# 3D Visual Question Answering

Leonard Schenk
leonard.schenk@tum.de

Munzer Dwedari
munzer.dwedari@tum.de

## Abstract

*In this work, we introduce a new Seq2Seq architecture for the task of 3D Visual-Question-Answering (3D-VQA) on the ScanQA [3] benchmark. We especially distinguish ourselves from the baseline model ScanQA by not choosing the answer among the collection of answer candidates but by creating a language model to predict the answer word by word. Moreover, we employ attention mechanisms, which provide additional explainability on the scene object proposals and question sequence in the answer module. We also enhance the fusion of both modalities with an additional graph module. Our model outperforms the current baseline on 6 out of 7 benchmark scores*[1]*. Apart from that, we shed light on a problem where models neglect the scene information during the answer prediction.*

## 1. Introduction

Multi-modality tasks have recently seen a significant surge in Deep Learning. The challenge is to create deep learning models that can process multiple modalities as input, fuse them, and use the combined features to create meaningful output. One prominent field combines visual and linguistic information in various tasks, like Image or Scene Captioning [5, 7], Visual Grounding [1, 4, 5, 14], Visual Entailment [6, 11, 17], and Visual Question Answering [2, 9, 16]. While the research field developed mainly on 2D images, the focus has been shifting towards 3D visual input recently. One interesting task is the aforementioned Visual Question Answering (VQA), which uses scene information to answer different questions about objects and their attributes in that scene. This is considered a key task when it comes to interaction between humans and machines.

In March 2022, Azuma et al. [3] published their paper with a new task of 3D Visual Question Answering. Within their paper, the authors create a new dataset based on ScanNet [8] scenes and set up a benchmark with their current model as the baseline. However, it comes with the constraint of predicting the final answer only from a set of previously given answer candidates. In possible use cases e.g. home robots this would mean a considerable limitation since the number of different questions and answers is almost infinite and requires a natural language generation. To accommodate this, we present a model that is equipped with a Seq2Seq [15] architecture for the answer generation. Our contributions are:

- Creating a Seq2Seq model that outperforms the current ScanQA [3] baseline on the benchmark scores and
- Outlying important limitations in the current benchmark scores and baseline model.

## 2. Related Work

**3D Scenes and Natural Language.** Up until recently, the research field of vision and language focused on two-dimensional images as input. In 2020, Chen et al. [4] pioneered the research field on 3D Scenes by introducing the task of Visual Grounding on a newly created dataset called ScanRefer, which uses the scenes from ScanNet [8]. Their dataset sets a foundation for new benchmarks in Visual Grounding and Dense Captioning. Shortly afterwards, Achlioptas et al. [1] introduced a fine-grained version of the task with two new datasets Sr3D and Nr3D, consisting of captions for multiple objects in ScanNet [8] scenes. Following that, numerous papers have achieved new state-of-the-art performance on one or more of the datasets (3DVG-Transformer [21], TransRefer3D [10], SAT [18], LanguageRefer [14], D3Net [5]).

**Visual Question Answering.** In general, the task of VQA is to provide answers to questions that are thematically related to the content of an image or 3D scene. As mentioned before, up until now the research focused on 2D images as the visual part of the task, with the first research being published in 2015 [2] and 2016 [9]. The current state of the art in this field is the general-purpose multimodal model BEiT-3 [16], which performs pre-training on a large amount of data and transfers one multiway transformer model to handle various tasks. In 2022, ScanQA [3] shifted the focus to 3D indoor scenes instead of images.

[1] https://eval.ai/web/challenges/challenge-page/1715/overview

## 3. Data

For the task, we use the newly created ScanQA [3] dataset, which is based on the indoor scenes of ScanNet [8]. The training split consists of 25,563 question-answer pairs on 562 scenes. Validation/test sets have 4675/6149 pairs and 71/97 scenes respectively. Some questions have multiple answers, but for simplicity, we train on the first answer only. As for testing, the authors proposed two sets, one with object ID annotations and one without. Object IDs are used to identify the object type that is referred to in the question. In this regard, we show our results here on the test set w/o object IDs (5.1). Nevertheless, our score values on both test sets are similar when compared to the baseline scores.

## 4. Method

### 4.1. Task

As described earlier, 3D Visual Question Answering is a multimodal task, which receives a 3D Scene and a corresponding textual question as input. After processing and fusing both of the modalities, there are three outputs: an answer, a bounding box around the object referred to in the question and its key object ID.

