CLIP-Guided Vision-Language Pre-training for Question Answering in 3D Scenes

Maria Parelli[*] Alexandros Delitzas[*] Nikolas Hars, Georgios Vlassis,
Sotirios Anagnostidis, Gregor Bachmann, Thomas Hofmann
ETH Zurich, Switzerland
{mparelli, adelitzas, nihars, gvlassis, sanagnos, gregorb}@ethz.ch

Abstract

Training models to apply linguistic knowledge and visual concepts from 2D images to 3D world understanding is a promising direction that researchers have only recently started to explore. In this work, we design a novel 3D pre-training Vision-Language method that helps a model learn semantically meaningful and transferable 3D scene point cloud representations. We inject the representational power of the popular CLIP model into our 3D encoder by aligning the encoded 3D scene features with the corresponding 2D image and text embeddings produced by CLIP. To assess our model's 3D world reasoning capability, we evaluate it on the downstream task of 3D Visual Question Answering. Experimental quantitative and qualitative results show that our pre-training method outperforms state-of-the-art works in this task and leads to an interpretable representation of 3D scene features.

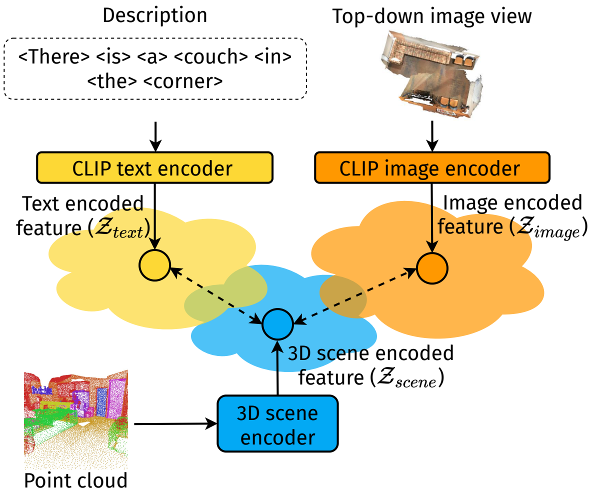

Description
Top-down image view
<There> <is> <a> <couch> <in> <the> <corner>
CLIP text encoder
CLIP image encoder
Text encoded feature ($\mathcal{Z}_{text}$)
Image encoded feature ($\mathcal{Z}_{image}$)
3D scene encoded feature ($\mathcal{Z}_{scene}$)
3D scene encoder
Point cloud

Figure 1. Our pre-training method encourages the alignment of the 3D scene representation to the corresponding text and image embeddings in CLIP space via a cosine similarity loss.

1. Introduction

Humans inherently have a coupled representation of textual and visual structures that is essential to perceiving the world. In recent years, vision and language research has demonstrated significant progress toward enabling models to bridge the semantic gap between the textual and visual modalities. In the 2D domain, many works aim to solve related tasks, such as image captioning [27], Visual Question Answering (VQA) [2] and Visual Commonsense Reasoning (VCR) [29] via aggregating multi-modal information. To this end, many elaborate pre-training techniques [5, 10] have been investigated to encourage fine-grained alignment between words and image regions and increase the robustness of Vision-Language (V-L) architectures. However, the 3D world is characterized by complex inter-object relationships and this restricted form of 2D image supervision limits the usability of such models. In this direction, a new area of research has emerged, whose primary goal is to endow models with 3D spatial reasoning abilities. Some characteristic lines of work are 3D object localization [4], 3D object captioning [25] and embodied question answering [7]. To increase downstream performance in 3D recognition and segmentation benchmarks, some recent pre-training methods employ contrastive losses to transfer 2D visual knowledge to 3D models [1, 13] or align 3D point clouds and voxel representations [30]. However, to our knowledge, no pre-training methods for question answering have been proposed that guide a model to correlate 3D visual input to language cues and corresponding 2D information. In this work, we aim to design a 3D V-L pre-training method to help a model learn language-grounded and semantically meaningful scene object representations, enhancing its performance on downstream 3D scene reasoning tasks. To evaluate our approach, we focus on the 3D Visual Question Answering (3D-VQA) setting, as pre-

