Supplementary Material for
Learning the What and How of Annotation in Video Object Segmentation

Thanos Delatolas[1,2] Vicky Kalogeiton[3] Dim P. Papadopoulos[1,2]
[1] Technical University of Denmark [2] Pioneer Center for AI
[3] LIX, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris
atde@dtu.dk, vicky.kalogeiton@lix.polytechnique.fr, dimp@dtu.dk

https://eva-vos.compute.dtu.dk/

1. Ablation study on the number of clicks

In the main paper, we experiment with two annotation types: *'mask drawing'* and *'corrective clicks'*. For *'corrective clicks'*, the annotator clicks 3 times to improve the segmentation quality of the selected frame f_* and determines the number of positive and negative clicks. In Fig. 1, we present an ablation analysis on the number of corrective clicks. Each compared method selects the next frame using an oracle and considers only a specific number of corrective clicks. We observe that 3 clicks outperform both 5 and 10 clicks at lower annotation budgets and have almost identical performance at higher budgets.

2. Video comparisons

We provide side-by-side comparisons of two videos from the MOSE-long test set under different annotation budgets in the attached video file *video_comparisons.mp4*. Similar to Sec. 5.3 in the main manuscript, we compare our full pipeline against the following methods: Oracle, which selects both the frame and the annotation type by using an oracle, and Mask-only, which selects the next frame randomly and considers only the *'mask drawing'* annotation type. The provided video consists of the following: (top-left) ground-truth masks, (top-right) predicted masks by Oracle, (bottom-left) predicted masks by EVA-VOS, and (bottom-right) Mask-only.

3. Human annotator simulation

In this work, we only perform experiments by simulating the human intervention. In Fig. 2, we present two frames with simulated clicks. We observe that our click-simulation algorithm correctly generates positive clicks where there is no previous mask and negative clicks where the previous mask does not belong to the target object.

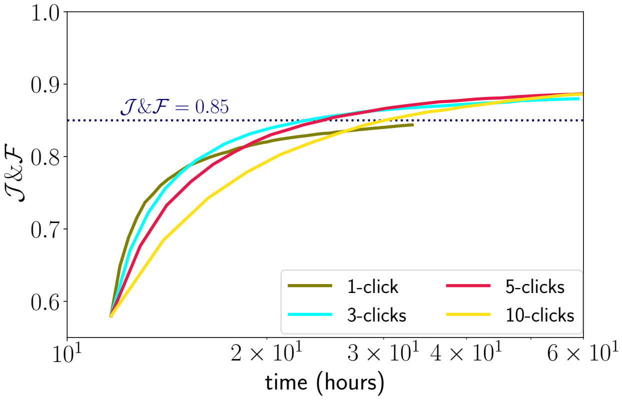

Figure 1. Ablation on the number of *'corrective clicks'*

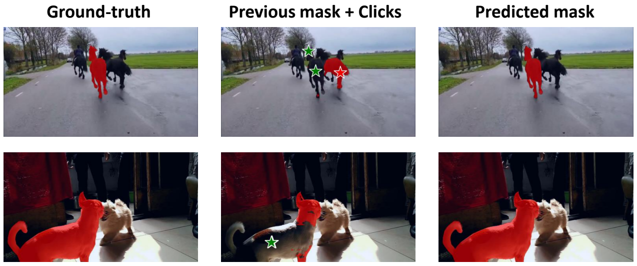

Figure 2. **Simulated clicks.** The green star denotes a positive click, while the red star denotes a negative click. The ground-truth images indicate the target object.

4. QNet frame selection

In this section, we visualize the Frame selection procedure (Sec. 3.2 in the main paper). In Fig. 3, we visualize, in the feature space, the frames with their corresponding masks of two videos after t=3 (above) and t=1 (below) iterations of annotation. We observe that our method is capable of selecting frames with poor segmentation quality. Furthermore, frames with similar segmentation quality tend to be closer to the feature space. This demonstrates the robustness and effectiveness of our training process.

