Joint Architecture for 3D Vision Language

Robin Borth
Technical University of Munich

Yaomengxi Han
Technical University of Munich

Abstract

3D visual grounding and 3D dense captioning are similar in that they both require an understanding of cross-modality relationships. Previous attempts to jointly solve these tasks have not fully exploited these relationships, instead only superficially enhancing one modality with the other. We propose a novel approach, VL3DNet, for jointly solving these two tasks. Our method utilizes a shared vision-language transformer module to enhance the vision and language modalities simultaneously, effectively exploiting the intermodal relations. Compared to previous attempts, such as D3Net [1] and 3DJCG [2], our method has significant improvement in visual grounding and dense captioning.

1. Introduction

There has been growing research [3–6] in the area of 3D vision-language, most of which focused on two tasks, 3D visual grounding and 3D dense captioning. 3D visual grounding refers to the task of identifying the object in a 3D scene based on a given query, while 3D dense captioning is the task of generating discriminative natural language descriptions for all objects in a 3D scene.

Previous studies attempted to leverage the complementary nature of the two tasks by either training them jointly [7] or enhancing one modality with the other [2]. However, they did not completely encompass the connections between the spatial relationships of objects and their associated text descriptions, which is vital for both tasks. Motivated by this, we propose a novel task-agnostic vision-language fusion module, which allows us to effectively exploit the relative relations between objects and the relationship between objects and descriptions. This module can enrich both vision and language modal with cross-modal information, hence we only need lightweight task-specific heads for both downstream tasks. Furthermore, by utilizing this fusion module and a joint loss function, we are able to train the two tasks jointly, resulting in improved performance compared with training them separately.

Our work has two major contributions: First, we employ a transformer-based vision-language fusion module, which enhances two modalities simultaneously, allowing for the use of lightweight MLP heads for two downstream tasks. Second, we propose a joint training scheme that improves performance for both tasks by incorporating a joint loss.

2. Related Works

3D Object Detection The task of 3D object detection involves identifying all objects present in a scene using a 3D point cloud as input. In recent years, various detection methods have been proposed, such as VoteNet [8], PointGroup [9], SoftGroup [10], among others.

D3Net [1] has demonstrated that a strong detector can significantly enhance the overall performance of a vision-language system. They proposed using PointGroup as the detection backbone. However, PointGroup, along with other backbones such as PointNet [11], its variant [12], and VoteNet, struggle with accurately distinguishing between instances located close to each other.

Recently, SoftGroup [10] has overcome this issue and demonstrated superior performance in the ScanNet [7] instance segmentation challenge. SoftGroup is designed to be more robust and flexible than traditional PointGroup clustering methods, as it uses a soft assignment approach which allows each data point to belong to multiple clusters with varying degrees of association with them. This results in a more nuanced representation of the data and can lead to improved performance. Based on its successful performance on the ScanNetv2 benchmark, SoftGroup is used as the detection backbone in our system.

3D Vision-Language Fusion There is a growing interest in research on 3D vision and language fusion, and the two most well-studied domains are 3D visual grounding and dense captioning. Most prior research on 3D visual grounding [4, 5, 15, 16] focuses on identifying relationships between objects and natural language queries. On the other hand, prior works in dense captioning [3, 17, 18] learn relationships between objects in the scene and translate them into natural language descriptions. Due to the complemen-

