SoftCap: Dense Captioning for 3D Scenes with SparseConv

Yunxiang Lu
yunxiang.lu@tum.de

Jiachen Lu
jiachen.lu@tum.de

Abstract

Recent work on dense captioning in 3D have achieved impressive results. Despite developments in this area, the limited performance of object detection constrains the quality of generated captions. To improve this issue, we propose the SoftCap model that uses a SoftGroup based detection backbone. With sparse convolution and soft grouping mechanism, better detection performance and denser object features can be achieved, which enables the later language model to generate more reliable captions. A message passsing graph model and an attention mechanism are used to aggregate object features with relational information. Our method can effecively localize and describe objects in 3D scenes and outperforms the existing baseline method with a significant improvement.

1. Introduction

Recently, the 3D vision-language field has been drawing increasing research interest in bridging 3D scene understanding and natural language processing. However, most existing 3D dense captioning methods such as Scan2Cap [4] use VoteNet [20] as the detection backbone. VoteNet uses PointNet++ [21] to extract seed points and then vote them to object centers for grouping the objects and predicting the 3D bounding box and semantic classs, which means some point information will be lost during the hard grouping process and the noise points near ambiguous objects would affect the detection performance.

To improve this issue, we propose a model using SoftGroup [26] based detection backbone. Our backbone uses U-Net with sparse convolution [9] to extract point features and considers soft semantic scores to perform grouping instead of hard one-hot semantic predictions, which leads to better detection performance on locally ambiguous objects.

By using SoftGroup based detection backbone, the language model is expected to generate more reliable captions with denser object features. In our work, we first adopt "Show and Tell" [25] method with a simple GRU [5] module to show the positive effect of our detection backbone on final captions. Then we apply a relational graph module and a more complex Context-aware attention captioning module (CAC) to show that aggregated object features with relational information contributes to more details in generated captions. In addition, ”REINFORCE with baseline” algorithm [23] is used to further improve the captioning performance.

2. Related work

The task of 3D dense captioning was first introduced by Chen et al. [4] in Scan2Cap. Most prior works [2, 12, 28] learn the multimodal relationships among the objects in the 3D scene and decode them as text outputs. Despite the improved captioning module, their work still suffers from poor quality detections due to the weak detection backbone. Our SoftGroup based backbone performs soft grouping mechanism on point-level and is thus able to predict more precise bounding boxes.

Image captioning has attracted a great deal of interest [7, 11, 14, 24]. Rennie et al. [23] propose a self-critical sequence training method with REINFORCE algorithm for image captioning. Many dense captioning works [13,15,27] are also closely related to our task. In contrast, we work directly on 3D scene input dealing with object attributes as well as 3D spatial relationships, rather than on 2D images.

3. Method

We propose an end-to-end architecture consisting of the following main components: 1) SoftGroup based detection backbone; 2) relational graph; 3) context-aware attention captioning. Figure 1 shows the pipeline of our work.

3.1. Detection Backbone

As input to the detection module, we assume a point cloud P of a scan from ScanNet consisting the coordinates and colors. We adapt SoftGroup-based network [26] to process the input 3D point cloud data and aggregate instance features. The output of this module is the final instances (Bounding boxes) and object features O. Since the grouping process is performed on point level, we can infer the bounding box from refined point set of each cluster. The proposal features aggregated from points also provide denser information. (see Appendix A).

