UniT3D: A Unified Transformer for 3D Dense Captioning and Visual Grounding Supplementary Material

Dave Zhenyu Chen[1] Ronghang Hu[2] Xinlei Chen[2] Matthias Nießner[1] Angel X. Chang[3]

[1]Technical University of Munich [2]Meta AI [3]Simon Fraser University

In this supplementary material, we provide detailed dense captioning results on the ScanRefer dataset in Sec. 1. To showcase the effectiveness of the proposed pre-training scheme and joint training objectives, we provide additional results and analysis in Sec. 2. We also include details about a re-evaluation of BUTD-DETR [6] in Sec. 3.

1. Detailed dense captioning results

	Captioning Precisions				Detection
	C@0.5IoU	B-4@0.5IoU	R@0.5IoU	M@0.5IoU	mAP@0.5
Scan2Cap [4]	10.21	5.85	9.70	4.67	32.09
X-Trans2Cap [10]	11.04	6.00	11.54	3.92	35.31
MORE [7]	10.30	5.49	11.14	3.87	33.75
3DJCG [1]	13.47	9.19	16.31	5.66	39.75
D3Net [4]	18.24	11.04	31.53	7.47	50.93
D3Net [4] (CIDEr loss)	30.83	16.70	26.24	11.48	53.85
Ours (from scratch)	19.92	10.25	19.43	9.28	53.91
Ours (w/ pre-training)	**22.41**	**13.70**	**23.07**	**11.11**	**54.03**

(a) 3D Dense Captioning Precisions

	Captioning Recalls				Detection
	C@0.5IoU	B-4@0.5IoU	R@0.5IoU	M@0.5IoU	mAP@0.5
Scan2Cap [4]	39.08	23.32	44.48	21.97	32.09
X-Trans2Cap [10]	43.87	25.05	44.97	22.46	35.31
MORE [7]	40.94	22.93	44.42	21.66	33.75
3DJCG [1]	**49.48**	**31.03**	50.80	24.22	39.75
D3Net [4]	46.07	30.29	**51.67**	**24.35**	50.93
D3Net [4] (CIDEr loss)	62.64	35.68	53.90	25.72	53.85
Ours (from scratch)	40.40	25.60	44.75	21.26	53.91
Ours (w/ pre-training)	46.69	27.22	45.98	21.91	**54.03**

(b) 3D Dense Captioning Recalls

Table 1.1: The 3D dense captioning precisions and recalls on Scan2Cap [4] validation set. All reported metrics are thresholded by IoU 0.5. Our method achieves strong dense captioning precisions and competitive dense captioning recalls in comparison to previous methods. Note that we compare to D3Net [3] trained only with the cross-entropy objective for a fair comparison.

We present the detailed dense captioning precisions and recalls in Tab. 1.1. To keep the comparison consistent and fair, all methods presented here are trained with the cross-entropy objective only, including D3Net [3]. As discussed in the Experiments Section in the main paper, the previous evaluation protocol in Chen et al. [4] mainly covers the dense captioning recalls without penalizing false positives. This protocol only takes the number of GT boxes into account, neglecting the fact that up to infinite predictions can be produced without being punished. We further detailed the dense captioning precisions and recalls, as displayed in Tab. 1.1a and Tab. 1.1b. For some previous methods with VoteNet [8] backbone such as MORE [7] and 3DJCG [1], their dense captioning precisions are notably lower than the other methods with stronger detection backbone such as D3Net [3]. To further showcase the impact of having cleaner box predictions, we visualize the predicted boxes with captions in Fig. 1.1. Despite having a slightly lower dense captioning recall, our method still produces much more plausible bounding box predictions, resulting in a strong dense captioning precision and F1-score compared with the previous methods.

2. Further training details and analysis

As the synthetic data contain many noise samples, we continue the joint training scheme with the bidirectional and seq-to-seq objectives after the convergence on the synthetic data. Additionally, to make sure the multimodal representation contains task-specific information, we further fine-tune the network with the training objective of the specific target task (*i.e*. bidirectional for grounding and seq-to-seq for captioning) on ScanRefer as the final training stage. To show the effectiveness of the joint training objective, we report the intermediate training steps for “joint from scratch” and “joint fine-tuned” in Sec. 4.5 of the main paper.