### 4.2. Model

We choose a model that consists of four modules, which are responsible for the feature extraction on both modalities, the fusion of these features, and generating the answer (fig. 1).

**3D feature extraction.** The input to the 3D feature extraction module is a point cloud $p \in \mathbb{R}^{n_p \times c}$, where $n_p$ is the number of points and $c$ is the number of features for each point. The minimum number of $c$ is 3, meaning only the coordinates for each point are used. For our final model, we train with color, normals, height, and multiview features, so $c$ is of size 135. To encode these features, we use VoteNet [13], which computes object proposals $p' \in \mathbb{R}^{m \times h}$ for each scene. Subsequently, the proposals are further refined concerning their relationships with each other using the transformer architecture 3DVG from Zhao et al. [21]. Together, the proposals represent the scene information in the fusion module. Each object proposal is an $h$-dimensional representation of one of the scene's objects.

**Language feature extraction.** Using pre-trained GloVe [12] embeddings, each question word is encoded into a 300-dimensional word embedding. Next, the embeddings are passed through an encoder LSTM (first part of Seq2Seq [15]). This results in an embedded representation of the question $q' \in \mathbb{R}^{n_q \times h}$, where $n_q$ is the number of words and $h$ is the hidden size for the contextualized feature vectors.

**3D scene and question fusion.** The fusion module is responsible for combining the unimodal features from the previous feature extraction modules into a bimodal representation of the question and the 3D scene. To fuse the modalities we stick to the MCAN [19] module, which is used in ScanQA [3], with an addition of a graph module after the self-guided attention on the object proposals (fig. 1). The graph module consists of multiple node-to-node message passing layers, where each node represents an object proposal and the edges are determined through a KNN graph using the center points coordinates of the object proposals. With the graph module, we encode contextual graph

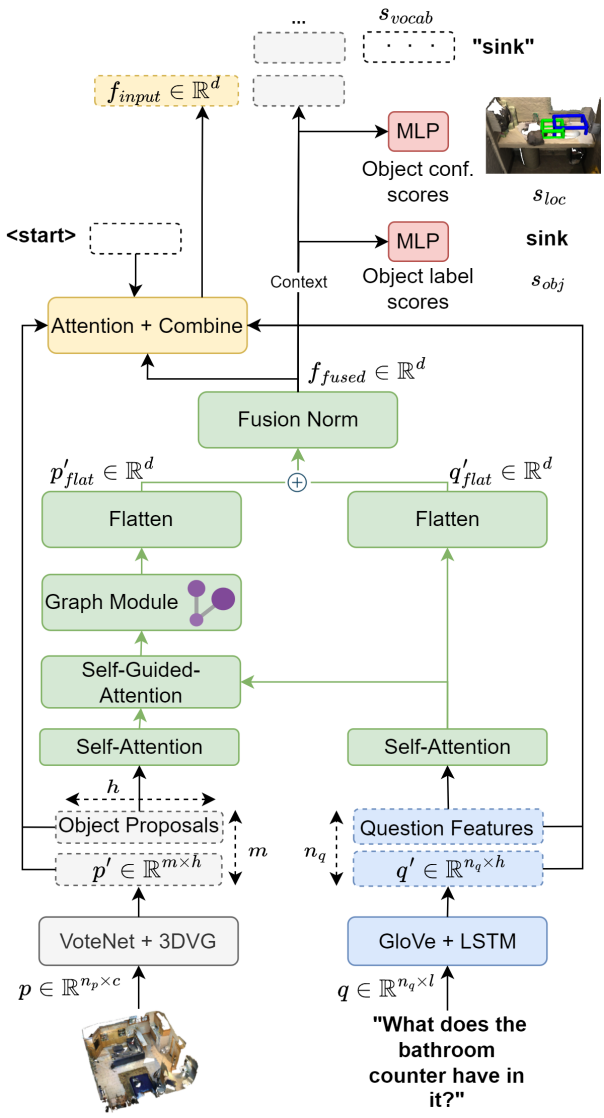


Figure 1. Model architecture. The model can be divided in 4 parts: The grey and blue branches are for the scene and question encodings respectively. The green part is MCAN [19] fusion with the graph module. Finally, at the top (yellow) is the answer module along with 2 MLPs (red) for the 2 subtasks object localization and classification.

information in node embeddings, i.e. object proposals, by iteratively combining neighboring nodes' features. After that, we flatten the object proposals and the question language features into single vectors then add and normalize them with a feed forward network. The output of this MLP Layer is the fused feature vector $f_{fused} \in \mathbb{R}^{512}$. The fused feature vector is used as input to 3 different networks: the answer module and two MLPs to predict the reference object's bounding box and its object ID.

