Mask Grounding for Referring Image Segmentation

Supplementary Material

1. Training and Evaluation Datasets

We evaluate MagNet on the three standard benchmark datasets: RefCOCO [10], RefCOCO+ [10], and G-Ref [6, 7]. RefCOCO and RefCOCO+ contain 19,994 and 19,992 images, respectively, with 142,209 and 141,564 annotated expressions that refer to 50,000 and 49,856 object regions. In contrast, G-Ref contains 26,711 images with 104,560 annotated expressions that refer to 54,822 object regions. Expressions in these datasets are concise with an average length of 3.5 words. However, RefCOCO+ is harder as it disallows location words. G-Ref is the hardest as it contain longer (an average length of 8.4 words) and more complex expressions. For G-Ref, we validate our method on both the UMD [7] and the Google [6] partition.

2. Evaluation Metrics

We assess our proposed method using three commonly used metrics: overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision values at 0.5, 0.7, and 0.9 IoU threshold levels. The oIoU metric is defined as the ratio of the total intersection area over the total union area between the predictions and ground truth of all test data. This metric is sensitive to the segmentation accuracy of large objects. The mIoU metric calculates the average IoU of all test data. This metric treats large and small objects equally. Lastly, the precision metric measures the percentage of test samples that passeses through an IoU threshold. All these metrics collectively provide a comprehensive evaluation of our proposed method.

3. Implementation Details

MagNet is built on top of LAVT and has a total computational cost of 159.14 GFLOPs and a FPS of 3.8 (on RTX 4090). We adopt Swin-Base [4] with patch size 4×4 and window size 12×12 as our image encoder, and BERT-Base [2] with hidden dimension 768 as our language encoder. We initialize Swin-Base with the official classification weights pretrained on ImageNet22K [3] and BERT-Base with the *bert-base-uncased* weights from HuggingFace's Transformer library [8]. The rest of the weights in our model are randomly initialized. We use exactly the same pixel decoder and Transformer decoder implemented in the open-source Mask2Former [1] codebase without any modification. Our pixel decoder has 4 layers and hidden dimension of 256. Our Transformer decoder has 9 layers and 1 object query as there is only 1 ground truth mask for each image in RIS. Following LAVT [9], we adopt the AdamW [5] optimizer with weight decay 0.01 and initial learning rate 5e-5. The learning rate is reduced using the polynomial learning rate decay. Our model is trained for 40 epochs using batch size of 32. All input images are resized to 480×480 without using any data augmentation and all text tokens are zero padded to a maximum length of 20 tokens. For Mask Grounding, we use a corruption probability of 0.15 and a masked token predictor with 8 BERT layers. For cross-modal alignment module, we use 4 different average pooling layers to pool the input feature map to output sizes of 1×1, 2×2, 3×3 and 6×6. For cross-modal alignment loss, we set $\tau_1 = 0.1$ and $\tau_2 = 0.3$.

4. Limitations

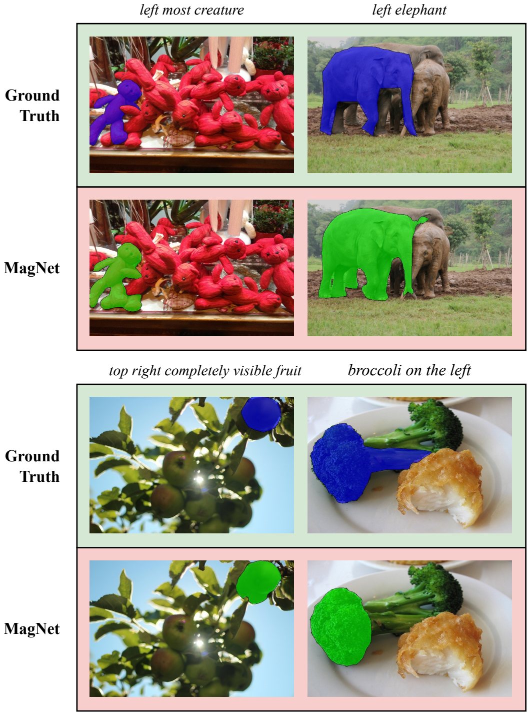

Figure 1. Visualization of MagNet's failure cases.

Although MagNet achieves impressive overall performance for referring image segmentation, its segmentation accuracy at the object boundaries still has some room for improvements. As shown in Fig. 1, when the objects are occluded or have unclear boundaries, MagNet's predicted masks might miss some target pixels or spill over to other objects. This problem can be attributed to a lack of boundary supervision and relatively low mask annotation quality in the training datasets. We anticipate future work to further improve the performance of MagNet by tackling these issues.

