SpatioAwareGrouding3D : A Spatio Aware Model For Improving 3D Vision Grouding

Cai Liang, Bo Li, Zhengming Zhou, Longlong Wang, Pengfei He, Liang Hu, Haoxing Wang
Robotics at Xiaomi
{liangcai,libo51,zhouzhengming,wanglonglong1,hepengfei1,huliang8,wanghaoxing1}@xiaomi.com

Abstract

This technical report presents our method, called SpatioAwareGrouding3D, which is a spatio aware model for the Multi-View 3D Visual Grounding track in the Autonomous Grand Challenge at CVPR 2024. Our method is built upon the baseline model [11] and contains three simple yet effective techniques to improve the performance. Firstly, a Visual Language Enhancer Layer is introduced into the model for further fusing the text and visual features. Secondly, a Spatio-aware Decoder is proposed to enhance the spatial reasoning ability of the model, including the designed spatio condition self-attention layer and vision 2d cross attention layer. Finally, we use the ensemble technique to further boost the performance. Experimental results on the Multi-View 3D Visual Grounding track demonstrate that the proposed SpatioAwareGrouding3D achieved 46.92 under the AP@0.25 metric.

1. Introduction

The indoor embodied 3D perception system, which aims to understand the object semantics and scene geometry grounded in language descriptions, is important for the embodied agent [11]. It faces more challenges compared to the system for driving scenes, including the multi-modal input with language instructions, more complex semantic understanding, diverse object categories and orientations, and different perceptual spaces and needs. Based on this, the Multi-View 3D Visual Grounding track in the Autonomous Grand Challenge is proposed, which is built upon the EmbodiedScan[11], a holistic multi-modal, ego-centric 3D perception suite. In this track, given language prompts describing specific objects, models are required to detect them in the scene and predict their oriented 3D bounding boxes.

It is noted that, the text-aware feature extracting and spatial reasoning abilities of the model are important for the model to tackle the visual grounding task. Thus, a novel model called SpatioAwareGrouding3D is proposed, which contains a Visual language Enhancer Layer and a Spatio-aware Decoder to understand the scenes better. Moreover, a ensemble technique is introduced to further improve the performance of the model.

2. Methodology

In this section, we firstly introduce the architecture of the proposed SpatioAwareGrouding3D, including the details of the Visual Language Enhancer Layer and the Spatio-aware Decoder. Then, we introduce the strategy of the ensemble technique.

2.1. Overall architecture

As shown in Fig. 1, SpatioAwareGrouding3D is built upon the Embodied Perception Model [11] and contains five modules: the image backbone, the point cloud backbone, the text encoder, the enhancer fuse module and the spatio-aware decoder. Given the multi-modal inputs, the image backbone is to extract vision 2D features from the multi-view images, while the point cloud backbone pursues vision 3d features from the input point cloud. The text encoder is designed for generating the text features from the language prompts. Then, the multi-modal features are gradually fused by the enhancer fuse module, which is consisted of the sparse fuse module, the neck 3d module and the Visual Language (VL) Enhancer Layer. Finally, the features updated by the enhancer fuse module are given to the spatio-aware decoder, which outputs the final prediction results.

2.2. Visual Language Enhancer Layer

Considering that the vanilla fuse module in the baseline only takes the vision 2d and point cloud features, we insert the Visual Language Enhancer Layer proposed in Grounding Dino [8] into the module, which is helpful for aligning features of the vision and text. As shown in Fig. 2, it takes the vision 3d features generated by the neck3d and the text feature produced by the text encoder as the input, adopting two cross attention layers for cross-modality fea-

