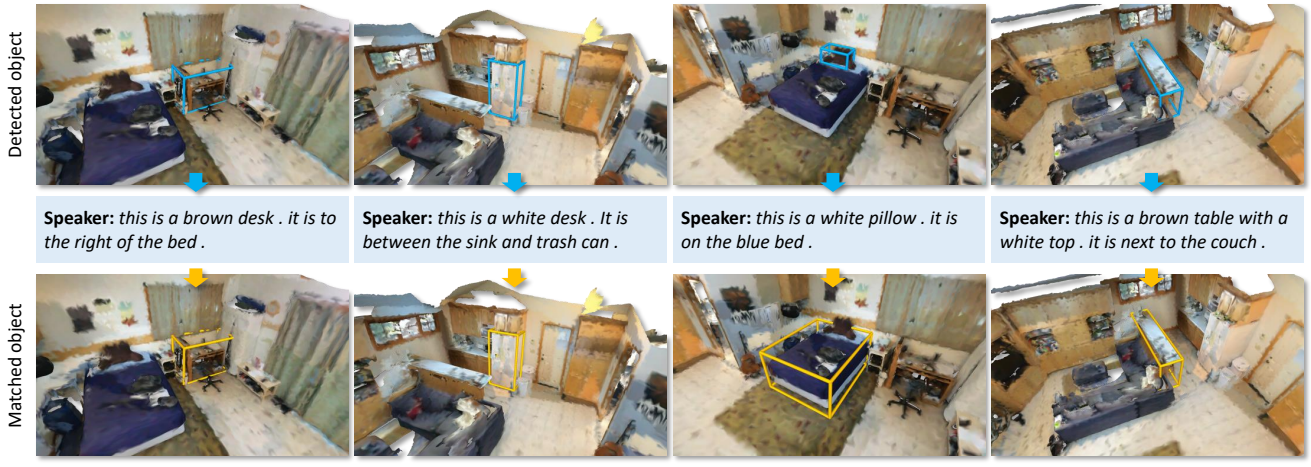
D.1. Quantitative results

Instance segmentation. Tab. 8 compares the instance segmentation results of our PointGroup implementation against the original PointGroup (first row). With positions and colors as input, our implementation of PointGroup (second row) gives a similar performance as original PointGroup. When replacing colors with multiview features and normals (last row), our PointGroup implementation significantly outperforms the original one. Our multiview-based PointGroup gives mAP@0.5 of 62.8, which is close to the performance of the current state-of-the-art model HAIS [9], which achieves 64.1 on the validation set of ScanNet v2. Our implementation also surpasses the original PointGroup in training speed: given point coordinates and colors as input, it takes less than two days to train the model in our implementation, while the original one could take up to three days until convergence.

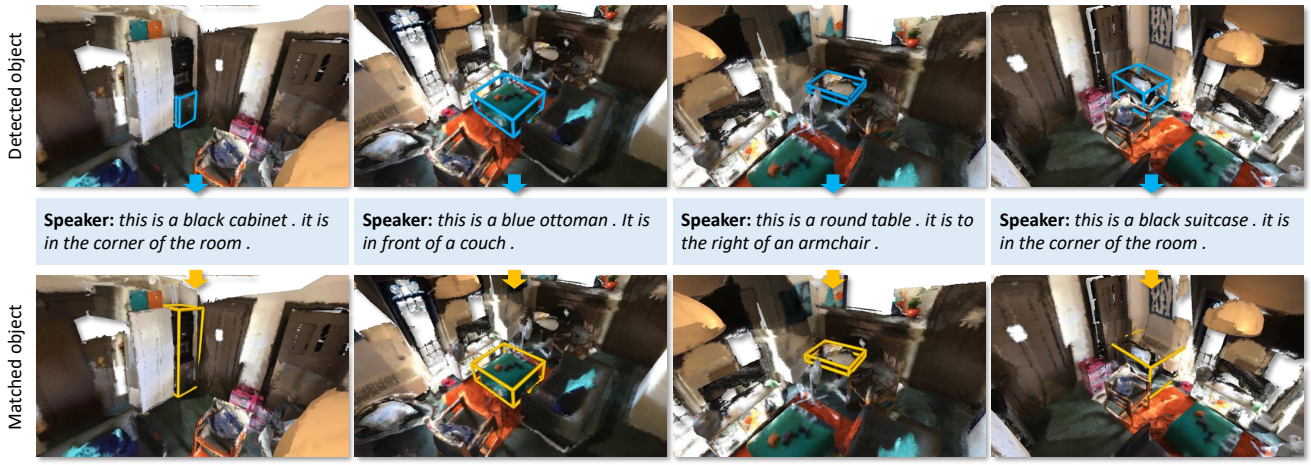
Object Detection. We compare our object detection results before fine-tuning with the speaker-listener architecture against the VoteNet [44] in Tab. 9. Given positions and colors as input, our PointGroup detector (second row) clearly outperforms VoteNet. Using multiview features and normals instead of RGB colors, our PointGroup based detector gives improved detection results of 50.7 mAP@0.5, which outperforms the current state-of-the-art detectors [41, 63] on the validation set of ScanNet v2 with gains of 3.7 and 2.6 respectively. Also, our PointGroup generates notably better detections for small and thin objects than VoteNet, such as picture ("pic.") and counter ("cntr").

