Supplementary of Rethinking Amodal Video Segmentation from Learning Supervised Signals with Object-centric Representation

1. Qualitative comparison between EoRaS and competitors

In the supplementary part, we show some qualitative comparisons of our model and competitors. For more comparisons, we also uploaded a folder with some GIF files comparing the predicted full masks generated by our model EoRaS and the ground truth full masks.

1.1. Qualitative comparison in Movi-B and Movi-D

Figure 1 provides a comprehensive qualitative comparison between our EoRaS and competitors across two datasets, Movi-B and Movi-D. The left column displays images from Movi-B, and the right column displays images from Movi-D, with the numbers in the upper-left corner indicating the source frame of each image. For example, 17-3 indicates this image is from the 3^{rd} frame of the 17^{th} video. Notably, we also highlight the objects with the largest predicted mask difference by framing them for ease of comparison.

When analyzing the images from Movi-B, our model outperforms competitors in many cases. For example, in the first image (17-3), our prediction for the green cylinder is superior to those of our competitors. Specifically, AISFormer predicts a full mask that extends beyond the ground truth, while SaVos predicts an incomplete mask. In the last image (26-11), only our EoRaS model accurately predicts the spout of the teapot.

Examining the Movi-D dataset, we note that AISFormer over-completes the predictions for the objects in the first two images (4-1 and 4-5), while SaVos delivers incomplete masks. Conversely, EoRaS accurately predicts the full masks of the books in the third (34-13) and fourth (34-19) images, while AISFormer and SaVos provide incomplete masks.

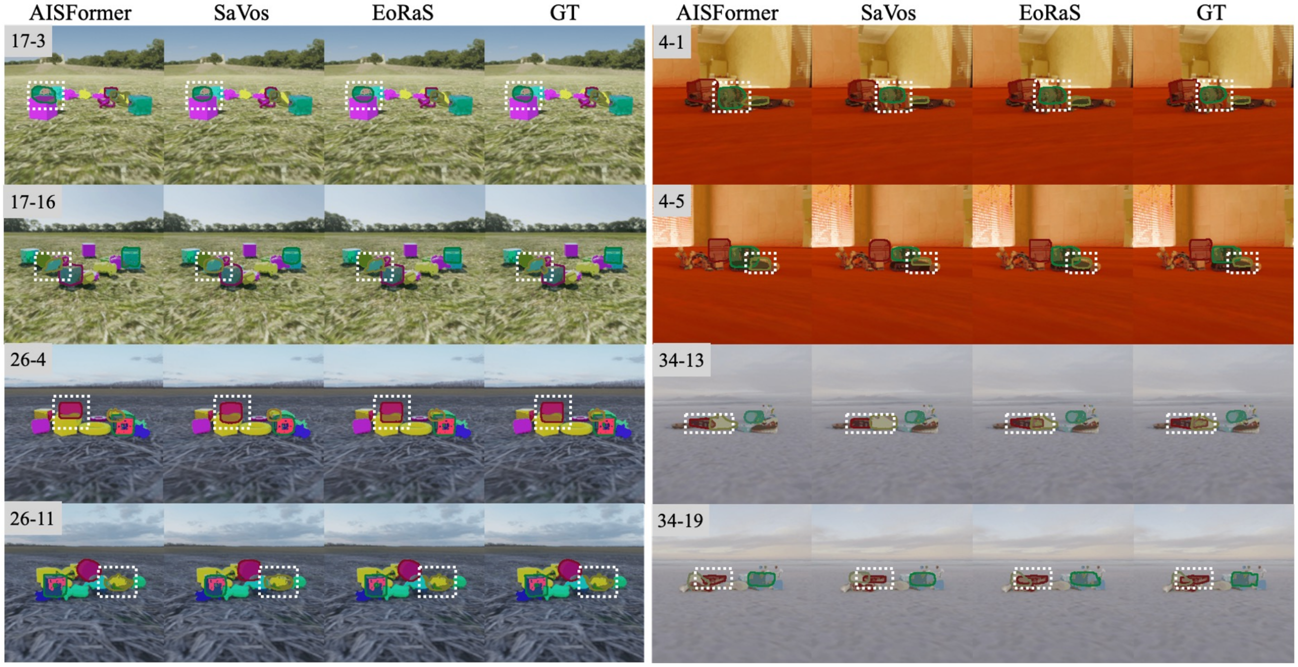

Figure 1: Qualitative comparison between our EoRaS and competitors in the Movi-B and Movi-D datasets. The images in the left column are from Movi-B, and those in the right column are from Movi-D. The numbers in each upper-left corner indicate where these images come from. For example, 17-3 indicates this image is from the 3^{rd} frame of the 17^{th} video. For convenience, we also put frames on those objects with the largest predicted mask difference.