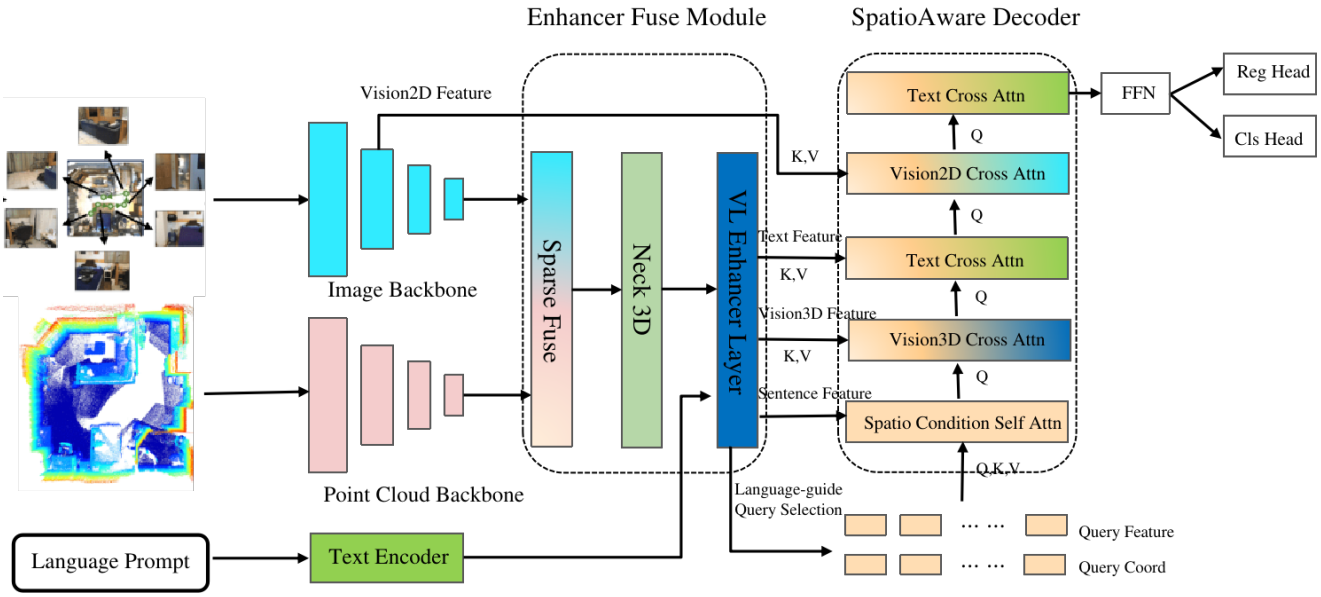

Figure 1. overview of SpatioAwareGrouding3D.

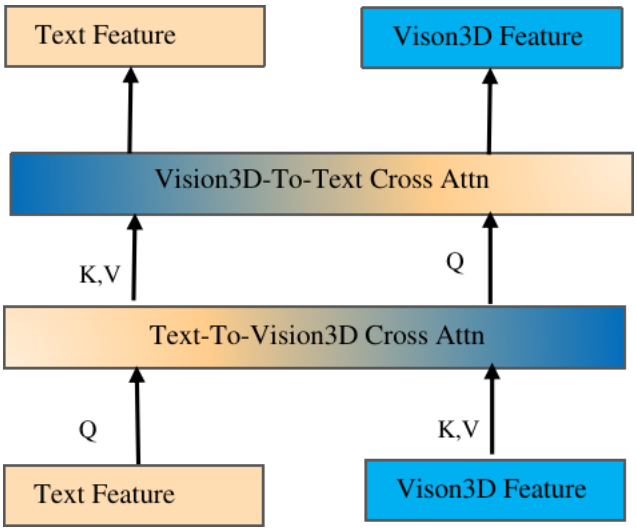

Figure 2. Visual Language Enhancer Layer.

ture fusion. Specifically, The Text-To-Vision3D Cross attention and Vision3d-to-Text cross attention are both vanilla attention [10]. Since we found that the text and vision self-attention in the original Visual Language Enhancer Layer almost have no effect in our experiments, they are removed for reducing the GPU Memory cost.

2.3. Spatio-aware Decoder

The Spatio-aware Decoder is to predict the final oriented 3D bounding boxes from the enhanced multi-modal features, which contains the Language-guide query selection layer and several attention layers. Please see the Embodied Perception Model [11] for more details. To dig the spatial information in vision features more effectively, we introduce the spatio condition self-attention layer and the vision2d cross attention layer in each decoder layer as shown in Fig. 1.

2.3.1 Spatial Condition self-attention

The spatial condition self-attention layer proposed in [2] is helpful for understanding and reasoning about spatial relations, which explicitly considers the relative spatial location in the attention map. It is used to replace the vanilla self-attention after the Language-guide query selection layer and takes the query features $\{Q_1, Q_2, ..., Q_n\}$, the query coordinates $\{O_1, O_2, ..., O_n\}$ and the sentence feature s_{cls} as the input.

Firstly, the spatial distance d_{ij} between each pair of queries $\{Q_i, Q_j\}$ is computed with their coordinates $\{O_i, O_j\}$ and used to generate the pairwise spatial feature f_{ij}^s as:

$$f_{ij}^s = [d_{ij}, norm(d_{ij})], \tag{1}$$

where $[\cdot]$ denotes the concatenation operation and $norm(\cdot)$ is the normalization operation. Then, the language conditioned weight g_i^s is generated for each query, which could be formulated as:

$$g_i^s = W_S^\top (s_{cls} + Q_i), \tag{2}$$

where $W_S \in \mathbb{R}^{d\times 2}$ is a learnable parameter and the bias term is omitted for simplicity. The spatial relevance of (Q_i, Q_j) is defined by fusing g_i^s and f_{ij}^s, which could be formulated as:

$$\omega_{ij}^s = g_i^s \cdot f_{ij}^s \cdot g_j^s. \tag{3}$$

Finally, the spatial condition attention map is calculated with the sigmoid softmax (SigSoftmax) fusion function:

$$\omega_{ij} = \frac{\sigma(\omega_{ij}^s)\exp(w_{ij}^o)}{\sum_{l=1}^{N}\sigma(\omega_{il}^s)\exp(w_{il}^o)}, \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function and w_{ij}^{o} is the vanilla attention weight.