[*] Equal contribution.

sented in [3], in which models have to answer questions about 3D scenes, given their RGB-D indoor scan. Inspired by [20], we hypothesize that the high-level semantics from pre-trained 2D visual and linguistic knowledge can benefit 3D scene understanding. We design a transformer-based 3D scene encoder module, which extracts a holistic scene representation by modeling the relations among the scene's object features. Our proposed pre-training method aims to train the scene encoder to project the appearance and geometric features of the 3D scan to an interpretable latent space. This can be achieved by aligning the scene embedding to the corresponding text and image representations extracted by the Contrastive Language-Image Pre-training (CLIP) [18] model. To measure the downstream performance of our approach on the 3D-VQA benchmark, we transfer the weights of the pre-trained 3D scene encoder to a novel 3D Vision-Linguistic architecture that processes the multi-modal representations and fine-tune the model in a supervised manner. Our fine-tuned model outperforms the state-of-the-art model in the ScanQA dataset [3] in both question-answering and referred object localization tasks. We also provide a visualization of the learned 3D scene features after pre-training, demonstrating our model's high-level semantic understanding.

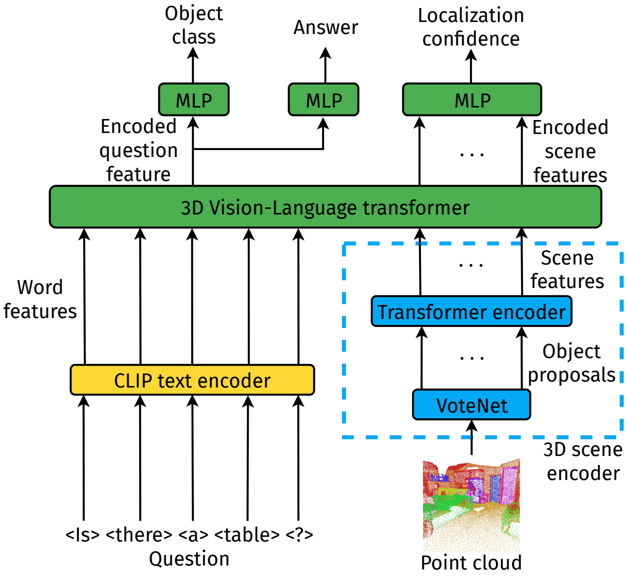


Figure 2. Our 3D-VQA model consists of a 3D scene encoder and a CLIP text encoder. The generated visual and linguistic tokens are fused via a 3D Vision-Language Transformer encoder, which predicts the target answers and localizes the objects referred to by the question.

2. Related Work

2.1. 3D Visual Question Answering

One of the first tasks at the intersection of language and 3D scene understanding was 3D language grounding, in which a model has to localize an object based on a textual description [4, 21]. Building upon this, a novel task has been proposed, namely 3D Visual Question answering. In this problem, a model receives 3D visual information, often in the form of a 3D scene scan and has to answer a perceptual question about the scene. A few approaches have been proposed, such as [3], which develops a new 3D-VQA dataset based on ScanNet [6] scenes and designs a fusion model that jointly processes 3D object and sentence embeddings to predict the correct answer. More recently, Ma et al. [14] introduced SQA3D, a dataset for embodied scene understanding, which requires the agent to understand its 3D location as described by the textual description and reason about its environment.

