6 Supplementary Material

6.1 Qualitative Analysis on the Human Pose Knowledge

We have shown the quantitative analysis on the effectiveness of human pose knowledge in Table 2. Additionally, as shown in Figure 6, we include a qualitative analysis to understand the advantages of incorporating human pose knowledge.

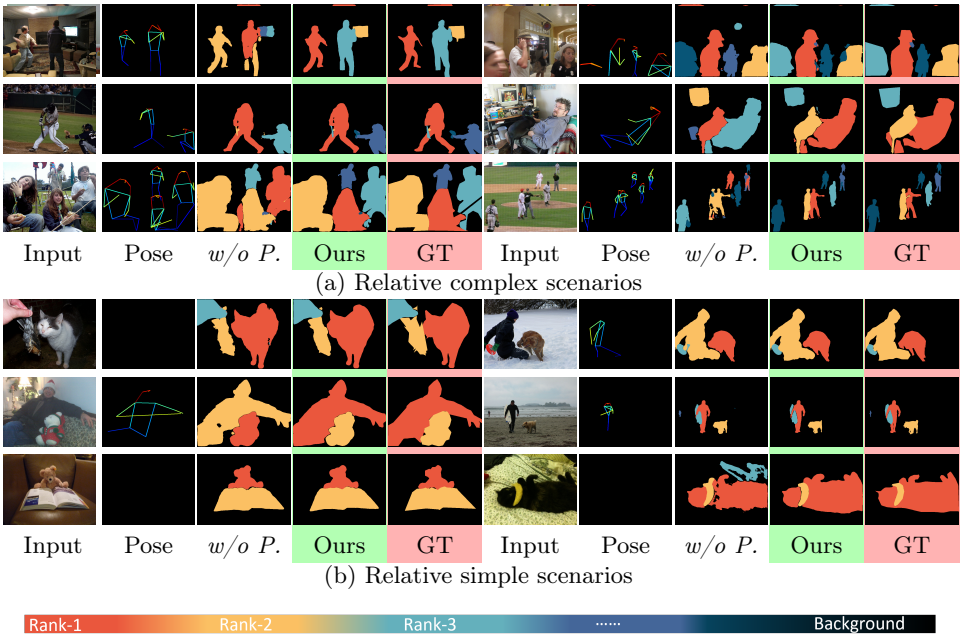

Fig. 6: Analysis on the Pose Knowledge. In general, our PoseSOR (*i.e.*, Ours) yields more favorable results by integrating human pose knowledge into SOR model. It is particularly evident in the complex scenarios (a), which typically encompass complex human activities and high-level interactions between human participants and their surroundings, while both *w/o P.* and Ours perform well in the simple scenarios (b). *w/o P.* is an ablated version of PoseSOR, achieved by removing the pose queries and the related components. Pose refers to the pose prediction made by our PoseSOR.

6.2 The Complexity of PoseSOR

We further include a complexity analysis of PoseSOR. Our PoseSOR has 220M parameters and the inference time is 61ms per image at 768x768 resolution and 33ms per image at 512x512 resolution on a machine with an RTX4090 GPU and an Intel i7-13700 CPU. We show detailed comparisons with PSR and OCOR in Table 5. Overall, our efficiency is comparable, and it exhibits more favorable SOR performance.

Table 5: Complexity Analysis of PoseSOR. We evaluate the below models on a machine with an RTX4090 GPU and an Intel i7-13700 CPU.

Method	Ours	Ours	PSR	OCOR
Resolution	768x768	512x512	640x480	800x800
FLOPs↓	470G	229G	325G	608G
Parameters↓	220M	220M	217M	338M
Runtime↓	61ms	33ms	30ms	68ms
SA-SOR↑	0.673	0.673	0.644	0.541
SOR↑	0.871	0.860	0.815	0.873
MAE↓	7.23	7.35	9.59	10.2

6.3 Generalization Ability of PoseSOR

We test PoseSOR on new data, as shown in Figure 7, where human poses are completely/partially occluded. We find PoseSOR generally performing well, even in scenes without humans.

