LoA-Trans: Enhancing Visual Grounding by Location-Aware Transformers

Ziling Huang[1,2] and Shin'ichi Satoh[2,1]

[1] The University of Tokyo, Tokyo, Japan
[2] National Institute of Informatics, Tokyo, Japan
{huangziling,satoh}@nii.ac.jp

1 Segmentation Heads

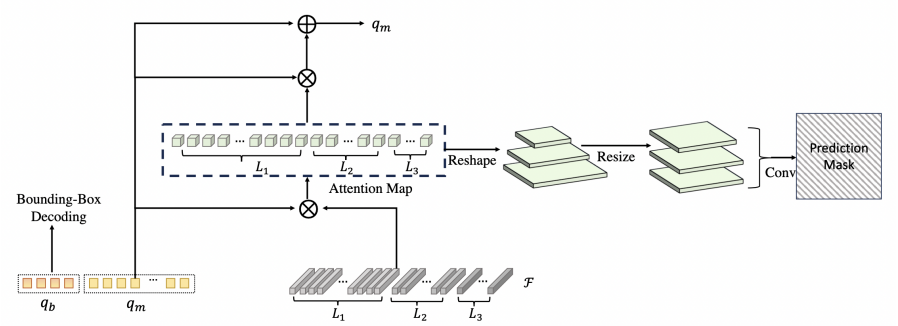

Fig. 1: The architecture of proposed segmentation head.

Fig. 1 shows the architecture of segmentation head in our work. We engage in cross-attention between mask queries and the multi-level multi-modality features $\mathcal{F}$. The cross-attention is performed several layers. The attention maps $\mathcal{A}$ of final layer are divided into segments $\mathcal{A}_i$, each matching the resolution of $\mathcal{F}_i$. Following this, all segments of the attention maps are upscaled to a resolution of $H/8 \times W/8$. Finally, these upscaled maps are concatenated along the channel dimension and project with convolution for final output. This process ensures that the mask prediction is informed by a comprehensive view of the features across different levels, enhancing the accuracy of the mask estimation. In our work, the layer of segmentation head is 2.

2 Limitation and Future Work

Our model currently faces two main limitations: ***Small target.*** The system estimates the center prompt on a final feature map of only 20×20 resolution in input size 640×640, which may not capture small objects effectively, leading to decreased performance. To

address this, we plan to incorporate larger feature maps in future center prompt estimations. ***Single target constraint.*** We strictly assume that each expression corresponds to only one object per image. This assumption is imposed by the major benchmarks such as Refcoco, Refcoco+, and Refcocog. We understand that this scenario does not accurately represent more complex real-world contexts. Moving forward, we aim to adapt our approach to the Generalized Referring Expression Segmentation task [1]. This adaptation will involve enhancing the Location-Aware Network not only to estimate the center prompt but also to ascertain the presence of the object in the image, broadening the applicability and effectiveness of our model.

3 Experiments

3.1 Model Size and Input Resolution

Table 1 shows model size and input resolution, calculated from the released code of each paper.

Model	M.S.	I.R.	Model	M.S.	I.R.
TransVG	149.8M	640	LAVT	218.9M	480
QRNet	273.3M	640	ReLA	225.5M	480
SeqTR	212.0M	640	SeqTR	212.0M	640
Ours-S	204.8M	640	Ours-S	204.8M	640

Table 1: I.R.=Image Resolution, M.S. = Model Size.

3.2 Center Prediction Analysis

As shown in Figure 2 and Figure 3, more details can be found in paper.

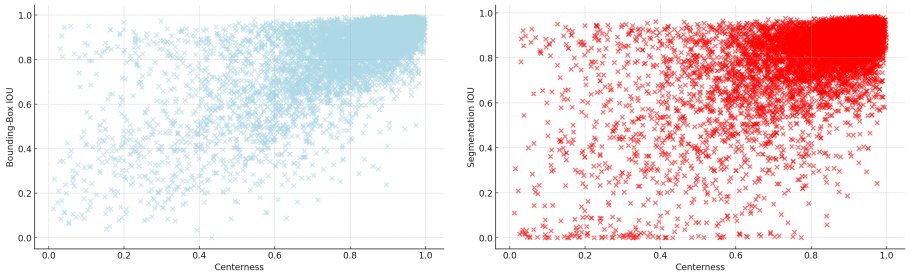

Fig. 2: Left:Correlation Between Targeting Quality and Bounding Box IoU. Left: Correlation Between Targeting Quality and Segmentation IoU. One point denotes one test sample. NOTE: This figure not include the case where l^*, r^*, t^* and b^* is smaller than 0. It means center outside bounding-box.

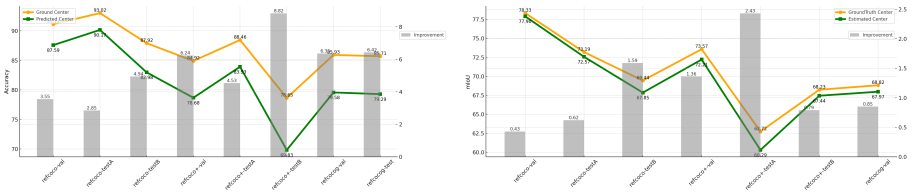

Fig. 3: Left:Comparison and Improvement Analysis of bounding-box results for GroundTruth Center vs. Estimated Center. Right:Comparison and Improvement Analysis of segmentation results for GroundTruth Center vs. Estimated Center.

References

1. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR (2023)