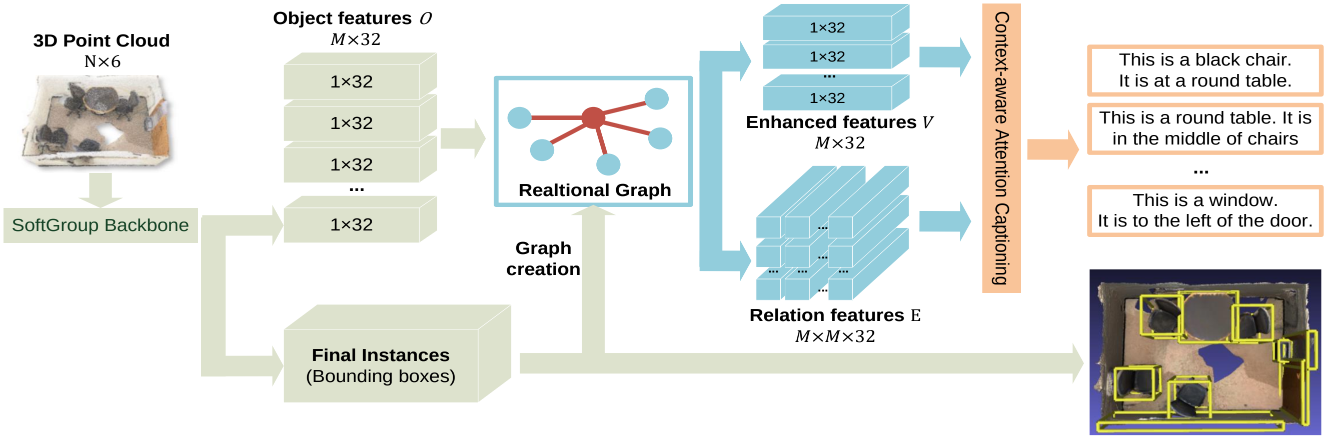
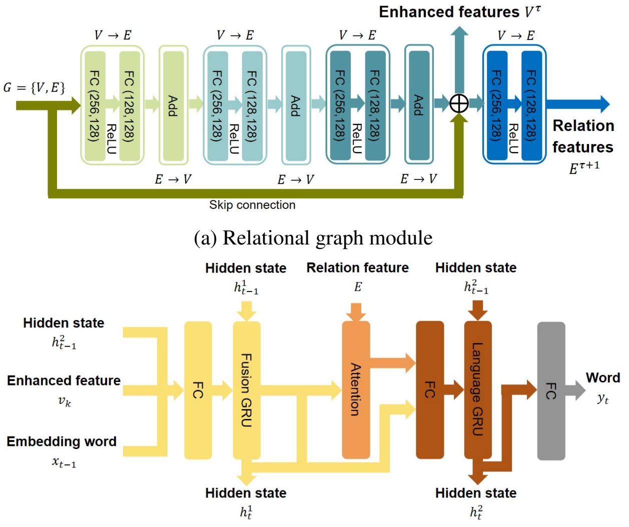


Figure 1. Our network takes the 3D point cloud as input, and a following detection backbone network, also known as SoftGroup, predicts the bounding boxes of final instances D_{bbox} and the object features of each instance O. After that, the relationship graph module generates the enhanced features V and relationship features E based on the object features from the SoftGroup. As the final step, the Context-aware attention captioning module uses the enhanced features and relational features to generate object description tokens.



(a) Relational graph module

(b) Context-aware attention captioning module

Figure 2. (a) Relational graph module generates the enhanced features V^{τ} and the relational features $E^{\tau+1}$ based on the created graph $G = \{V, E\}$ and object features O from the SoftGroup; (b) At time step t, the Context-aware attention captioning module takes the enhanced features v_k, relational features e_k and embedding word x_{t-1} of the target object k as input and then uses them to predict the next token y_i with the help of two GRU modules and attention mechanism.

3.2. Relational Graph

In order to capture relational features, we propose a relational graph module following Scan2Cap [4], also known as $G = \{V, E\}$, which can enhance the object features and capture the relational features between objects [4]. As shown in Figure 2a, we consider each object proposal in the scene as a node of the graph G and construct the graph edges using the relationship between the objects, and we use standard neural message passing [8]. The output of the Relational Graph module is enhanced object features $V \in \mathbb{R}^{M \times 32}$ and object relation features $E \in \mathbb{R}^{M \times M \times 32}$.

3.3. Context-aware Attention Captioning

Inspired by Scan2Cap [4], our language module is composed of two single-layer GRU modules and an attention module, as shown in Figure 2b.