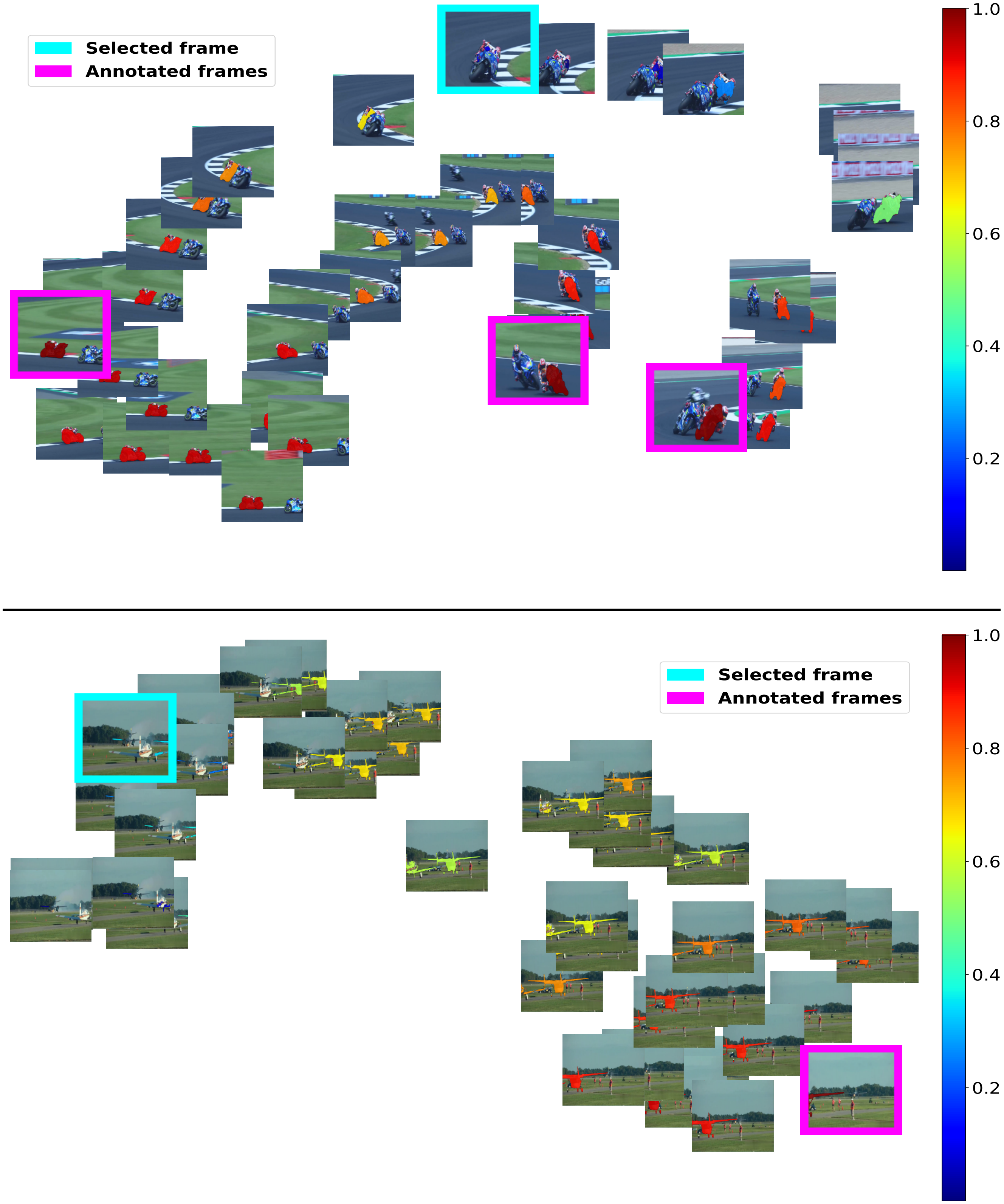

Figure 3. **QNet Frame selection.** We first extract embeddings using QNet for all frames of the video and their corresponding mask. Then, we select the frame f_* as the one with the maximum distance in the feature space from its closest previously annotated frame. Here, we visualize the frames and their masks in the 2D space using T-SNE. The color of each mask denotes its $\mathcal{J}\&\mathcal{F}$ while the outline (where it exists) of each image denotes whether it is annotated or it is the selected frame. We observe that the selected frame (cyan outline) has a very low $\mathcal{J}\&\mathcal{F}$, while frames with similar $\mathcal{J}\&\mathcal{F}$ tend to be close to the feature space.