CoupAlign: Coupling Word-Pixel with Sentence-Mask Alignments for Referring Image Segmentation

Anonymous Author(s)
Affiliation
Address
email

1 Ablation Study

Effect of Backbones. To demonstrate the effectiveness of our approach, we change the image backbone of CoupAlign to different networks, like Resnet101 [3] and Darknet53 [7], and evaluate it on RefCOCO validation set. In Tab. 1, we compare our results with the methods using Resnet101 as image backbone. In Tab.2, we compare the methods using Darknet53. The results show that CoupAlign still suppresses previous methods when using the same image backbone, which indicates that our CoupAlign is compatible with popular backbones.

Methed	oIoU	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9
LSCM-Resnet101 [5]	61.47	73.95	69.58	62.59	49.61	20.63
CMPC-Resnet101 [4]	61.36	71.27	64.44	55.03	39.28	12.89
EFN-Resnet101 [2]	62.76	73.95	69.58	62.59	49.61	20.63
LAVT [8]-Resnet101	68.10	80.55	75.25	67.27	55.05	25.37
CoupAlign-Resnet101 (Ours)	**68.93**	**81.26**	**76.54**	**69.69**	**57.07**	**27.44**

Table 1: Comparison with different methods using Resnet101.

Methed	oIoU	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9
LTS-Darknet53 [6]	65.43	75.16	69.51	60.74	45.17	14.41
VLT-Darknet53 [1]	65.65	76.20	-	-	-	-
CoupAlign-Darknet53 (ours)	**68.16**	**81.02**	**76.35**	**69.53**	**56.91**	**25.95**

Table 2: Comparison with different methods using Darknet53.

Effect of mask number N. We further investigate the impact of the number of segments N on model performance. In Tab.3, the first line represents the results of model without sentence-mask alignment. The results indicate that 1) the effect of the presence or absence of SMA on model performance is greater than the effect of the number of segments on model performance. No sentence-mask alignment degrades model performance by about 1.6% on mIoU, while the impact of number of segments on mIoU is around 1%. 2) When N is 100, we obtain the best performance. If N is too large or too small, the performance of the model will degrade.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Num of segments	mIoU	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9
0	73.85	84.86	80.85	74.62	62.29	29.25
20	74.77	85.46	81.81	76.22	64.83	30.91
50	74.98	85.67	81.76	76.20	64.86	31.88
100	**75.49**	**86.40**	**83.41**	**77.59**	**65.94**	**32.40**
150	74.12	85.33	81.71	75.86	64.74	31.30
200	74.38	84.91	81.05	75.36	64.23	31.67

Table 3: Ablation of different mask numbers.

2 Reproducibility

Our code is reproducible and can be implemented based on several open-source repositories[1][2][3][4][5] following the method description in Section 3 as well as implementation details in Section 4.

3 Code Release

According to the code authorization rule of our institution, we are not allowed to directly attach the code in the anonymous submission. We will apply for a code release license after our paper is accepted.

4 Potential Negative Societal Impacts

Our method has no ethical risk on dataset usage and privacy violation because all the datasets and tools are publicly available and transparent.

5 Limitations

Although CoupAlign achieves remarkable performance on referring image segmentation, it still does not segment finely enough in some scenes. As shown in Fig. 1, when the boundary of the object is blurred, or existing occlusion, the segmentation prediction is roughly accurate, but not precise enough. We believe that by adding the boundary enhancement module in the future, our method will perform better in these scenarios.

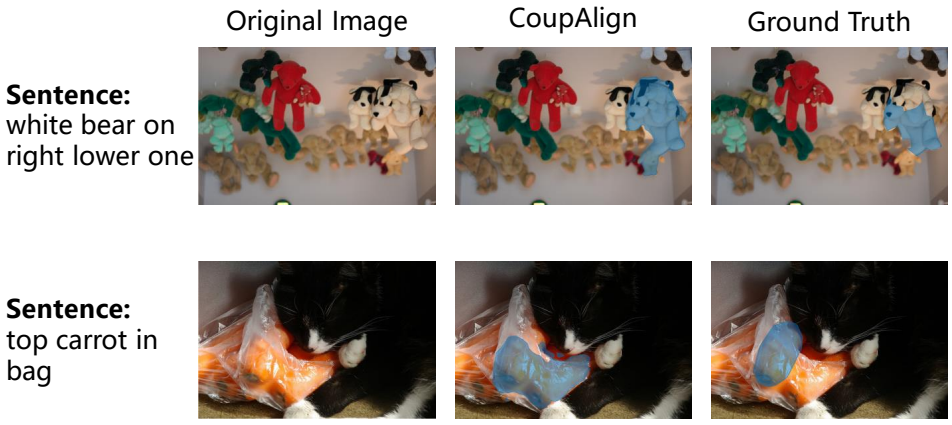

Figure 1: Visualization failure cases of CoupAlign.