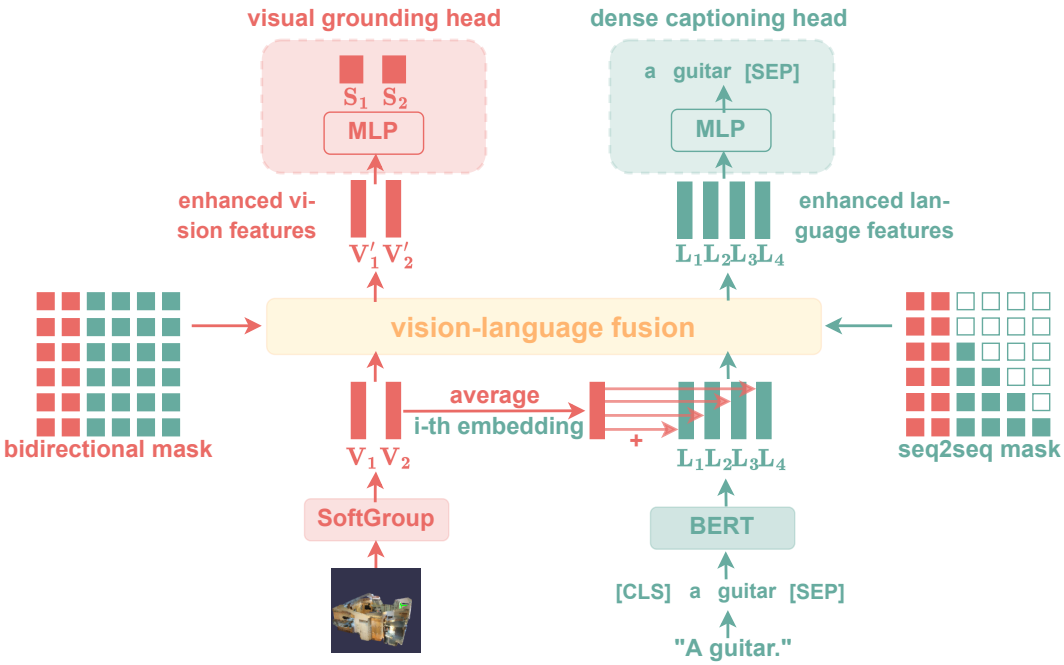

Figure 1. Overall architecture: SoftGroup detector extracts coordinates and features of each proposal and concatenates them before fed into a vision backbone mapper, which outputs 768-dimensional feature embeddings. On the other hand, queries are tokenized by the BERTtokenizer [13], and fed into BERT with a triangular mask. The vision and language embeddings are then concatenated and fed into the fusion module, a stack of few self-attention [14] layers. Then the enhanced vision representations are fed into the visual grounding head, which is a simple MLP and outputs a score for each detected object; while the enhanced language representations are fed into the dense captioning head, which generates the caption for the requested object autoregressively with teacher forcing.

tary nature of these two tasks, the study of combining vision and language information to enable the joint training of grounding and captioning has become a highly debated topic in both 2D [19–22] and 3D communities. In the field of 3D vision-language understanding, 3DJCG attempted to enhance the vision features using a transformer-based fusion model that also considers language features, however, it was unable to effectively enhance the language modalities, resulting in the use of complex task-specific head. Conversely, D3Net focused on building a speaker-listener pipeline. Both approaches failed to handle both modalities in a single transformer-based architecture and to fully leverage intermodal relations for enhanced representations.

3. Methods

3.1. Model Architecture

Motivated by the lack of learning for unified representation of vision and language modality in existent works, our model consists of five components: a SoftGroup detection backbone, a BERT language backbone, a transformer-based task-agnostic vision-language fusion module, and two task-specific heads. The overall architecture is shown in Fig.1 SoftGroup takes a 3D point cloud $P \in R^{K+3}$ from ScanNet [7] as input, and outputs the coordinates and a 32-dimensional feature embedding for each object detected in the scene. The feature and coordinates are then concatenated to form 38-dimensional embeddings, represented as $V_i \in R^{32+6}$, which are subsequently fed into a mapper for alignment with BERT language embeddings. In contrast, the language backbone of the system uses a query from the ScanRefer [4] as input and produces an encoded language embedding, represented as $L_i \in R^{m \times D}$. Here, m represents the token length, including the [CLS] and [SEP] tokens, and D denotes the dimension of each token embedding, which is typically set to 768 when utilizing a standard BERT model. It is worth noting that a triangular attention mask is also incorporated into the BERT module to enforce sequence-to-sequence language modeling. In dense captioning mode, we aim to generate a unique description for each object by performing separate forward passes. To keep track of which object is being described in each pass, we've introduced a global vision cue. This cue is the embedding of the object being described, which is added to all language token embeddings. Additionally, to ensure consistency in the language input for the fusion module, we've also added a global vision cue to the token embeddings in grounding mode. This cue is calculated as the average of all

proposal embeddings in that particular scene. Finally, the vision-language fusion module combines both vision and language embeddings to produce enhanced representations. The visual grounding and dense captioning heads take the enhanced vision and language representations as separate inputs and produce the final results.

3.2. Different Training Schemes

The model outlined in the previous section can be trained using three different approaches. The first approach is training only on the visual grounding task, in which the total loss $L = L_{\text{VG}} + L_{\text{CLS}}$ is set as the sum of the cross entropy loss for visual grounding, L_{VG}, and the query classification loss, L_{CLS}, which acts as a proxy loss. Additionally, prior to feeding the language embeddings into the fusion module, the embedding of the queried object is added as a global vision cue. The second approach is training only with the dense captioning task, using the cross entropy loss, L_{DC}, between the generated description and reference. To align with visual grounding, the average vision embedding is added to the language embeddings before the fusion module, and a seq2seq language mask and teacher forcing are employed during training. The final approach is joint training, using the combined loss of $L = L_{\text{VG}} + L_{\text{CLS}} + L_{\text{DC}}$.

