A APPENDIX

A.1 TRAINING DETAILS

We conduct all the experiments on 24 $\times$ Nvidia V100 GPUs. Following Xu et al. (2022), we set the batch size to 4096, and the learning rate is initialized to 0.0016 and decayed via the cosine schedule (Loshchilov & Hutter (2016)). We use the Adam optimizer (Kingma & Ba (2015)) with a weight decay of 0.05, and train ViewCo for 30 epochs with 5 initial epochs containing linear warm-up. The number M of generated text prompts is set to 3, and the text templates are the same as Radford et al. (2021). Additionally, we perform contrastive learning using the semantic embeddings outputted by the student network f_s and the text branch. They are mapped to the same latent space through two-layer MLPs.

A.2 ADDITIONAL QUALITATIVE RESULTS

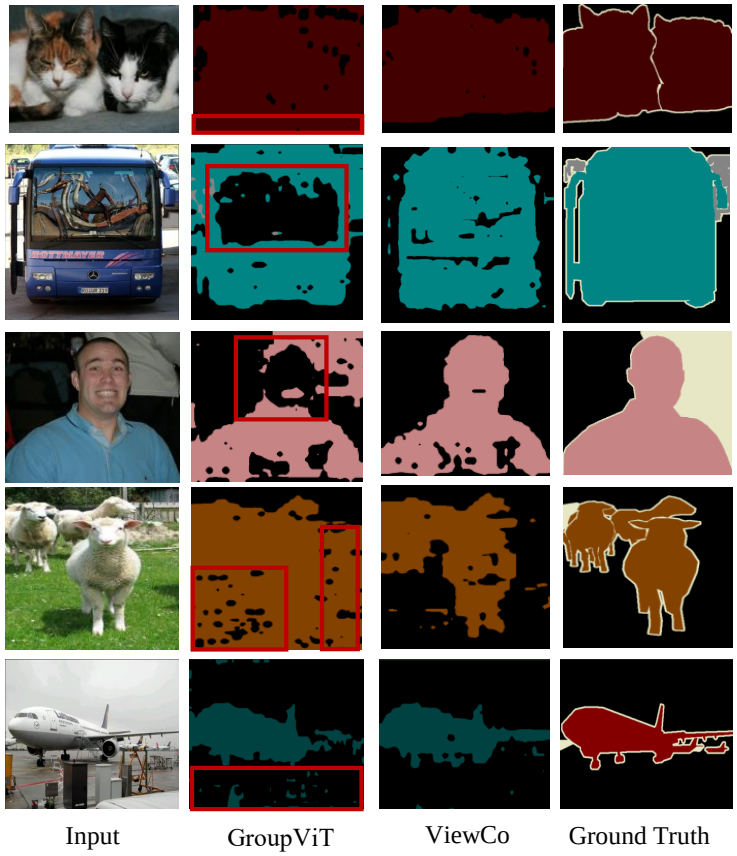

Figure 7: Qualitative comparison on PASCAL VOC 2012. The red boxes indicate obvious segmentation errors. ViewCo has better segmentation details than GroupViT (Xu et al. (2022)).

Visual comparison on PASCAL VOC 2012. Figure 7 shows the semantic segmentation visualization results of five groups of GroupViT and ViewCo. ViewCo can obtain better segmentation results than GroupViT.

A.3 MULTI-VIEW SEGMENTATION CONSISTENCY COMPARISON

The segmentation consistency comparison between GroupViT and ViewCo on multiple views is shown in Figure 8. We obtain two augmented views by cropping or rotating augmentation and then adopt the weights trained on CC12M by GroupViT and ViewCo to evaluate the segmentation consistency of the two models on multiple views. It can be seen that the segmentation of GroupViT on