Fusion GRU. At time step t of caption generation, the Fusion GRU layer will take the concatenated feature vectors u_t^1 as the input and output the first hidden state of the current time step h_t^1.

Attention module. The attention module attends not only to the enhanced features of the target object and the embedding word, but also to the relational features of the target object.

Language GRU. After applying the attention mechanism, we use the Language GRU and MLP to predict the token of the current time step y_i based on the attended input features u_t^2.

3.4. Training Objective

Object detection loss. We use the instance segmentation loss introduced in SoftGroup [26] to train the detection backbone. The detection loss is composed of five parts: $L_{det} = L_{sem} + L_{off} + L_{class} + L_{mask} + L_{mask\_score}$. L_{sem} is a cross-entropy loss supervising semantic label prediction for each point. L_{off} is a L_1 regression loss constraining the point shifted to corresponding instance center. L_{class} is a cross-entropy loss supervising the predicted class of each proposal. The binary cross-entropy loss L_{mask} supervises the instance mask within each proposal and the L_2 loss L_{mask_score} supervises the IoU score of the predicted mask with the ground truth.

Relative orientation loss. To stabilize the learning process of the relational graph module, we follow Scan2Cap [4] to apply the relative orientation loss on the message passing module as a proxy loss. The output augular deviations ranging from 0° to 180° degree are discretized into 6 classes and a simple cross-entropy loss L_{ori} supervises the relative orientation predictions.

Cross entropy caption loss. During the training, we use tercher forcing mechanism and predict the next word based on GT word. For each training sample containing a pair of GT bounding box and associated GT description, we optimize the description associated with the predcited bounding box which has the highest IoU score with the current GT bounding box. A conventional cross entropy loss function L_{cap} is applied on the generated token probabilities, as in previous work [14, 24].

CIDEr loss using reinforcement learning. Following prior work [17, 18, 22], generating descriptions is treated as a reinforcement learning task in which the language module acts as the 'agent' and the previously generated words and input visual signal I are the 'environment'. The generated caption $\hat{C} = \{c_1, ..., c_T\}$ is treated as the result of action based on the policy p_θ, which is defined by the weight θ of the network. The optimization goal is to maximize the reward function $R(\hat{C}, I)$. In order to reduce the variance of the loss function, we follow the idea of self-critical sequence training (SCST) [23] to baseline the algorithm with the reward $R(C^*, I)$ of the caption C^* independent of $\hat{C}$. During the training, we apply beam search to sample captions $\hat{C}$ and use greedily decoded inference caption C^* as the baseline. The CIDEr score of the caption defines the reward function R. Since the reward function is not differentiable, we apply policy gradient to get final loss:

$$L_{cap\_rl} \approx -(R(\hat{C}, I) - R(C^*, I)) \sum_{t=1}^{T} \log p(\hat{c}_t | I, \theta) \quad (1)$$

Final loss. We combine all these loss terms in a linear manner as our final loss function. Equation 2, 3 show the final loss function using cross entropy loss and CIDEr loss.

$$L_{XE} = 5L_{det} + L_{ori} + 0.5L_{cap} \quad (2)$$

$$L_{RL} = 5L_{det} + L_{ori} + L_{cap\_rl} \quad (3)$$

3.5. Training and Inference

A stage-wise training strategy is used for stable training. We first pretrain the SoftGroup backbone on all training scans in ScanNet via detection loss L_{det}. Then we train the complete pipeline with the pretrained detection backbone with loss L_{XE}. After convergence, we trian the model further with reinforcement learning mechanism with loss L_{RL}.