4. Experiments

We obtain all following results from the validation set of the ScanRefer [4] dataset.

4.1. Implementation details

We use SoftGroup implemented with Minkowski Engine as our detection backbone. The best checkpoint is obtained by training on ScanRefer dataset for 5 epochs with batch size 16. During the joint training, we use Adam optimizer and set the learning rate of our language backbone as 10^{-6} and $10^{-}3$ for the rest modules. Our architecture contains 141M trainable parameters and all experiments are conducted on GeForce RTX 2080 Ti GPU with PyTorch.

4.2. Evaluation Metrics

Visual Grounding For this task, we use Acc@kIoU as measuring criteria. This method involves setting a threshold for the intersection over union (IoU) between predictions and ground truths, and only counting a prediction as positive if the IoU is higher than the threshold. We use threshold values of 0.25 and 0.5 following Chen et al. [4]

Dense Captioning We use four captioning metrics, CIDEr, BLEU-4, ROUGE-L and METEOR following Scan2Cap [3]. Similar to visual grounding metrics, we also use IoU scores as threshold for these metrics. We define the precision and recall scores for dense captioning for each metric M separately as following:

$$\text{P@0.5IoU} = \frac{1}{|B^{\text{pred}}|} \sum_{i=1}^{|B^{\text{pred}}|} m_i u_i,$$

$$\text{R@0.5IoU} = \frac{1}{|B^{\text{gt}}|} \sum_{i=1}^{|B^{\text{gt}}|} m_i u_i.$$

Note that u_i above is an indicator $1\{\text{IoU}(B_i) \geq 0.5\}$. Finally, to compute the F1 score of the task we use the following formula:

$$\text{F1-score} = \frac{2 \times \text{P@0.5IoU} \times \text{R@0.5IoU}}{\text{P@0.5IoU} + \text{R@0.5IoU}}$$

4.3. Comparison with State-of-the-art Methods

We compare the performance of our model with other state-of-the-art methods for visual grounding and dense captioning and the results are shown in Tab.1 and Tab.2. For visual grounding, we also split the validation into 'Unique', 'Multiple' and 'Overall', where the 'Unique' split measures the localisation accuracy when in the current scene, there is only one instance of the same semantic label as the queried object, and the 'Multiple' split measures the accuracy when there are multiple instances of the same semantic labels as the queries object. For dense captioning, we measure the performance with the metrics describe in section **4.2**.

Methods	Unique		Multiple		Overall	
	a@0.25IoU	a@0.5IoU	a@0.25IoU	a@0.5IoU	a@0.25IoU	a@0.5IoU
ScanRefer [4]	0.6859	0.4352	0.3488	0.2097	0.4244	0.2603
D3Net [1]	**0.7923**	**0.6843**	0.3905	0.3074	0.4806	0.3919
3DJCG [2]	0.7675	0.6059	**0.4389**	0.3117	**0.5216**	0.3776
Ours	0.7291	0.6633	0.3926	**0.3125**	0.4940	**0.4172**

*Our result is obtained on the validation set instead of the benchmark test set.

Table 1. Quantitative results on visual grounding. Overall, our model performs better compared with the previous state-of-the-art models in some metrics. The results of other methods are obtained from the ScanRefer [4] benchmark.

Methods	Captioning F1-Score			
	C@0.5IoU	B-4@0.5IoU	R-L@0.5IoU	M@0.5IoU
Scan2Cap [3]	0.0849	0.0576	0.1073	0.0492
D3Net [1]	0.2088	0.1335	**0.2237**	0.1022
3DJCG [2]	0.1918	**0.1350**	0.2207	0.1013
Ours	**0.2849**	0.0698	0.2115	**0.2003**

*Our result is obtained on the validation set instead of the benchmark test set.

Table 2. Quantitative results on dense captioning. As is shown in the table, our model performs significantly better in terms of CIDEr and METEOR metrics compared with previous models.

4.4. Ablation Studies

In order to measure how the joint training scheme improves the performance on both tasks, we also carry out ablation studies with three different training schemes, as mentioned in section **3.2**. As is shown in Tab. 3, joint training can improve the performance of both tasks.

