Self-Supervised Spatial-Temporal Feature Learning for Video Correspondence
—– NeurIPS 2022 Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

The supplementary material contains: 1) ablation study of different contrastive models; 2) ablation study of entropy-based selection; 3) more qualitative examples for video object segmentation.

1 Ablation study of different contrastive models

Given a query point randomly sampled in the target frame, we visualize the result of computing the local correlation and global correlation map $w.r.t.$ reference frame. The dashed line in red represents the range of computing correlation map $w.r.t.$ query point. The reference frame is randomly sampled in the memory bank of inference strategy. Given a query point randomly sampled in the target frame, we visualize the result of computing the local correlation and global correlation map $w.r.t.$ reference frame. The dashed line in red represents the range of computing correlation map $w.r.t.$ query point. The reference frame is randomly sampled in the memory bank of inference strategy.Given a query point randomly sampled in the target frame, we visualize the result of computing the local correlation and global correlation map $w.r.t.$ reference frame. The dashed line in red represents the range of computing correlation map $w.r.t.$ query point. The reference frame is randomly sampled in the memory bank of inference strategy

2 Ablation study of entropy-based selection

We make detailed ablation study for our entropy-based selection in terms of both visual perception and quantitative comparison. As shown in Figure 1, the entropy map has a higher response on moving objects involved in severe deformation and occlusions, which should be paid more attention to. In Table 1, we adopt different thresholds to generate the mask m with high entropy. The baseline is to apply local correlation distillation for all queries, $i.e.$, T = 1.0. When setting T with 0.1, the performance drops to 67.4% due to the underutilization of the supervision from finest pyramid level. The results of setting T with 0.4 and 0.7 indicate applying distillation in the region with high entropy exhibits a performance gain.

Method	Dataset	$\mathcal{J}\&\mathcal{F}_m \uparrow$
T = 0.1	YTV	67.4
T = 0.4	YTV	68.3
T = 0.7	YTV	69.0
T = 1.0	YTV	68.1

Table 1: **The quantitative results on the validation set of DAVIS-2017.** The Dataset represents dataset(s) used for training. YTV:YouTube-VOS []

3 More qualitative examples for video object segmentation

We do the most of experiments

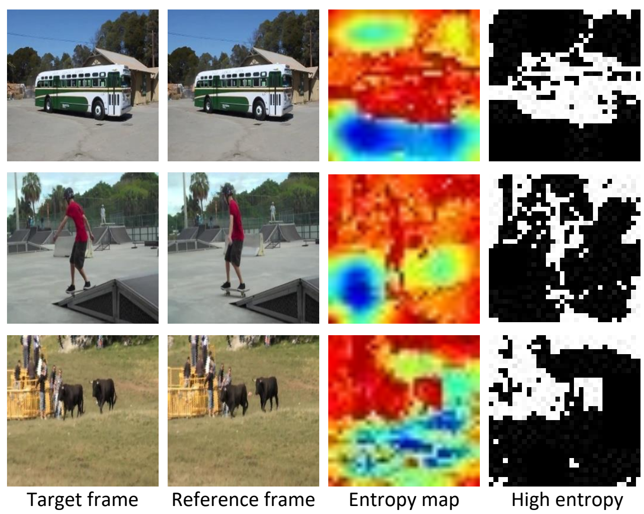

Figure 1: **Visualization of the entropy map.** We compute the entropy for each query in target frame using Eq 8 in our main paper. The mask with high entropy is generated by setting a threshold.