In particular, for “joint from scratch”, we follow a two-stage training strategy. We first train the network from scratch on ScanRefer with both bidirectional and seq-to-seq objectives (a), then continue training the network on ScanRefer with the target objective (b). As shown in Tab. 2.1 and Tab. 2.2, such two-stage *joint-to-target* training strategy can effectively improve both visual grounding accuracy and dense captioning results. These improvements indicate that our network is capable of learning and sharing a strong joint representation across two downstream tasks.

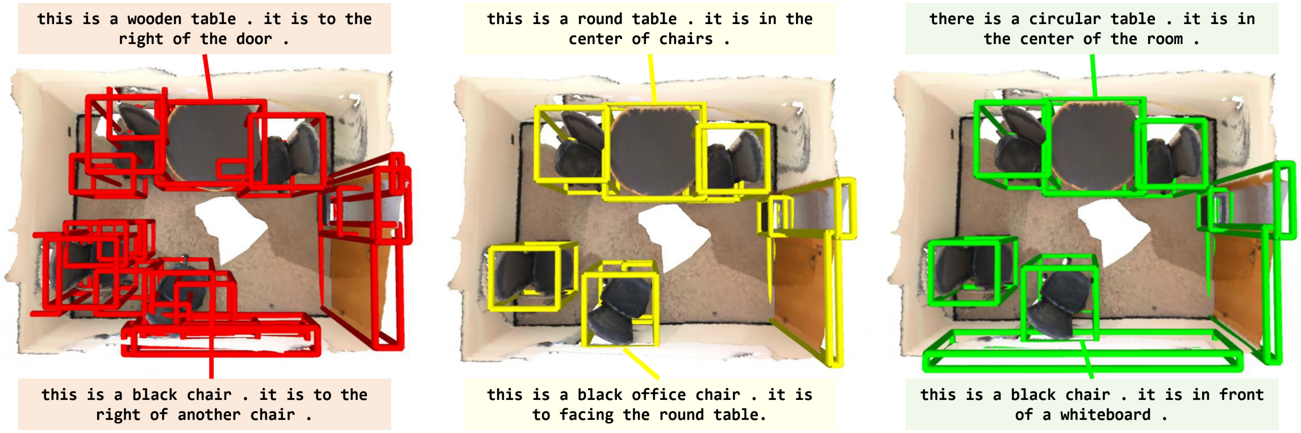
this is a wooden table . it is to the right of the door .
this is a round table . it is in the center of chairs .
there is a circular table . it is in the center of the room .
this is a black chair . it is to the right of another chair .
this is a black office chair . it is to facing the round table.
this is a black chair . it is in front of a whiteboard .

Figure 1.1: Detected boxes with captions from 3DJCG [1] (in red boxes), our method (in yellow boxes), and ground truths (in green boxes). Our method generates much fewer and cleaner box predictions when compared to those from 3DJCG. This results in a much higher dense captioning precision. Best viewed in color.

	Training Dataset(s)		Training Objective(s)		Visual Grounding Accuracy		
Training setup	Synthetic	ScanRefer	Bidirectional	Seq-to-Seq	Unique@0.5IoU	Multiple@0.5IoU	Overall@0.5IoU
(a) joint from scratch		✓	✓	✓	72.30	28.41	36.85
(b) continue from (a)		✓	✓		**73.68**	28.84	37.45
(c) joint from pre-trained	✓	✓	✓	✓	72.63	30.67	38.81
(d) continue from (c)		✓	✓		73.14	**31.05**	**39.14**

Table 2.1: 3D visual grounding results on ScanRefer [2] with detailed pre-training and joint training steps. (a) When trained from scratch on ScanRefer [2] with joint training objectives, our model already has a strong performance on par with the previous SOTA. (b) Continuing fine-tuning from (a) solely with the bidirectional objective improves the visual grounding results. (c) Jointly training with both objectives from pre-trained weights on the synthetic data, it achieves better visual grounding results in comparison with jointly training from scratch (a). (d) Continuing fine-tuning from (c) with the bidirectional objective, our final setting achieves the best overall visual grounding results.

