Supplemental Materials on
Interactive Video Object Segmentation Using Global and Local Transfer Modules

Yuk Heo, Yeong Jun Koh, and Chang-Su Kim

S-1 Detail Network Structures

Fig. S-1 shows more detailed structures of the proposed A-Net and T-Net, in which we express the spatial resolutions of signals in terms of the width W and the height H of an input frame and also specify the number C of channels.

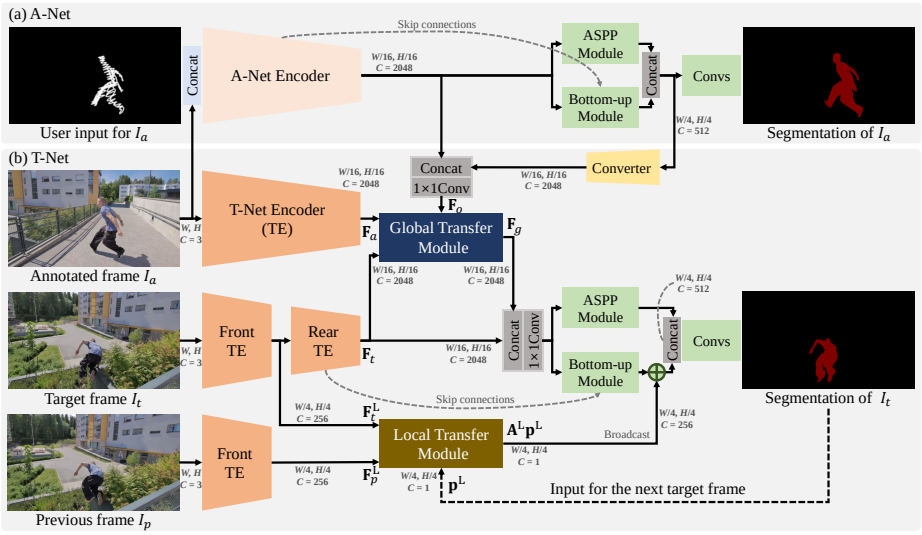

Fig. S-1: Detailed structures of the proposed A-Net and T-Net.

S-2 User Study

We did a user study, in which 10 participants were asked to perform VOS using the proposed algorithm. Note that an example of such interactive segmentation is available in the supplementary video clip. For each video sequence in the DAVIS2016 dataset [33], Table S-1 lists the average running time, the average number of rounds, the mean J score, and the mean F score of all the participants. Except for the "car-roundabout" and "motocross-jump" sequences, the participants were satisfied with the VOS results within three rounds. The J and F scores on the "bmx-trees," "kite-surf," and "paragliding-launch" sequences are relatively low, even though participants were satisfied with their results after a few rounds. This is because thin parts such as the kite strings and the suspension

lines of the paraglider were not properly segmented as shown in Fig. S-2. Nevertheless, the proposed algorithm segments out the main subjects (*i.e.* bike and rider, surfer, and pilot) accurately, which makes users satisfied with the results. Fig. S-3 provides more interactive VOS results in this user study.

Table S-1: User study results on the DAVIS2016 validation set.

Video sequence	Average time	Average round	Mean J	Mean F
blackswan	11.0	1.0	0.930	0.965
bmx-trees	41.9	2.2	0.546	0.720
breakdance	65.2	2.7	0.809	0.807
camel	20.8	1.3	0.925	0.969
car-roundabout	47.1	3.2	0.947	0.929
car-shadow	30.1	2.7	0.949	0.986
cows	19.6	1.2	0.925	0.948
dance-twirl	48.4	2.4	0.816	0.837
dog	18.5	1.6	0.912	0.933
drift-chicane	29.5	2.5	0.850	0.939
drift-straight	30.1	2.2	0.879	0.833
goat	14.1	1.0	0.853	0.840
horsejump-high	13.0	1.0	0.820	0.886
kite-surf	9.6	1.0	0.641	0.474
libby	12.4	1.1	0.808	0.920
motocross-jump	67.2	3.9	0.863	0.716
paragliding-launch	12.4	1.0	0.614	0.206
parkour	42.9	2.4	0.872	0.921
scooter-black	29.3	2.0	0.845	0.746
soapbox	32.3	1.6	0.869	0.842

Fig. S-2: Examples of segmentation results after the 1st rounds on the "bmx-trees," "kite-surf," and "paragliding-launch" sequences in the user study. Positive scribbles, which were provided by users, are depicted in green.