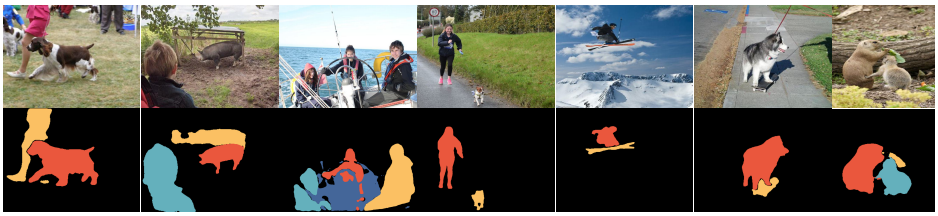

Fig. 7: Generalization Ability of PoseSOR. We evaluate PoseSOR on new data.

6.4 More Visual Results

We present more visual results in Figure 8, Figure 9 and Figure 10.

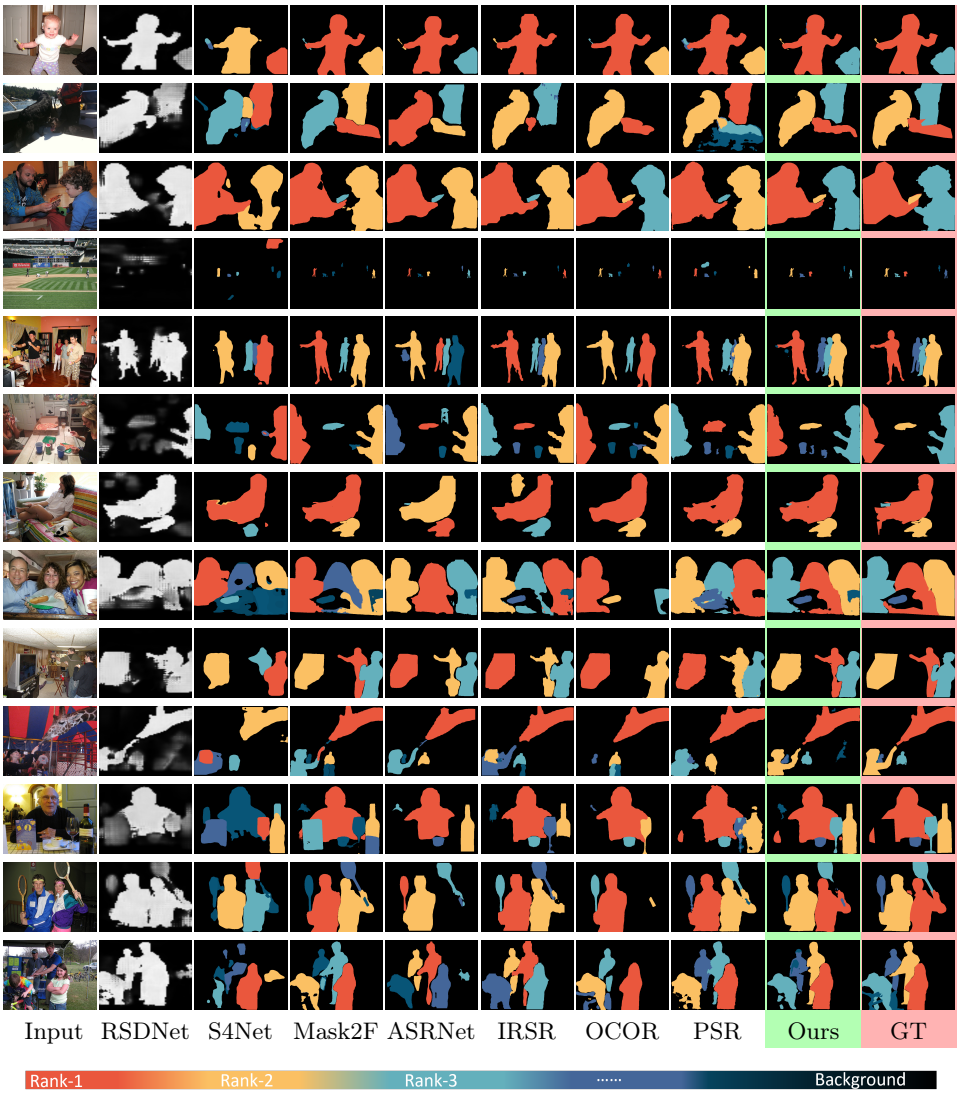

Fig. 8: Visual Results. Our method generally produces more accurate results that align with human labels (GT). Salient instances are colorized using varying color temperatures, ranging from warm to cold, to indicate the shifting order of human attention.