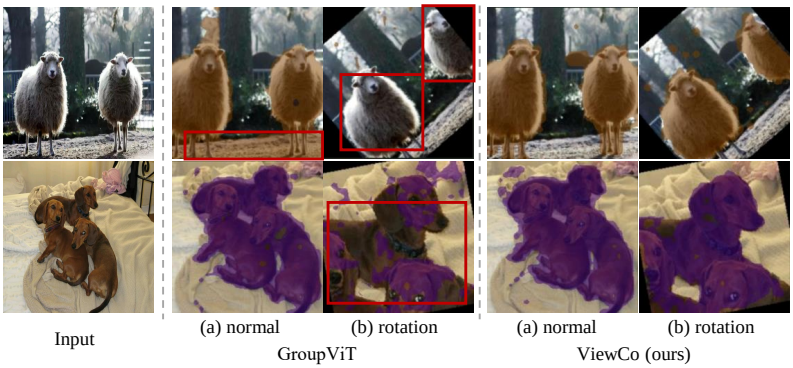

Figure 8: Visual comparison between GroupViT and ViewCo on multi-view (normal and rotated view) segmentation consistency. The red boxes indicate obvious segmentation errors. Compared with GroupViT, ViewCo has better segmentation consistency across multiple views and stable segmentation performance for rotated views.

normal and rotated views is inconsistent, and GroupViT performs poorly on rotated views compared to normal views. In contrast, ViewCo's segmentation on normal and rotated views maintains good consistency and is not disturbed by image rotation. This shows that ViewCo's cross-view segmentation consistency module learns different view consistency semantics well.

A.4 Cross-view segmentation consistency visualization analysis

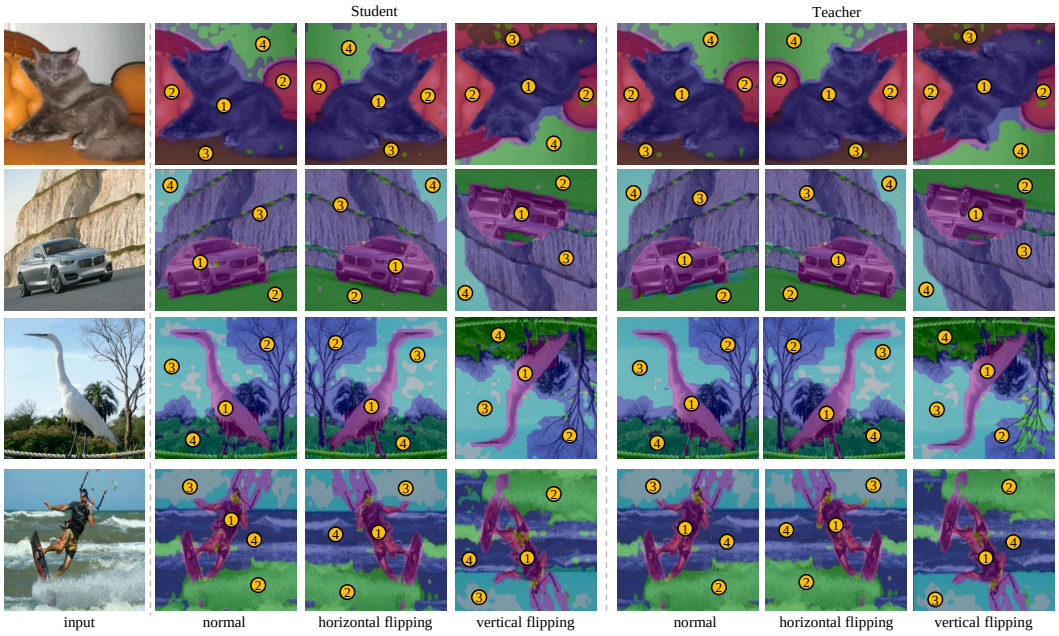

Figure 9: A visual example of ViewCo's cross-view segmentation consistency.

In addition, we further analyze the effect of image flip augmentation on ViewCo's cross-view segmentation consistency module by visualization in Figure 9. Specifically, we visualized the segmentation results of the 8 segment semantics output by the second grouping block in the teacher and student networks of different views. It should be noted that we did not perform any post-processing on the visualization results, but only numbered the main semantics in the image for the convenience of visualization. Image segmentation shows less than 8 semantics, mainly because in the grouping

block the model determines which segment semantics each image patch belongs to by computing an attention map of shape 196×8 (196 is the number of image patches). This means that some segments may not have corresponding image patches. And, in the segmentation results from different views of the same image, image regions with the same color represent the same location IDs and the same semantics. From the visualization results in Figure 9, it can be found that the semantics in the segmentation maps of different views output by the teacher and student networks maintain a good consistency in the corresponding position IDs, which again justifies the design of ViewCo's cross-view segmentation consistency module.