SeqRank: Sequential Ranking of Salient Objects

Anonymous submission

In this supplementary material, we provide an introduction to multi-head attention. Then we discuss the inference cost of our SeqRank, and offer more visual results at the end of this material.

Multi-Head Attention. SeqRank is an innovative SOR method that includes two novel modules: the Fovea Module (FOM) and the Sequential Ranking Module (SRM). Both of these modules utilize the multi-head attention layer (Vaswani et al. 2017), as their basic components. Multi-head attention is a powerful mechanism for modeling complex relationships between tokens, and the standard multi-head attention can be expressed as follows:

$$MultiHead(Q, K, V) = concat(head_1, head_2, ..., head_h)W_o \tag{1}$$

$$head_i = Attention(Q_i, K_i, V_i),\ where\, i \in \{1, 2, ..., h\} \tag{2}$$

where $W_o \in R^{d\times d}$ is a weight matrices for aggregating different heads, h is the number of heads, $concat$ is the concatenate operation and Q_i, K_i, V_i are calculated as:

$$Q_i = QW_i^Q,\ K_i = KW_i^K,\ V_i = VW_i^V \tag{3}$$

where $Q \in R^{N\times d}$ is the query sequence, $K \in R^{L\times d}$ is the key sequence and $V \in R^{L\times d}$ is the value sequence. $W_i^Q \in R^{d\times \frac{d}{h}}, W_i^K \in R^{d\times \frac{d}{h}}, W_i^V \in R^{d\times \frac{d}{h}}$ are linear projections for mapping each sequence to different sub-spaces. $Attention$ is computed as:

$$Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d}})V_i \tag{4}$$

where $softmax$ is the softmax function. We use multi-head attention for building our cross-attention and self-attention layers.

Discussion on Inference Cost. We notice that the computation cost of our SeqRank varies from image to image due to the sequential ranking strategy. Specifically, the SRM predicts one attention shift at a time and predicts no object in the last run. Thus, the SRM is required to run for $n + 1$ times for an image with n salient objects. Fortunately, the backbone network, the pixel decoder, and the FOM only run once since the image features are shareable, and the segmentation of salient objects can be handled individually. Besides, our SRM is lightweight, bringing about 0.6% additional FLOPs for predicting one more attention shift. Therefore, the inference procedure of our SeqRank can still be efficient.

More Visual Results. We present more visual results in Figure 1 and Figure 2. In general, our model produces more favorable results compared to the other methods.

References

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Figure 1: Visual Comparison. Salient instances are colorized using varying color temperatures, ranging from warm to cold, indicating the order in which they are visited.

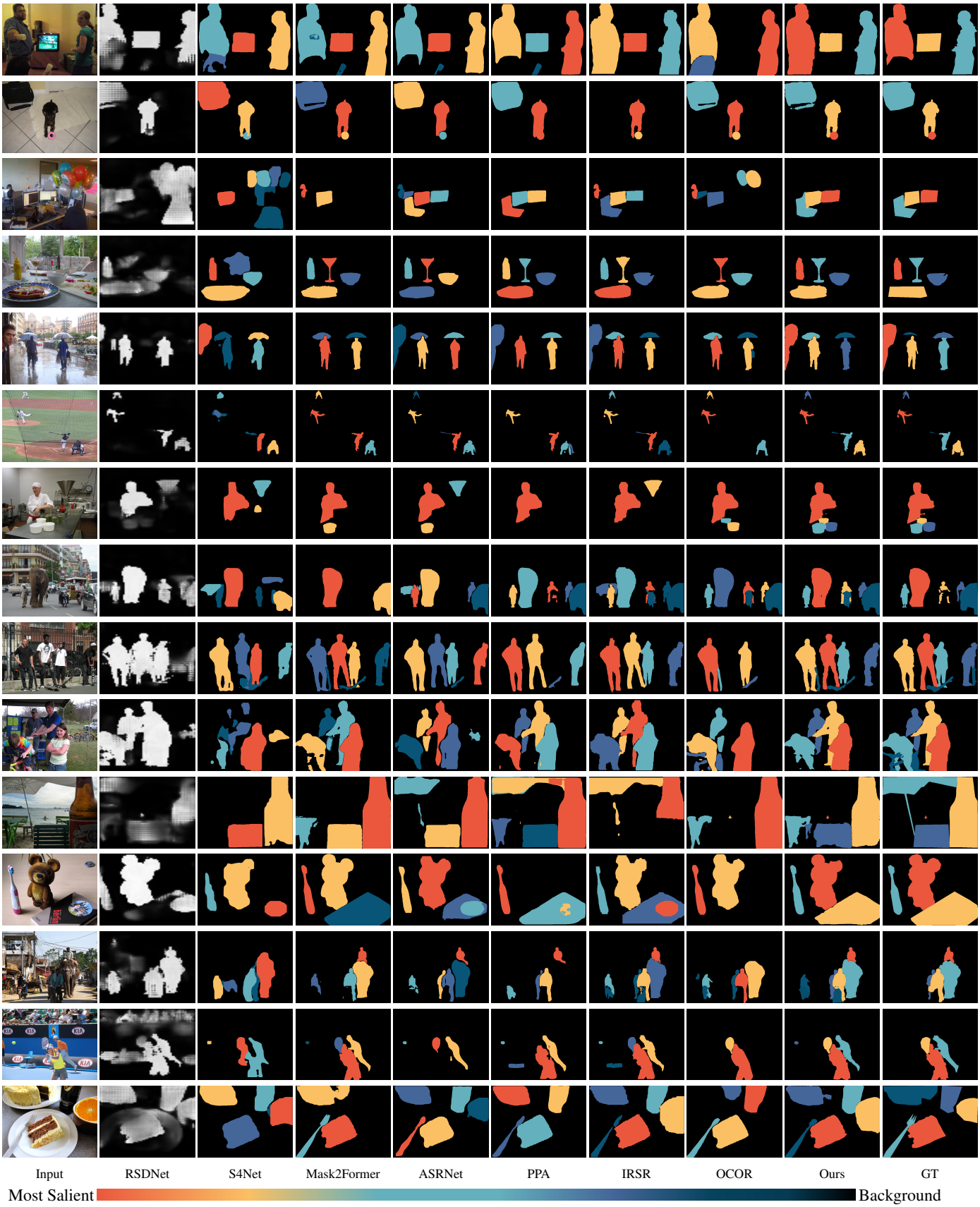

Figure 2: Visual Comparison. Salient instances are colorized using varying color temperatures, ranging from warm to cold, indicating the order in which they are visited.