D.2. Qualitative results

Instance segmentation. We present our instance segmentation results in Fig. 9. Our PointGroup trained with multiview features and normals clearly generates better instance segmentation masks than our model with raw point colors as input, as it better segments out tiny objects leveraging the



(a) Scan scene0000_01



(b) Scan scene0002_01

Figure 8. Intermediate dense captioning and visual grounding results in the Speaker-listener architecture for RGB-D scans where no GT object descriptions are provided in ScanRefer dataset [6]

Method	mAP@0.5	cab.	bed	chair	sofa	tab.	door	wind.	booksh.	pic.	cntr	desk	curt.	refrige.	s. curt.	toil.	sink	batht.	other
VoteNet	33.5	8.1	76.1	67.2	**68.8**	42.4	15.3	6.4	28.0	1.3	9.5	37.5	11.6	27.8	10.0	86.5	16.8	**78.9**	11.7
PG (Color)	44.6	25.2	69.1	77.1	67.3	53.3	32.7	32.2	36.8	26.9	30.0	52.1	**33.5**	26.7	37.4	87.8	32.3	69.6	13.6
PG (Multiview)	**50.7**	**36.4**	**77.6**	**80.9**	66.1	**59.2**	**40.2**	**33.1**	**37.0**	**27.7**	**32.0**	**56.5**	32.2	**62.1**	**70.0**	**91.1**	**33.8**	60.2	**16.0**

Table 9. Comparison of object detection performance of PointGroup (PG) and VoteNet. We report mAP with IoU threshold 0.5 on the ScanNet v2 validation set. PointGroup produces more accurate bounding boxes than VoteNet, and using multiview features further improves performance over incorporating color as input directly.

higher resolution of the multiview images.

Object Detection. Fig. 10 showcases the effectiveness of our PointGroup in object detection over VoteNet. Our PointGroup implementation produces much more accurate object bounding boxes due to the fine-grained per-point segmentation. Also, training with multiview normal features can further improve the quality of the generated bounding boxes in comparison with PointGroup trained with the raw point colors (the third column vs. the first column).

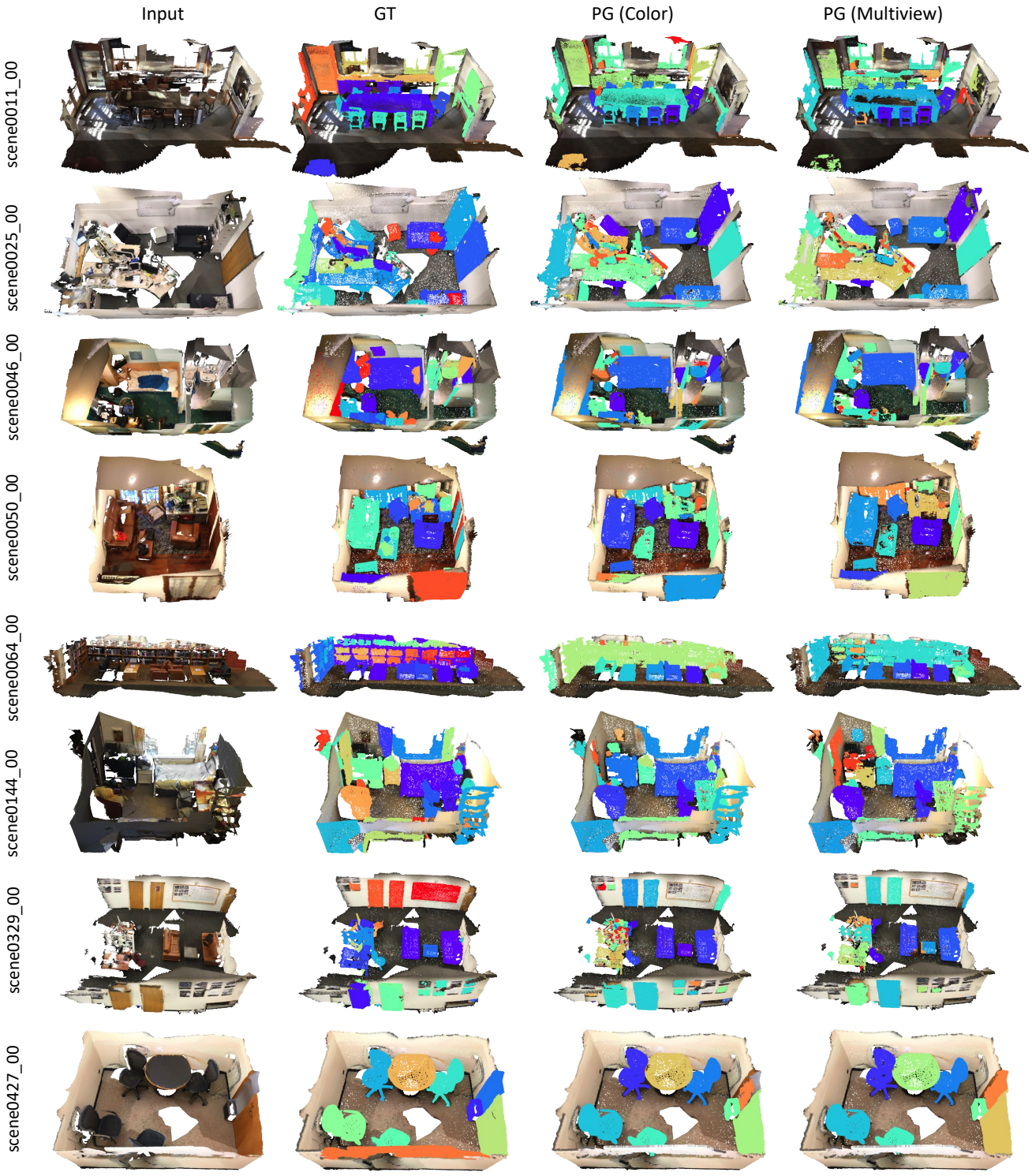


Figure 9. Qualitative results in instance segmentation task on the ScanNet v2 validation set.

Figure 10. Qualitative results in object detection task on the ScanNet v2 validation set.