5. How Mask Grounding can be used in any existing RIS method?

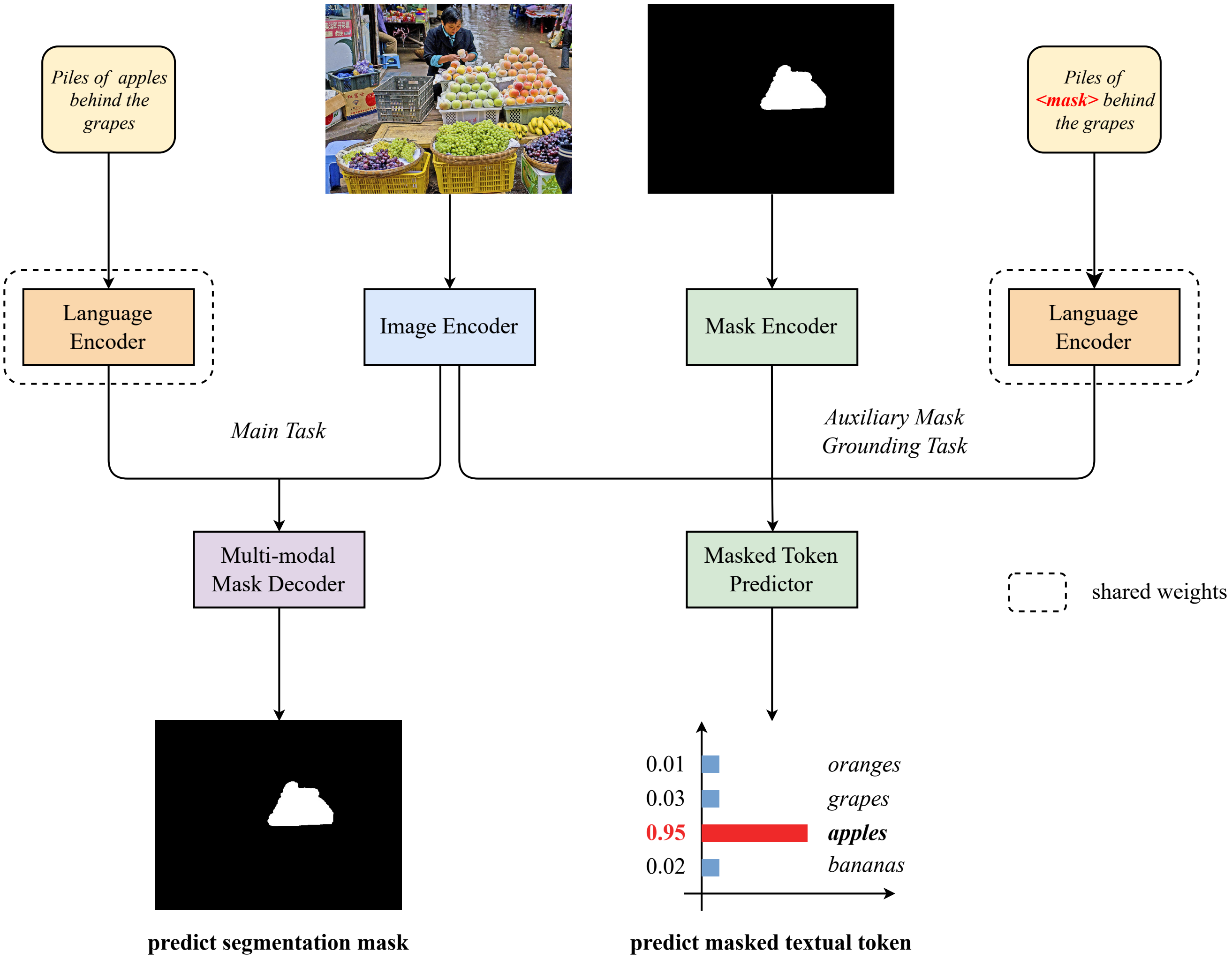

Figure 2. Mask Grounding can be used as an auxiliary task in any existing RIS method to boost performance by improving fine-grained visual grounding. The main RIS task is to predict the segmentation mask based on referring expression, whereas the auxiliary Mask Grounding task is to predict the identity of the masked textual token. In order to perform Mask Grounding, we only need to add a light-weight Mask Encoder to encode the segmentation mask and a light-weight Masked Token Predictor to predict the masked textual token based on the visual, textual and segmentation information.