Method	Detection	Captioning F1-score C	B-4	M	R	Detection mAP
Scan2Cap [4]	VoteNet	15.71	9.01	7.18	14.92	32.09
X-Trans2Cap [28]	VoteNet	17.64	9.68	7.21	15.25	35.31
MORE [12]	VoteNet	16.46	8.86	7.12	14.71	31.93
Ours (CE)	SoftGroup	30.76	16.30	**13.83**	28.41	57.22
Ours (CIDEr)	SoftGroup	**36.27**	**18.66**	13.82	**29.13**	**57.38**

Table 1. Quantitative results of 3D dense captioning on ScanRefer [3]. All metrics are thresholded by **IoU 0.5**. Our method outperforms all baselines with a remarkable margin.

Network Architecture	C	B-4	M	R	mAP
SoftGroup [26] + GRU [5]	25.69	13.74	12.96	26.88	55.64
SoftGroup [26] + RG [4] + GRU	26.12	14.09	12.74	26.98	55.13
SoftGroup [26] + RG + Att2GRU	26.77	14.81	13.13	27.48	56.48
SoftGroup [26] + RG + CAC [4]	**30.76**	**16.30**	**13.83**	**28.41**	**57.22**

Table 2. Ablation study 1: Comparison of 3D dense captioning results obtained by our method with different components (RG means relational graph) using **CE loss** based on **F1 scores and 0.5IoU**. Clearly, our method with graph module and CAC module is shown to be most effective.

Network Architecture	C	B-4	M	R	mAP
SoftGroup [26] + GRU [5]	33.24	17.45	13.57	28.36	55.80
SoftGroup [26] + RG [4] + GRU	34.12	17.38	13.62	28.61	55.29
SoftGroup [26] + RG + Att2GRU	34.78	17.62	13.46	28.23	56.64
SoftGroup [26] + RG + CAC [4]	**36.27**	**18.66**	**13.82**	**29.13**	**57.38**

Table 3. Ablation study 2: Comparison of 3D dense captioning results obtained by our method with different components (RG means relational graph) using **CIDEr loss** based on **F1 scores and 0.5IoU** . Clearly, our method with graph module and CAC module is shown to be effective. Also, our network performs better captioning when using CIDEr loss than using CE loss.

During the inference, for simplicity we only infer the final instances corresponding to class with highest probability in the refinement stage of detection backbone.

4. Experiments

4.1. Experimental Settings

Datasets. We use the ScanRefer dataset and follow the official ScanRefer benchmark split [3]. All results and analysis are conducted on the val split.

Evaluation Metrics. To jointly measure the quality of generated captions and detected bounding boxes, we evaluate the standard image captioning metrics like CIDEr and BLEU under different Intersectionover-Union (IoU) scores between predicted bounding boxes and the matched ground truth bounding boxes.

We follow the protocal in Unit3D [1] that measures both the captioning precision $M^{\mathrm{P}}@k\mathrm{IoU} = \frac{1}{N^{\mathrm{pred}}} \sum_{i=1}^{N^{\mathrm{pred}}} m_i u_i$ and captioning recall $M^{\mathrm{R}}@k\mathrm{IoU} = \frac{1}{N^{\mathrm{GT}}} \sum_{i=1}^{N^{\mathrm{GT}}} m_i u_i$, where $u_i \in \{0, 1\}$ is set to 1 if the IoU score for the i^{th} box is greater than k, otherwise 0. We use m to represent the captioning metrics. Note that the previous evaluation protocol in Scan2Cap [4] only covers the dense captioning recalls without penalizing false positives. Finally, we combine them to get the final captioning F1-score:

$$M@k\mathrm{IoU} = \frac{2 \times M^{\mathrm{P}}@k\mathrm{IoU} \times M^{\mathrm{R}}@k\mathrm{IoU}}{M^{\mathrm{P}}@k\mathrm{IoU} + M^{\mathrm{R}}@k\mathrm{IoU}} \quad (4)$$

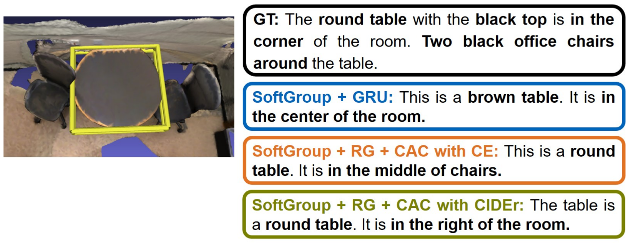
We use mean average precision (mAP) thresholded by IoU as the object detection metric.