	Training Dataset(s)		Training Objective(s)		Dense Captioning F1-Scores			
Training setup	Synthetic	ScanRefer	Bidirectional	Seq-to-Seq	CIDEr@0.5IoU	BLEU-4@0.5IoU	ROUGE-L@0.5IoU	METEOR@0.5IoU
(a) joint from scratch		✓	✓	✓	26.48	14.64	27.10	12.92
(b) continue from (a)		✓		✓	27.28	17.22	29.12	13.74
(c) joint from pre-trained	✓	✓	✓	✓	29.77	17.78	30.10	14.28
(d) continue from (c)		✓		✓	**30.28**	**18.23**	**30.72**	**14.74**

Table 2.2: 3D dense captioning results on ScanRefer [2] with detailed pre-training and joint training steps. (a) Our model without pre-training already demonstrates competitive performance against the previous SOTA. (b) Continuing fine-tuning from (a) solely with the seq-to-seq objective improves the dense captioning results. (c) Jointly training the network with pre-trained weights on synthetic data, it achieves better dense captioning results in comparison with jointly training from scratch (a). (d) Continuing fine-tuning from (c) with the seq-to-seq objective, our final setting achieves the best overall dense captioning results.

Similarly, to show the advantage of pre-training on the distillate 2D priors, we report the intermediate results of the two-stage training steps for "joint fine-tuned". We first train the network from scratch on the synthesized data with both bidirectional and seq-to-seq objectives (c), then fine-tune the pre-trained network on the downstream tasks with the respective target objective (d). By comparing (c) with (a), we observe a clear performance boost in both downstream tasks. Further improvements can be observed after the final fine-tuning step on the downstream task. Such improvements further validate the effectiveness of expanding the multimodal representation learning to distillate 2D data.

	Val Acc@0.25IoU			Val Acc@0.5IoU		
	Unique	Multiple	Overall	Unique	Multiple	Overall
Original	84.20	46.60	52.20	66.30	35.10	39.80
Re-evaluated	82.77	44.01	49.69	63.81	33.51	38.01

Table 3.1: 3D visual grounding accuracy of BUTD-DETR [6]. We re-evaluated BUTD-DETR [6] by removing the GT object labels in the text queries from the original implementation.

3. Re-evaluation of BUTD-DETR

We notice that the input text queries of BUTD-DETR [6] in the official implementation differ from the evaluation protocol in the other work [2, 5, 9, 11, 1, 3], where the GT object labels are manually added to the text. For instance, given a query for a table "this is a round wooden object. it is between two black chairs.", the official implementation adds the GT object label "table" to the query as "this is a round wooden object. table. it is between two black chairs.". Such augmentation during evaluation leads to three problems: 1) Using the GT object labels during inference results in unfair comparison; 2) The rich relationships in the language cues are neglected, as the grounding model tends to rely on the object names to distinguish objects in the scene; 3) Some difficult cases are aided by exposed GT information where the target is simply referred as "object" in the query, as in the aforementioned example. For the purpose of having a fair and consistent comparison, we re-evaluate BUTD-DETR [6] by removing the additional object names in the input texts. The re-evaluated visual grounding results against the ones in the original paper [6] are displayed in Tab. 3.1.

References

[1] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3DJCG: A unified framework for joint dense captioning and visual grounding on 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022. 1, 2, 3

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. 2, 3

[3] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. *arXiv preprint arXiv:2112.01551*, 2021. 1, 3

[4] 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

[5] 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. 3

[6] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022. 1, 3

[7] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. *arXiv preprint arXiv:2203.05203*, 2022. 1

[8] 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

[9] 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. 3

[10] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3D dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022. 1

[11] 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. 3