**Answer module.** In ScanQA [3], the model proposed by the authors predicts the answer as a probability distribution over the collection of the training answers. However, inspired by the Seq2Seq model of Sutskever et al. [15], we treat the answer prediction as a sequence generation from the training vocabulary, which gives the model more freedom to generate unique answers that better fit the question than pre-defined answer choices. The output of the modified MCAN [19] fusion module is used as a context vector $f_{fused} \in \mathbb{R}^{512}$ (initial hidden vector) to the GRU decoder (second part of the Seq2Seq [15] model). However, before predicting the next word, we concatenate the context vector and the current word embedding $f_{embed} \in \mathbb{R}^{300}$ (for instance "<start>") and process them in an MLP to calculate the attention weights on the encoded question sequence and object proposals. The attention weights are used to process both modalities' sequences into 2 final vectors $f_{lang}$ and $f_{obj} \in \mathbb{R}^{256}$. We then concatenate $f_{lang}$, $f_{obj}$, and $f_{embed}$, forward them to an MLP and use the output as the input vector $f_{input} \in \mathbb{R}^{512}$ in the GRU. The output of the GRU is processed by another MLP to calculate the probability distribution over the next word, while the new hidden vector is used as the new context vector. With the attention mechanism we allow the model to put focus on important parts of the two different modalities at each step of the answer generation. We discuss the attention mechanism later in section 5.3.

### 4.3. Losses

We keep our final loss similar to ScanRefer [4] (and ScanQA [3]), which is a linear combination of the localization loss $L_{loc}$, the object detection loss $L_{det}$, the object classification loss $L_{obj}$, and the answer loss $L_{ans}$. Unlike ScanQA, our answer loss $L_{ans}$ is a multi-class cross entropy loss over each of the predicted answer words including the "<end>" token:

$$L_{ans} = -\sum_{a \in A} \sum_{v \in V} y_{a,v} log(\hat{y}_{a,v}) \quad (1)$$

where $A$ is the ground truth answer including the "<end>" token and $V$ is the training vocabulary. $y_{a,v}$ has the value 1 when the current ground truth token $a$ matches the vocabulary $v$ and 0 otherwise. $\hat{y}_{a,v}$ is the predicted probability of the token $v$ in the Softmax output for the word $a$.


Figure 2. 9 examples from the validation set including the predicted (blue) and ground truth (green) bounding boxes for the qualitative results. "A" stands for the predicted answer, "GT" for the ground truth.

## 5. Results

### 5.1. Quantitative Analysis

The evaluation of our model compared to the baseline model of ScanQA [3] can be seen in table 1. With our final model, we outperform the current baseline model on 6 out of 7 scores. It is important to mention that we exclude the metric EM@10 (top 10 answer accuracy) since our model predicts one answer only instead of choosing the 10 most probable answer candidates. Furthermore, each model is trained twice: once with and once without the scene information. It can be seen that in this case, the performance of our model decreases more than ScanQA, indicating that our model makes more use of the scene information. In section 5.3, we go into more detail about this finding. Due to simplicity, we only show the scores on the test dataset w/o object annotations. However, results from the test dataset w/ object annotations are similar.

### 5.2. Qualitative Analysis

For the qualitative analysis, we show some interesting examples of the predicted answers and bounding boxes on the validation set (fig. 2). We see for instance in 2) that our model picks a different kitchen cabinet than the ground

| Model | EM@1 | BLEU-1 | BLEU-4 | ROUGE | METEOR | CIDEr | SPICE | avg. % change |
|---|---|---|---|---|---|---|---|---|
| ScanQA | **20.90** | 30.68 | 10.75 | 31.09 | 12.59 | 60.24 | 11.29 | |
| ScanQA (no scene) | 19.45 | 29.02 | **12.04** | 29.34 | 11.70 | 55.84 | 9.94 | ↓ 5.75% |
| Ours (no scene) | 14.62 | 29.00 | 8.82 | 26.01 | 10.51 | 47.01 | 7.5 | ↓ **21.75%** |
| Ours (no graph) | 14.96 | 31.35 | 9.97 | 27.15 | 11.10 | 50.99 | 9.35 | |
| Ours (no attention) | 16.59 | 30.57 | 8.05 | 27.65 | 11.25 | 51.92 | 9.73 | |
| Ours | 19.17 | **35.19** | 10.84 | **32.23** | **13.12** | **61.32** | **11.48** | |

Table 1. Scores on the ScanQA test benchmark w/o objects. For all scores, the higher the better.