Implementation Details. In the pre-training stage, the SoftGroup backbone is trained on ScanNet [6] using Adam optimizer [16]. A relabeling processs is essential before training to adapt for the instance classes of ScanRefer Dataset [3]. Then we use Adam optimizer with learning rate 1e-3 to train the whole pipeline on ScanRefer dataset with batch size 4 for 90k iterations, until convergence. Finally we apply loss L_{RL} to train the model again with Adam and leraning rate 5e-4 for 50k iterations until convergence. We follow [4], [26] to set other hyper-parameters and data augmentation in model. All our experiments are conducted on an RTX 3070 GPU with PyTorch Lightning [19].

4.2. Quantitative Analysis

We compare our method with other 3D dense captioning works. As shown in the Table 1, our method with SoftGroup based detection backbone has a significant improvement on detection performance (mAP) compared to others that use VoteNet for object detection. Furthermore, as shown in the captioning F1 scores of the Table 1, our method also significantly outperforms previous works in captioning performance.

4.3. Qualitative Analysis



GT: The **round table** with the **black top** is **in the corner** of the room. **Two black office chairs around** the table.

SoftGroup + GRU: This is a **brown table**. It is **in the center of the room.**

SoftGroup + RG + CAC with CE: This is a **round table**. It is **in the middle of chairs.**

SoftGroup + RG + CAC with CIDEr: The table is a **round table**. It is **in the right of the room.**

Figure 3. Result visualization for one table

We see from Figure 5 that our SoftCap model is able to generate almost correct descriptions for each object in the scene.

From Figure 3, we see that our SoftCap model, after trained with CE, can infer descriptions containing more relational features, e.g. it is in the middle of chairs, compared to using SoftGroup + GRU inferred: it is in the center of the room. This also proves that the relational graph module and the attention mechanism can indeed capture more information about the relationships between objects. However, after trained with CIDEr loss, we see a decrease in the accuracy of the descriptions of the positions and relationships of the objects. This is because when CIDEr loss is used, the network is forced to generate more similar and simple descriptions to reduce the differences with GT and improve the captioning score, which leads to low discriminability. This is one disadvantage of using CIDEr loss, for which Chen et al. [2] has proposed a corresponding solution.

4.4. Ablation Study

In order to demonstrate the performance of different network architectures and training losses in 3D dense captioning task, we performed a number of ablation studies as shown in Table 2 and 3.

Does the relational graph help? By comparing the first row with the second row in Table 2 and Table 3, namely the network w/o and w/ the relational graph module, we can conclude that the enhanced object features contribute to better captions. To study the influence of relation features, we use Att2GRU language module, which concatenates the attended relation features with enhanced features as input, showing that relation features improve captions further.

Does the context-aware attention captioning help? The second, third, and fourth rows of Table 2 and Table 3 show the performance of using different dense caption generation modules, namely the GRU module, the Att2GRU module, and the context-aware attentino captioning modules (CAC). We can see that SoftCap, which uses the CAC module to generate object descriptions, does have the best performance compared to the other network architectures because of its higher capacity and fusion mechanism.

Does the ”REINFORCE with baseline” help? By comparing the different performances of the same network architecture when using CE loss and CIDEr loss in Table 2 and Table 3, we can conclude that SoftCap has better caption performance when using CIDEr loss.

5. Conclusion

In this work, we propose an end-to-end SoftGroup-based network SoftCap, which addresses the issue of unideal detection backbone in previous 3D dense captioning methods. Also, we apply relational graph module, CAC language module and “REINFORCE with baseline” algorithm to generate final object descriptions. Our network outperforms the previous works in both detection and captioning performance with significant improvement.