1.2. Qualitative comparison in KITTI

In addition, we showcase the performance of our EoRaS model in the KITTI dataset (Figure 2). Given the sparsely annotated nature of the KITTI dataset, only a few frames have annotations, with no full ground truth masks available for the selected images. Nevertheless, we observe that our model outperforms competitors in certain cases. In the upper-right image (22-160), AISFormer gives a weirdly shaped mask, while SaVos gives an over-completed mask.

To add that, we also noticed that for the cases when there is no occlusion in front of one object, EoRaS can give a more accurate mask than our competitors, as shown in the yellow mask of the lower-right image (22-402), which further shows the robustness of our model.

Overall, the results presented in Figure 1 and Figure 2 suggest that our EoRaS model outperforms competitors in terms of accuracy, completeness, and robustness across various datasets.

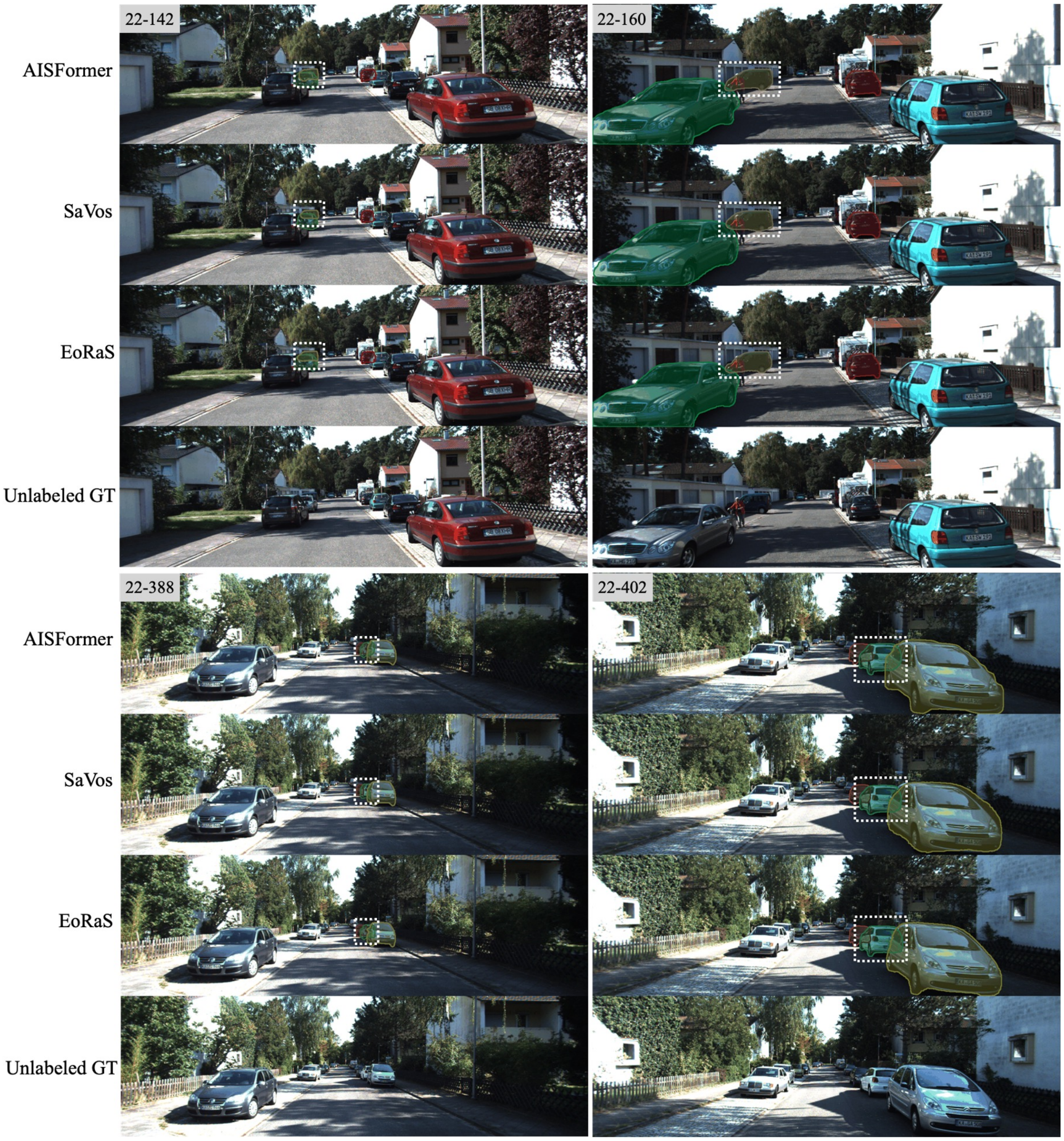

Figure 2: Qualitative comparison between our EoRaS and competitors in the KITTI dataset. The numbers in each upper-left corner indicate where these images come from. For example, 22-142 indicates this image is from the 142^{nd} frame of the 22^{nd} video. For convenience, we also put frames on those objects with the largest predicted mask difference. Due to the sparse labeling of the KITTI dataset, many images do not have ground-truth full masks.