2.2. 3D Vision Pre-training

2D V-L pre-training has been thoroughly studied [5, 10], pushing the state-of-the-art on V-L benchmarks. To the best of our knowledge, pre-training methods for visual reasoning tasks aiming to jointly model textual, 2D and 3D visual modalities have not been explored. In the 3D domain, current pre-training approaches have focused on learning enhanced 3D scene representations to solve downstream tasks, such as object detection and segmentation [19, 26, 30]. Zhang et al. [30] jointly pre-train point cloud and voxel architectures by using an extension of contrastive learning to multiple data formats. Other works leverage 2D knowledge from large-scale 2D datasets [1, 11, 13]. For instance, Liu et al. [13] map pixel-level and point-level features into the same embedding space via a pixel-to-point contrastive loss.

3. Proposed Method

We propose a pretext task (Fig. 1) that aligns the 3D scene embedding to the corresponding text and image representations in CLIP space [18] via a cosine similarity loss. To demonstrate its effectiveness, we (a) pre-train a 3D scene encoder with this objective and (b) transfer the learned pre-training weights to a novel 3D-VQA model and fine-tune it for the downstream task of 3D-VQA.

3.1. Pre-training Framework Overview

3.1.1 3D Scene Encoder

The 3D scene encoder processes the input scene's 3D point cloud $s \in \mathbb{R}^{N \times 3}$ and generates a holistic representation of the scene objects. It consists of a VoteNet [17] with a pre-trained PointNet++ [30] backbone, which computes k 3D object proposal features $p \in \mathbb{R}^{k \times h}$ of hidden dimension h. As in [3], we set $k = 256$ and $h = 256$. These features are fed to a transformer encoder layer [23] that refines the representations by modeling the global inter-object relations.

3.1.2 3D Scene Encoder Pre-training

The objective of pre-training is to inject the rich 2D visual and linguistic information of CLIP into the model. To this end, we aim to align the 3D scene-level features, extracted by our 3D scene encoder, to the corresponding CLIP text and image embeddings Z_{text} and Z_{image}. We use rendered top-down RGB images of the scene as input to the CLIP image encoder and textual descriptions of the scene for the CLIP text encoder. To obtain the scene embedding, we follow the practice of [8] and append a learnable classification token to the input sequence of object features extracted by VoteNet, whose state at the output of the Transformer encoder serves as the scene representation Z_{scene}. The loss for our pre-training method consists of three terms, the cosine distance $\mathcal{L}_{image}$ between the image and the scene representation, the cosine distance $\mathcal{L}_{text}$ between the text and the scene representation, and the object detection loss $\mathcal{L}_{det}$ of [17], to constrain the scene embedding space to be meaningful for the task of 3D-VQA. Formally, the final loss for the pre-training is defined as $\mathcal{L} = \mathcal{L}_{det} + \alpha\mathcal{L}_{text} + \beta\mathcal{L}_{image}$ where we set $\alpha, \beta = 0.02$.

3.2. Model Architecture for 3D-VQA

Our proposed model for the downstream 3D-VQA task consists of three modules, the pre-trained 3D scene encoder, which processes the 3D scene point features, a CLIP text encoder that extracts the question linguistic representation and a 3D Vision-Language Transformer that fuses the visual and question modalities. The model is tasked with finding the correct answer to the question and accurately localizing the target object related to the answer. Fig. 2 illustrates our 3D-VQA model.

3.2.1 Multi-Modal Fusion Module

To process the question, we use a pre-trained CLIP text encoder and obtain 512-dimensional word-level embeddings. The word and 3D scene features, extracted by the scene encoder, are concatenated and fused via a two-layer transformer encoder that leverages self-attention to simultaneously model intra- and inter-modal relations. The updated scene object features are forwarded to a fully-connected layer that is used for target object localization by determining the likelihood of each object box being related to the question. Following CLIP, we treat the updated EOT embedding (last token in the sequence) of the question as the pooled question feature Q' and use it as input to two linear classifiers. The first one predicts the correct answer by projecting Q' into a vector $a \in \mathbb{R}^n$ for the n answer candidates. The second predicts which objects from 18 ScanNet classes are associated with the question.

