Image-Text Co-Decomposition for Text-Supervised Semantic Segmentation

Supplementary Material

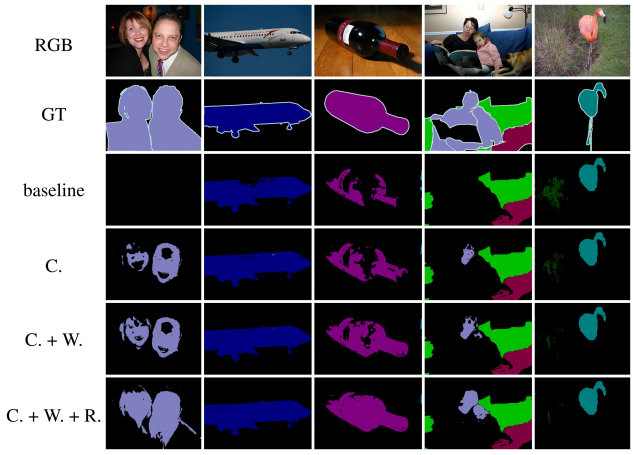

Figure 5. **Ablation studies.** We improve the baseline model by incrementally including (C.) the image-text co-decomposition module, (W.) the word highlighting prompt, and (R.) the region highlighting prompt. We present the segmentation results of the resulting models on the images of the PASCAL VOC [13] dataset.

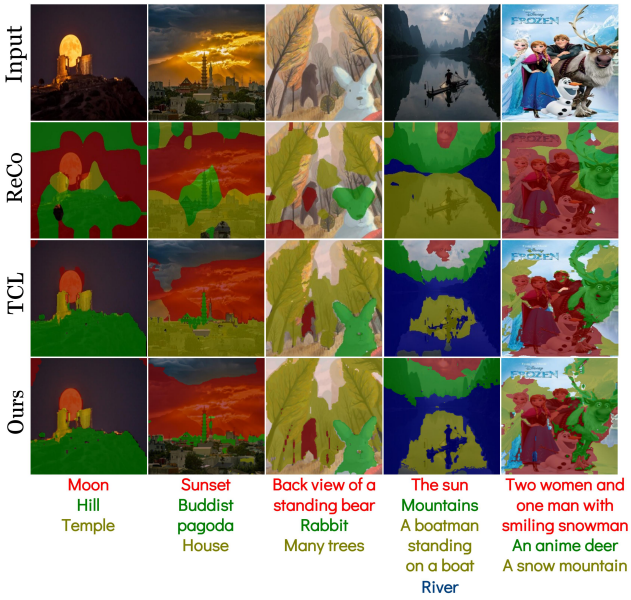

Figure 6. **Examples in the wild.** We show predictions on wild images with free-form text queries. Texts used as target classes are shown at the bottom of the images.

7. Additional Qualitative Results

7.1. Ablation study visualization

In the following, we conduct ablation studies by visualizing the effects of the proposed components in our method, including the image-text co-decomposition method, the word highlighting prompt, and the region highlighting prompt. To this end, Fig. 5 offers the visual comparison of segmentation results produced by the variants of our method on five images of the PASCAL VOC [13] dataset.

The image-text co-decomposition module equips the model with the region-word alignment ability to localize objects in the images accurately. This module aligns words with corresponding regions in the image, leading to more precise segmentation results. Furthermore, both the word and region highlighting prompts contribute to feature extraction, improving the model's ability to capture the details of the objects. Hence, the resultant model is more effective in segmenting the whole objects of interest.

7.2. Multi-noun queries

Fig. 6 shows predictions on wild web images with various text queries using the same images and queries selected from Fig. 5 of TCL [5]. Although our method is primarily designed and trained for single-noun queries, the figure demonstrates its effectiveness in processing more complex queries.

7.3. Failure case visualization

In Fig. 7, we show several failure cases of our method and two competing methods, TCL [5] and SimSeg [49], on the images of the PASCAL VOC [13] dataset.

The first example in Fig. 7a shows a common limitation of existing methods: When segmenting the "person" class, most methods focus on the most distinctive areas, namely the face in this example, and suffer from the variations in the clothes, resulting in the segment that does not cover the entire person. The second example in Fig. 7b depicts a scenario, where unexpected variations are present, *i.e.*, people showing in a television monitor. All three methods segment the outer borders of the monitor. Compared to TCL and SimSeg, our method can further segment the individuals within the monitor. Although the ground truth covers the entire TV monitor, this example validates the effectiveness of our model in localizing the individuals present on the screen.

Fig. 7c, Fig. 7d, and Fig. 7e showcase instances where co-occurrent objects, such as trains and tracks, airplanes and contrails, and boats and water, tend to be segmented together even though they are of different semantic categories. This is a challenge for our method and the two competing methods TCL [5] and SimSeg [49]. These visualization examples emphasize the difficulties of accurate

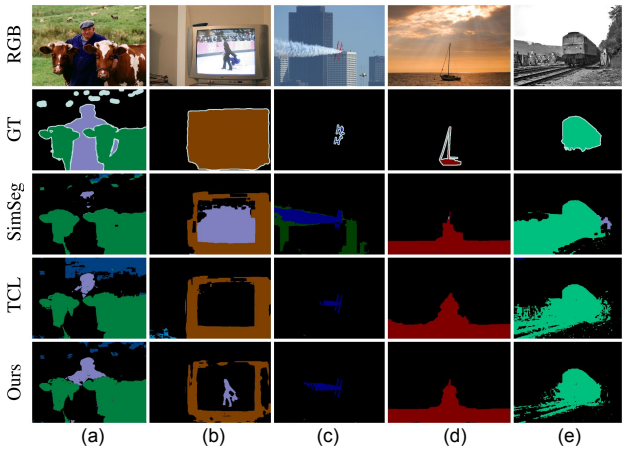

Figure 7. **Failure cases.** The proposed method is compared with the two most competitive methods, TCL [5] and SimSeg [49], on the images of the PASCAL VOC [13] dataset.

segmentation and the challenges in aligning model predictions with ground truth annotations. They provide insights into the limitations of current segmentation approaches and suggest future research directions.

8. More Implementation Details

Training time. On four NVIDIA 2080Ti GPUs, it takes eight hours to train the baseline model with only the image segmenter. On the same devices, it takes twelve hours to train our image-text co-decomposition method, which requires training an additional text segmenter. In light of the improved performance as described above, the longer training period can be justified.