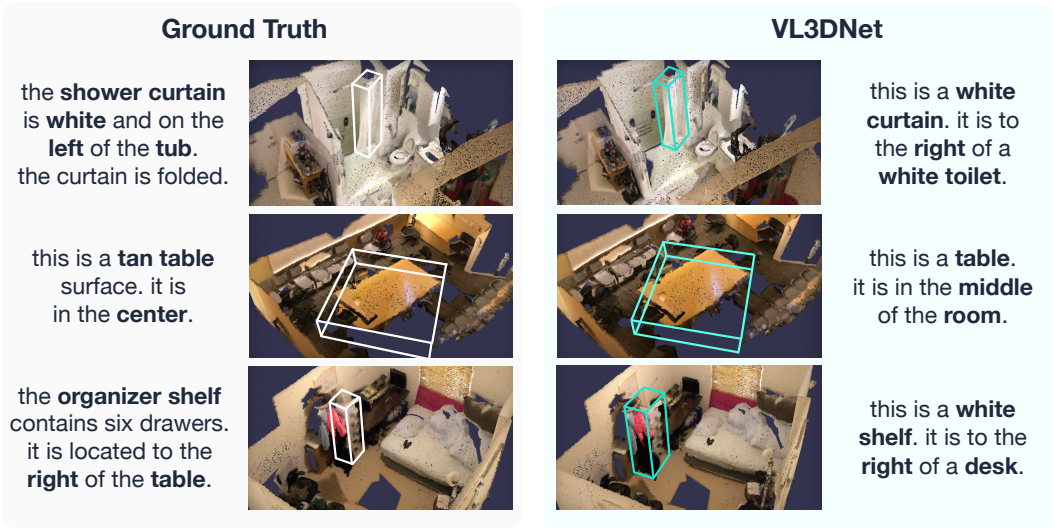

Figure 2. Qualitative results of our model on the ScanRefer dataset.Given a natural language query, the model is able to localize the described object accurately. The model also generates fluent and discriminative descriptions for the objects. We can see that the descriptions capture the spatial relations with other objects.

Training Scheme	Grounding acc@0.5IoU			Captioning F1-Score			
	Unique	Multiple	Overall	C@0.5IoU	B-4@0.5IoU	R-L@0.5IoU	M@0.5IoU
Grounding Only	0.6553	0.2980	0.4040	-	-	-	-
Captioning Only	-	-	-	0.2800	**0.0702**	0.2091	0.1995
Joint Training	**0.6633**	**0.3125**	**0.4172**	**0.2849**	0.0698	**0.2115**	**0.2003**

Table 3. Ablation studies show that joint training improves the performance of our model on both tasks, results on all metrics are significantly better than trained only on a single task, except for the BLEU metric.

4.5. Qualitative Result Analysis

Fig. 2 shows the qualitative analysis for our final model. The first column is one description from scanrefer, and the second column shows the 3D scene together with the corresponding ground truth object with that description. The third column shows our localization result based on the query and the last column shows the caption for the predicted object with the highest overlapping with the ground truth box.

5. Conclusion

In conclusion, the report proposes a novel approach for joint solving of 3D visual grounding and 3D dense captioning tasks. Our method, VL3DNet, utilizes a shared vision-language transformer module and a joint loss function to effectively exploit intermodal relationships between objects and descriptions. The use of SoftGroup as the detection backbone and the task-agnostic vision-language fusion module allow for the two tasks to be trained efficiently with lightweight task-specific heads. The results show that VL3DNet outperforms D3Net [1] by a significant margin, 6% in grounding and 35% in captioning respectively.

Limitation It is also possible to further replace SoftGroup with Mask3D, which promises to perform even better than the current non transformer-based backbone. This project can also be extended by enabling beam search for caption generation. In addition, the generated descriptions may not exhibit a high degree of originality or diversity, but this could be related to the limitations of the training data.

References

[1] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. 1, 3, 4

[2] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16443–16452. IEEE. 1, 3

[3] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-d scans. 1, 3

[4] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3d object localization in RGB-d scans using natural language. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12365, pages 202–221. Springer International Publishing. Series Title: Lecture Notes in Computer Science. 1, 2, 3

[5] Jiaming Chen, Weixin Luo, Xiaolin Wei, Lin Ma, and Wei Zhang. HAM: Hierarchical attention model with high performance for 3d visual grounding. 1

[6] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2908–2917. IEEE. 1

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2

[8] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. 1

[9] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-set point grouping for 3d instance segmentation. 1

[10] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. SoftGroup for 3d instance segmentation on point clouds. 1

[11] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. 1

[12] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. 1

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. 2

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. 2

[15] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-View Transformer for 3D Visual Grounding, April 2022. arXiv:2204.02174 [cs]. 1

[16] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-Language Model for 3D Visual Grounding. 1

[17] 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. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13695, pages 528–545. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science. 1

[18] Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. 1

[19] 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, March 2018. arXiv:1707.07998 [cs]. 2

[20] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs, May 2021. arXiv:2011.15124 [cs]. 2

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, June 2021. arXiv:2102.03334 [cs, stat]. 2

[22] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-Language Transformer and Query Generation for Referring Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16301–16310. IEEE, October 2021. 2