6. More Visualizations

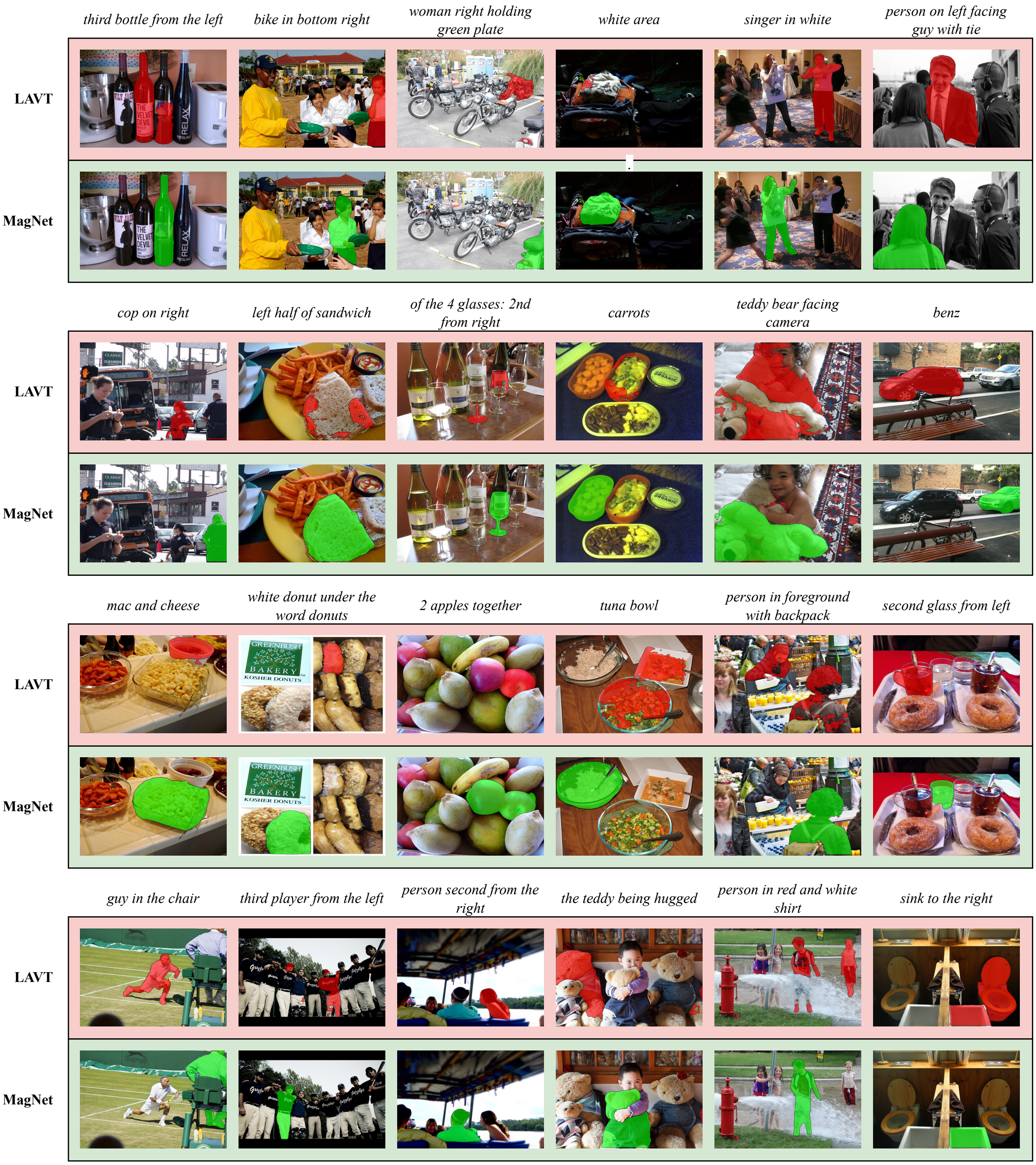
third bottle from the left
bike in bottom right
woman right holding green plate
white area
singer in white
person on left facing guy with tie
LAVT
MagNet
cop on right
left half of sandwich
of the 4 glasses: 2nd from right
carrots
teddy bear facing camera
benz
LAVT
MagNet
mac and cheese
white donut under the word donuts
2 apples together
tuna bowl
person in foreground with backpack
second glass from left
LAVT
MagNet
guy in the chair
third player from the left
person second from the right
the teddy being hugged
person in red and white shirt
sink to the right
LAVT
MagNet

Figure 3. More visualizations of MagNet predictions on RefCOCO.