Algorithm 1 Model Ensemble

Input: Ensemble model Box Prediction $B_1, B_2, ..., B_n$, Ensemble model score prediction $S_1, S_2, ..., S_n$, Ensemble weights $W_1, W_2, ..., W_n$, overlap thresh T.
Output: Ensemble model Predictions R
procedure ENSEMBLE(n)
    $bboxes, scores, res = [], [], []$
    for $i = 1$ to n **do**
        $scores.append(S_i)$
        $bboxes.append(B_i)$
    end for
    $idxs = Argsort(scores)$
    for $i = 1$ to $Len(idxs)$ **do**
        for $j = 1$ to $Len(idxs)$ **do**
            if $Overlap(B_i, B_j) >= T$ **then**
                $B_i = Average(W_i * B_i, W_j * B_j)$
            end if
        end for
        $res.append(B_i)$
    end for
    return res
end procedure
R = ENSEMBLE(B, S, W, T)
return R

2.3.2 Vision2d Cross Attention

Considering that the baseline model simply projects the points to vision 2D features by camera intrinsics and extrinsics for sampling and aggregating the image features into the vision 3D features, it might loses some high level semantic information potentially. Addressing this problem, we introduce a vision 2D cross attention layer before the text cross attention layer in the Spatio-aware Decoder. Inspired by SparseBev [7] and Deformable Attention [12], the vision 2D cross attention layer firstly generates a set of sampling offsets $\{(\Delta x_i, \Delta y_i, \Delta z_i)\}$ from the input query feature. These offsets are transformed to 3D sampling points based on the query coordinate (x_i, y_i, z_i):

$$\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = \begin{bmatrix} \Delta x_i \\ \Delta y_i \\ \Delta z_i \end{bmatrix} + \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{5}$$

Then, the sampling points are projected into the image features for aggregating them into the corresponding query by the cross attention mechanism.

2.4. Model Ensemble

The ensemble technique combines several individual models to obtain better performance [4]. Considering that the positions of the predicted boxes generated by a single model might be inaccurate, we average some overlapped 3D boxes from different models for achieving better performances under the AP@0.25 and AP@0.5 metrics. The details of the ensemble strategy are shown in Alg. 1

Table 1. Multi-view 3D visual grounding benchmark. the experiment was evaluated on mini validate dataset

Methods	Input	Epoch	AP_{25}	AP_{50}
Embodied Perception [11]	RGB-D	12	33.59	14.40
Ours	RGB-D	24	**39.00**	**15.43**

3. Experiments

3.1. Dataset

We use EmbodiedScan [11] to train the proposed model, which is a multi-modal, ego-centric 3D perception dataset, and benchmark for holistic 3D scene understanding. It consists of ScanNet[3], 3RScan[6] and Matterport3D[1] datasets. It encompasses over 5k scans encapsulating 1M ego-centric RGB-D views, 1M language prompts, 160k 3D-oriented boxes spanning over 760 categories, some of which partially align with LVIS, and dense semantic occupancy with 80 common categories[11].

The Multi-View 3D Visual Grounding challenge support two different scale datasets: the mini scale and the full scale. Due to the limited of GPU resources, only the mini scale dataset is used for training.

3.2. Implementation Details

We use Adam optimizer with cosine annealing policy to train the model, while the max learning rate is set to 5×10^{-4}, with 0.0005 weight decay. The model is trained on 8 NVIDIA A100 GPUs with 24 epochs. Follow Embodied Perception Model [11], the ResNet50[5] and MinkResNet34[9] are used as the multi-view and point cloud backbone. The number of visual language enhancer layer is set to 2, and the sampling offset points in the vision2d cross attention layer is set to 3 in order to reduce the training time.

3.3. Comparative evaluation

We evaluate the proposed SpatioAwareGrouding3D in the validation set in comparison to the baseline model [11] and the results are reported in Tab.1. It can be seen that SpatioAwareGrouding3D outperforms the baseline model and achieves 39.00 under the AP_{25} metric.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 3

[2] Shizhe Chen, Makarand Tapaswi, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, 2022. 2

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

[4] Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[6] Nassir Navab Federico Tombari Matthias Niessner Johanna Wald, Armen Avetisyan. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[7] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 3

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[9] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022. 3

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[11] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3