References

[1] DaveZhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and AngelX. Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. Nov 2022. 3

[2] DaveZhenyu Chen, Qirui Wu, Matthias Nießner, and AngelX. Chang. D3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. Dec 2021. 1, 4

[3] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in RGB-D scans using natural language. *CoRR*, abs/1912.08830, 2019. 3, 4

[4] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-D scans. *CoRR*, abs/2012.02206, 2020. 1, 2, 3, 4, 6

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. 1, 3

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *CoRR*, abs/1702.04405, 2017. 4

[7] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 677–691, Mar 2017. 1

[8] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017. 2

[9] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. *CoRR*, abs/1711.10275, 2017. 1, 6

[10] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *CoRR*, abs/2004.01658, 2020. 6

[11] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. *Cornell University - arXiv*, Jul 2018. 1

[12] 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. 1, 3

[13] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2016. 1

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 664–676, Mar 2017. 1, 3

[15] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and InSo Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. *Cornell University - arXiv*, Mar 2019. 1

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 4

[17] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. *Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data*, page 353–369. Dec 2017. 3

[18] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. *Cornell University - arXiv*, Mar 2018. 3

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach 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, 2019. 4

[20] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. *CoRR*, abs/1904.09664, 2019. 1

[21] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 1

[22] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *International Conference on Learning Representations*, Dec 2015. 3

[23] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *CoRR*, abs/1612.00563, 2016. 1, 3

[24] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2015. 1, 3

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *CoRR*, abs/1609.06647, 2016. 1

[26] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds, 2022. 1, 2, 3, 4, 6

[27] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. *Cornell University - arXiv*, Nov 2016. 1

[28] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Zhen Li, and Shuguang Cui. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. 1, 3