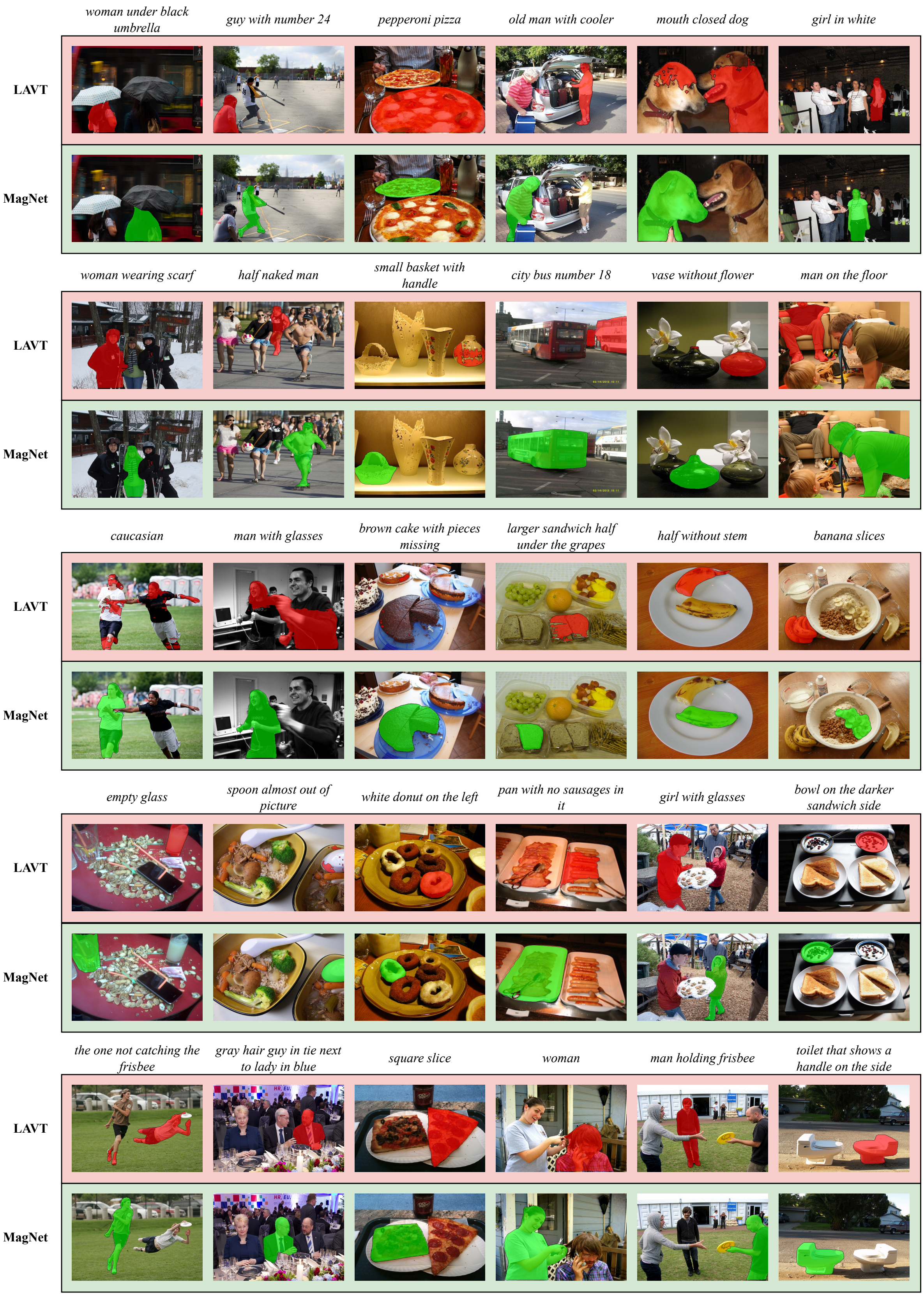

Figure 4. More visualizations of MagNet predictions on RefCOCO+.

References

[1] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NACL: HLT*, 2019. 1

[3] D. Jia, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[5] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[6] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1

[7] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 1

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *ACL*, 2020. 1

[9] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 1

[10] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1