3.2.2 Loss Functions

We model the final loss as a linear combination of four terms. We use the object localization loss $\mathcal{L}_{loc}$, as defined in [4], and the object detection loss $\mathcal{L}_{det}$ of VoteNet [17]. To further supervise the training, we include an object classification loss $\mathcal{L}_{obj}$, which is modeled as a multi-class cross-entropy loss and an answer classification loss $\mathcal{L}_{ans}$, which is a binary cross-entropy (BCE) loss function as there are multiple candidate answers. Thus, the total loss is defined as $\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{obj} + \mathcal{L}_{ans} + \mathcal{L}_{loc}$.

4. Experiments

In this section, we evaluate whether transferring the weights of the 3D network pre-trained by our method to the downstream task of 3D-VQA can boost performance compared to training from scratch. We pre-train the 3D scene encoder and fine-tune our 3D-VQA model for the 3D-VQA task in a supervised manner.

4.1. Datasets

We evaluate our approach on the ScanQA dataset [3], which consists of 41,363 diverse question-answer pairs and 3D object localization annotations from 800 3D ScanNet [6] scenes. The ScanQA dataset includes two test sets with and without object annotations. ScanNet is a large-scale annotated dataset of 3D mesh reconstructions of interior spaces. In the pre-training phase, we use the ScanNet annotated point cloud data to render the RGB images with the Open3D software [31]. To obtain the textual descriptions, we use the ScanRefer dataset [4], which contains 51,583 descriptions of 800 ScanNet scenes.

4.2. Implementation Details

We pre-train the 3D scene encoder for 6 epochs with the Adam optimizer using a batch size of 16, a learning rate of 1e-4 and a weight decay of 1e-5. We use the pre-trained weights of the scene encoder and we fine-tune the 3D-VQA network on ScanQA for 25 epochs with an initial learning rate of 5e-4. To mitigate overfitting, we applied rotation about all three axes using a random angle in $[-5°, 5°]$ and randomly translated the point cloud within 0.5 m in all directions. Additionally, we used a random cuboid augmentation, similar to [15], which extracts random cuboids from the input point cloud.

4.3. Evaluation

To measure the downstream performance of our model on 3D-VQA, we report the EM@1 metric, which is the percentage of predictions in which the predicted answer exactly matches any of the ground-truth answers. Following the practice of [3], we also include the widely used sentence evaluation metrics BLEU [16], ROUGE-L [12], ME-

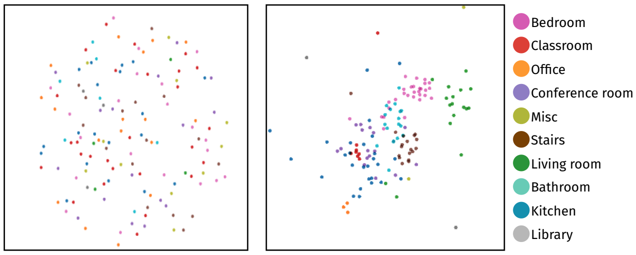
Method	EM@1	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Test set w/ objects						
Scanrefer + MCAN	20.56	27.85	7.46	30.68	11.97	57.36
ScanQA w/o multiview	22.49	30.82	9.66	33.37	13.17	64.55
ScanQA	23.45	31.56	12.04	34.34	13.55	67.29
Ours w/o pre-training	22.76	31.08	13.31	33.84	13.28	65.81
Ours	**23.92**	**32.72**	**14.64**	**35.15**	**13.94**	**69.53**
Test set w/o objects						
Scanrefer + MCAN	19.04	26.98	7.82	28.61	11.38	53.41
ScanQA w/o multiview	20.05	30.84	12.80	30.60	12.66	59.95
ScanQA	20.90	30.68	10.75	31.09	12.59	60.24
Ours w/o pre-training	20.71	31.22	11.49	31.35	12.80	60.75
Ours	**21.37**	**32.70**	**11.73**	**32.41**	**13.28**	**62.83**

Table 1. Comparison of question answering results on the ScanQA test datasets.