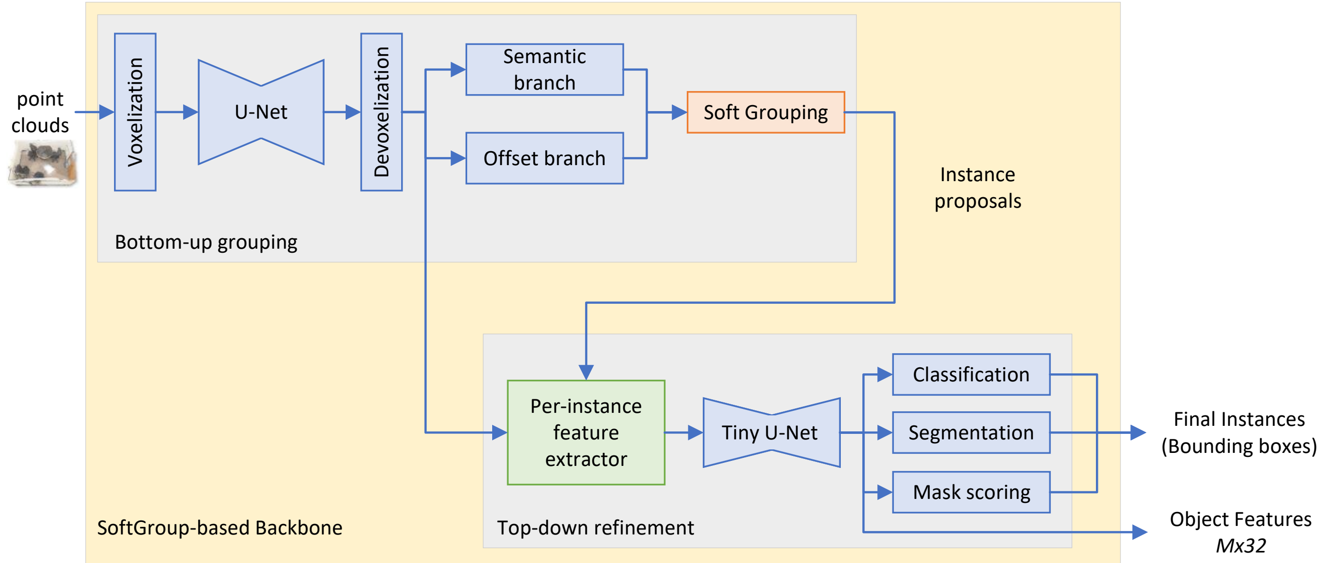


Figure 4. SoftGroup-based detection backbone

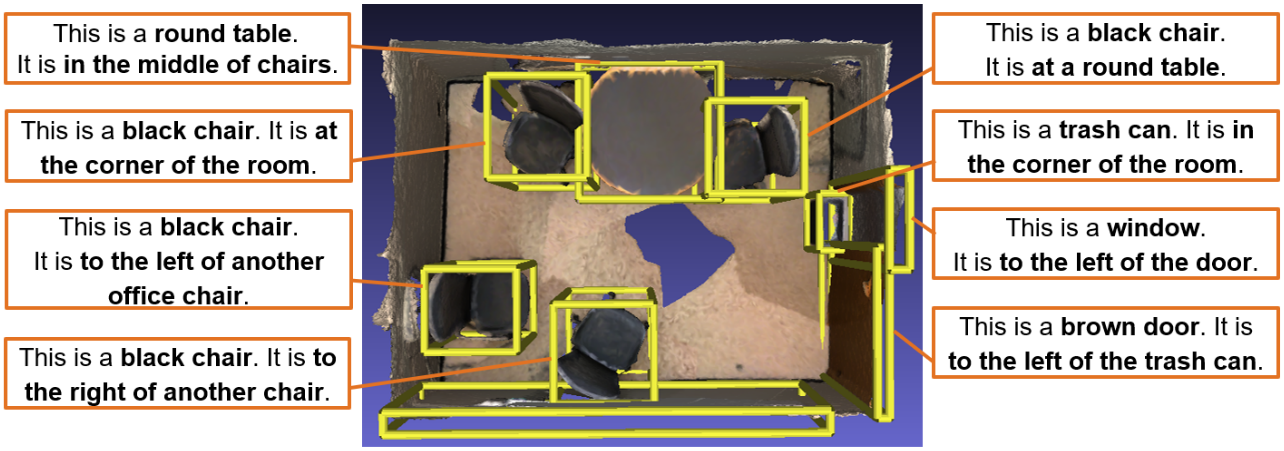

This is a **round table**. It is **in the middle of chairs**.
This is a **black chair**. It is **at the corner of the room**.
This is a **black chair**. It is **to the left of another office chair**.
This is a **black chair**. It is **to the right of another chair**.
This is a **black chair**. It is **at a round table**.
This is a **trash can**. It is **in the corner of the room**.
This is a **window**. It is **to the left of the door**.
This is a **brown door**. It is **to the left of the trash can**.

Figure 5. Result visualization for one scene

A. Supplementary Material

In the supplementary material, we provide additional details about this work.

As shown in Figure 4, the SoftGroup-based network consists of bottom-up grouping and top-down refinement stages.

It first voxelizes the input points and uses a U-Net style backbone with Submanifold Sparse Convolution [9] to obtain enhanced point features. After devoxelization, it performs soft grouping mechanism based on soft semantic prediction score of points to generate instance proposals. The above mentioned network architecture is known as bottom-up grouping and has been used in many point-based grouping works [10]. However, in SoftGroup work [26], this is only the pre-grouping part.

The next part is called the top-down refinement phase. First, the feature extractor module extracts the backbone features from the generated instance proposals. Then, each proposal's features are fed into a tiny U-Net to generate M refined instance features. At the final step, the three MLP branches, i.e., the classification branch, the segmentation branch, and the mask scoring branch, predict the final instance proposal (Bounding boxes).

Exact details and formula explanations about SoftGroup-based network, relational graph module and context-aware attention captioning module can be found in corresponding paper SoftGroup [26] and Scan2Cap [4], which we highly recommend readers to refer to.