[1] https://github.com/open-mmlab/mmdetection
[2] https://github.com/microsoft/Swin-Transformer
[3] https://github.com/huggingface/transformers
[4] https://github.com/pytorch/vision
[5] https://github.com/zichengsaber/LAVT-pytorch

6 Visualization

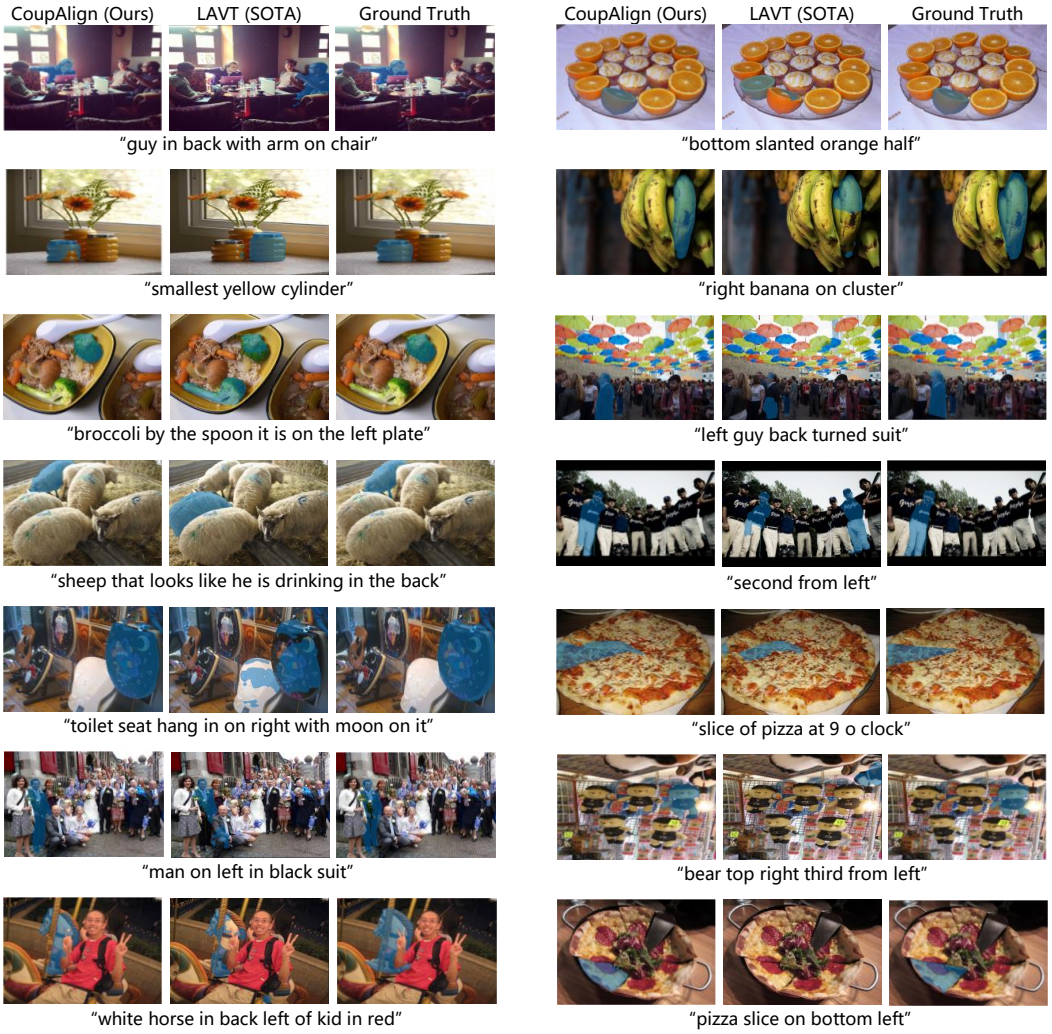

Figure 2: Visualization examples of the correct predictions of CoupAlign while LAVT typically fails.

References

[1] H. Ding, C. Liu, S. Wang, and X. Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[2] G. Feng, Z. Hu, L. Zhang, and H. Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10488–10497, 2020.

[5] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han. Linguistic structure guided context modeling for referring image segmentation. In *European Conference on Computer Vision*, pages 59–75. Springer, 2020.

[6] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021.

[7] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[8] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr. Lavt: Language-aware vision transformer for referring image segmentation. *arXiv preprint arXiv:2112.02244*, 2021.