Method	Acc@0.25	Acc@0.5
Scanrefer + MCAN	23.53	11.76
ScanQA w/o multiview	25.17	16.21
ScanQA	24.96	15.42
Ours w/o pre-training	26.57	18.58
Ours	**29.61**	**21.22**

Table 2. Comparison of referred object detection results on the ScanQA valid dataset.

TEOR [9] and CIDEr [24]. These metrics are significant for evaluating robust answer matching since some questions have multiple possible answer expressions. To assess the target object localization accuracy, we report the Acc@0.25 and Acc@0.5 metrics, which are the percentage of bounding box predictions that have a higher IoU with the ground truths than the threshold 0.25 and 0.5 respectively. As a baseline, we use the current state-of-the-art method of ScanQA (with and without multiview features) [3]. An additional baseline is ScanRefer + MCAN [28], where a pre-trained ScanRefer [4] model identifies the referred object and the MCAN model is applied to the image surrounding the localized object. We also compare to the performance of our model trained from scratch. The results are compiled in Tab. 1 and Tab. 2. With our pre-training method, we report a significant increase in the question answering metrics and a 3.04% and 2.64% gain in the Acc@0.25 and Acc@0.5 metrics, respectively, compared to the model without pre-training. This validates the effectiveness of our pre-training strategy in both question-answering and referred object localization performance. We also observe that our proposed method improves notably over the ScanQA baseline, even though we do not employ preprocessed multiview image features to achieve a more lightweight pipeline.



(a) Random initialization (b) Our pre-training method

Figure 3. T-SNE visualizations of scene-level features in ScanNet. The 3D scene encoder weights learned during pre-training lead to a structured feature representation space.

4.4. Visualization

We provide the T-SNE [22] visualization of the learned features of the pre-trained 3D scene encoder without fine-tuning in Fig. 3. We observe that semantically similar scenes (i.e., scenes of the same type) cluster nicely together in the embedding space. This highlights the high-level semantic understanding ability acquired by the model when it is trained with rich 2D visual and linguistic information.

5. Conclusion

In this work, we propose a novel V-L pre-training strategy that helps a model learn semantically meaningful 3D scene features by aligning them to the corresponding textual descriptions and rendered 2D images in the CLIP embedding space. Our quantitative and qualitative results on the downstream task of 3D-VQA demonstrate the efficacy of our approach in learning useful 3D scene representations. While we observe that a single top-down view already suffices during pre-training for significant downstream improvements, we believe that incorporating multiple views is a promising direction for future work.

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9902–9912, June 2022. 1, 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 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. 2, 3, 4

[4] 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, 4

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1, 2

[6] 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. 2, 3

[7] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[9] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT@ACL*, 2007. 4

[10] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 1, 2

[11] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 1500–1508, 2022. 2

[12] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. 3

[13] Yueh-Cheng Liu, Yu-Kai Huang, HungYueh Chiang, Hung-Ting Su, Zhe Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2D: Pixel-to-point knowledge transfer for 3D pretraining. *CoRR*, abs/2104.04687, 2021. 1, 2

[14] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes, 2022. 2

[15] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end Transformer model for 3D object detection. In *ICCV*, 2021. 3

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, July 2002. 3

[17] 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 International Conference on Computer Vision*, 2019. 2, 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2

[19] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[20] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. *arXiv preprint arXiv:2203.08063*, 2022. 2

[21] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3D objects. *CoRR*, abs/2107.12514, 2021. 2

[22] Laurens van der Maaten and Geoffrey Hinton. Vialuzing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008. 4

[23] 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. 2

[24] 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 (CVPR)*, 2015. 4

[25] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided Transformer for 3D dense captioning on point clouds. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022. 1

[26] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 2

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. 1

[28] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 4

[29] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[30] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10232–10243, 2021. 1, 2

[31] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 3