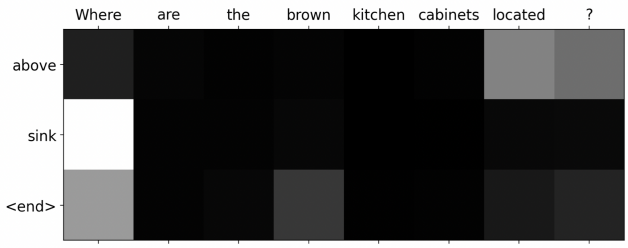


Figure 3. Attention weights for the predicted answer on the question from example 2), figure 2. Note that the lighter the color, the higher the attention value.

truth and gives a correct answer accordingly in the context of the scene. However, when calculating the scores on the benchmark, such answers are considered wrong. Furthermore, we can see that our model predicts the correct answer but chooses a different, but still meaningful bounding box, like the highlighted shelves in 8). Moreover, in 7) we see a case where the model misses both the bounding box and the answer.

As for visualizing the attention weights, we discover that using attention in the answer module improves our overall performance (table 1). However, the attention weights for both modalities are not always as comprehensible as in figure 3, which gives an example for the textual attention. In this instance, we see that when predicting the spacial adverb "above", most attention is put on the question word "located". As for the object "sink" it fully attends on the question word "where". We hint in section 5.4 at this subject matter as a potential field of interest in future works as attention in this case for both modalities can provide better explainability of the predictions.

### 5.3. Ablation Studies

**Does the graph module help?** The usage of graph architectures can be found in several models that deal with 3D spatial understanding [5, 7, 20]. In our case, we use a graph module to encode contextual information into each object proposal from its neighbouring ones. This allows the model to capture relations for better scene understanding. After finding the best position of a graph module in our architecture, we observe that it improves our results on the final test scores as can be seen in table 1.

**Does attention help in the answer module?** To do this ablation study, we simply process the input word embedding $f_{embed} \in R^{300}$ without the attention vectors in an MLP to get the input vector $f_{input} \in R^{512}$ for the GRU. Not only does attention allow us to visualize the focus of the model for each generated word (see section 5.2), it also yields better overall results on the test scores (table 1).

**Does our model learn from the scene?** In the course of the study, it has become apparent that either of the models can learn to answer the questions to a certain extent by just training on the questions (no scene). Hence we train our and ScanQA's models once with and once without scene information by using a fixed dummy array of random values as point cloud input. We note the average percentage change of the scores and see that our model performs worse and thus learns more from the scene information than ScanQA. Still, there is room for further research to improve upon this important issue in the future.

### 5.4. Conclusion

In this project, we tackle the newly proposed 3D Visual Question Answering task. With our Seq2Seq answer module, attention mechanisms, and graph module we outperform the current baseline model on the benchmark scores. During our studies, we find that models of this particular task can suffer from neglecting the scene information and solely use the question to predict the answer. Therefore, we propose an additional training step with dummy point clouds to show how much worse a model performs when only given the question. In that regard, as can be seen in table 1, we perform better than the baseline. Apart from that, we also see that one question may have many distinct possible answers (fig. 2). However, ScanQA includes up to two answers only for most of the questions in their dataset. This can worsen the performance of some models on the final scores of the benchmark. Finally, we look forward to future works on this task to include more expressive attention mechanisms that make the answer predictions more explainable to humans.

## References


[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, pages 422–440. Springer, 2020. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[4] 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. 1, 3

[5] 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. *CoRR*, abs/2112.01551, 2021. 1, 4

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740, 2019. 1

[7] 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, 4

[8] 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

[9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1

[10] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 1

[11] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. *CoRR*, abs/2012.15409, 2020. 1

[12] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[13] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds, 2019. 2

[14] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding, 2021. 1

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. 1, 2, 3

[16] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 1

[17] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706, 2019. 1

[18] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 1

[19] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. *CoRR*, abs/1906.10770, 2019. 2, 3

[20] 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. 4

[21] 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