Fig. 9: Visual Results. Our method generally produces more accurate results that align with human labels (GT). Salient instances are colorized using varying color temperatures, ranging from warm to cold, to indicate the shifting order of human attention.

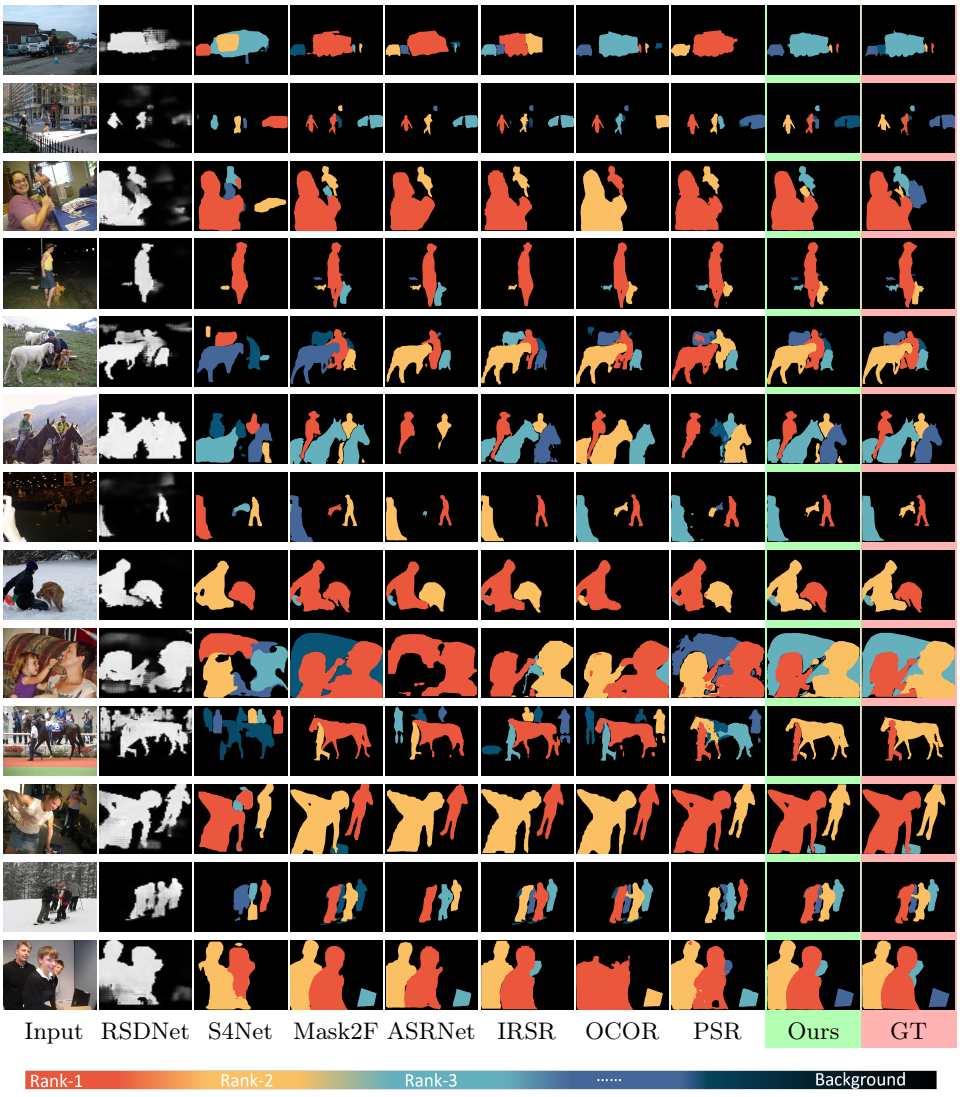

Fig. 10: Visual Results. Our method generally produces more accurate results that align with human labels (GT). Salient instances are colorized using varying color temperatures, ranging from warm to cold, to indicate the shifting order of human attention.