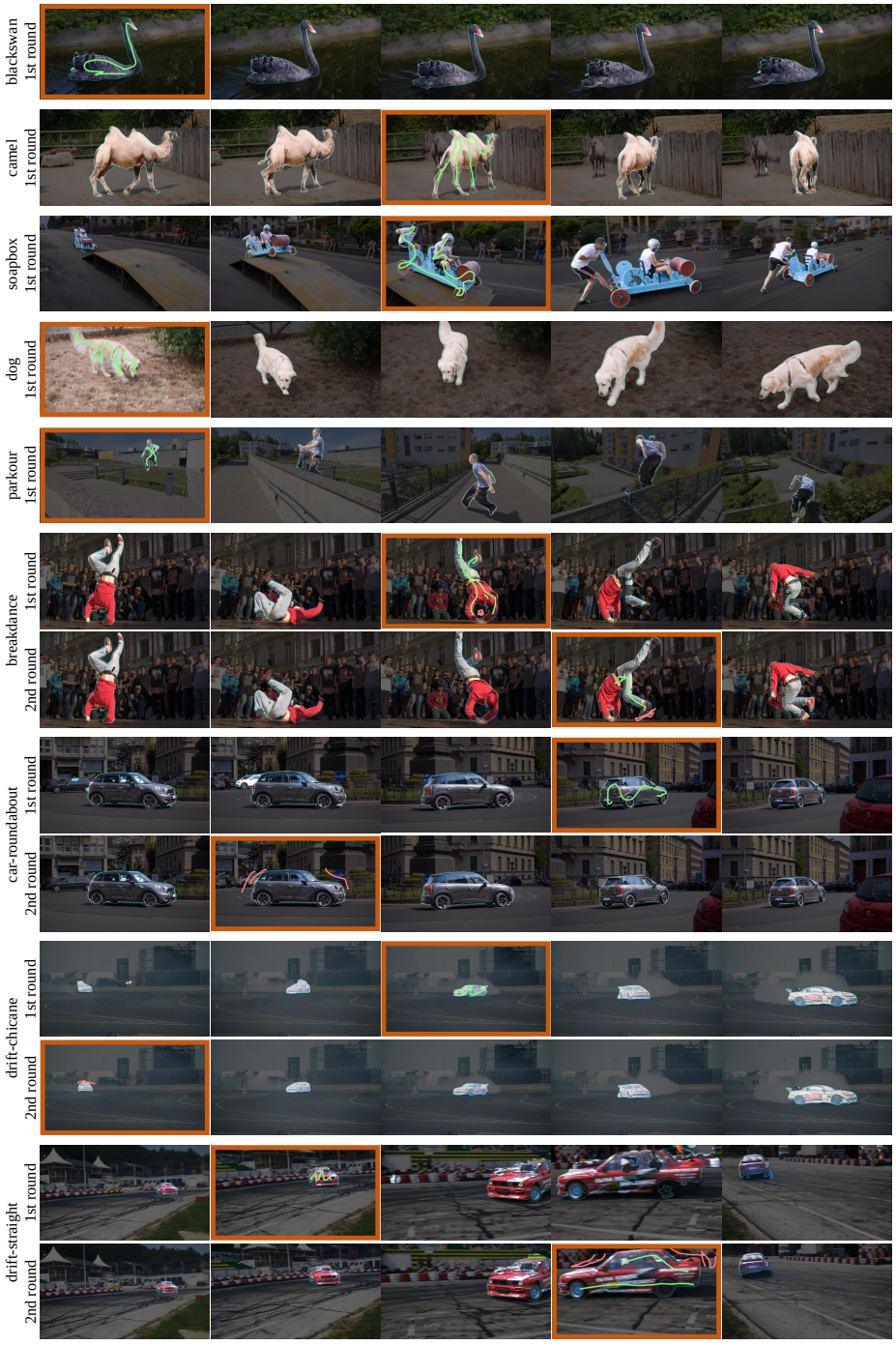

Fig. S-3: More qualitative results of the proposed interactive VOS algorithm in the user study.

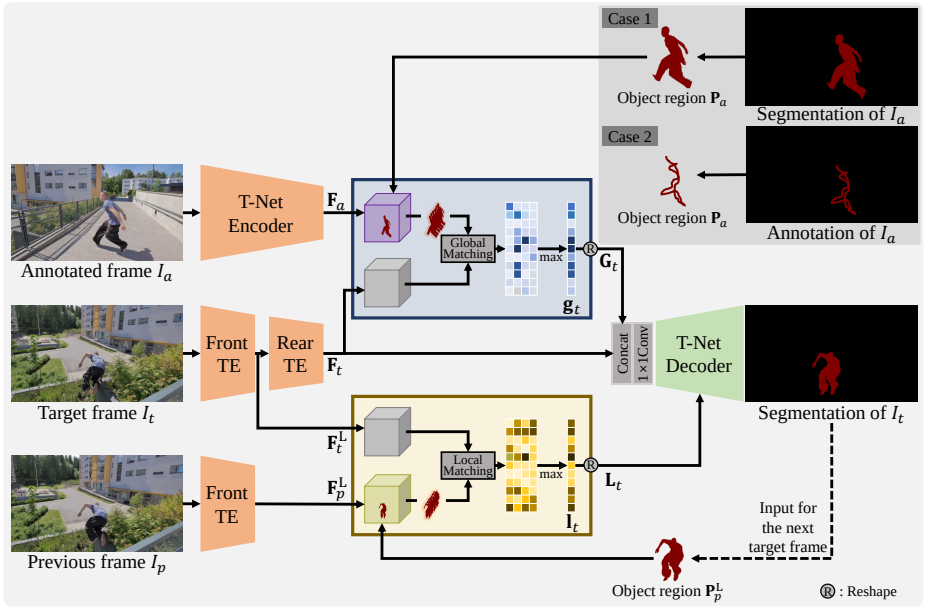

Fig. S-4: Detailed structure of the modified network in the ablation study. In the global matching, the region predicted by A-Net is chosen for the target object region as an example.

S-3 Details of Modified Network in Ablation Study

In Section 4.3 in the paper, through an ablation study, we verified that our global and local transfer modules are more effective than the global and local matching in [43] for interactive VOS. We replaced the proposed global and local modules with the matching method in [43], as shown in Fig. S-4.

Let $\mathbf{S}$ denote the similarity matrix between the annotated and target frames, whose element $s(i,j)$ is the cosine similarity between the feature of ith pixel in $\tilde{\mathbf{F}}_t$ and that of the jth pixel in $\tilde{\mathbf{F}}_a$. Then, we consider the similarity for pixels within a target object region in the annotated frame. Let $\mathbf{P}_a$ denote the set of pixels that belong to a target object region in the annotated frame. Note that there are two ways to determine the target object region: 1) the region predicted by A-Net or 2) the set of scribble-annotated pixels, as illustrated in Fig. S-4. We compute the global matching vector $\mathbf{g}_t \in \mathbb{R}^{H_1 W_1 \times 1}$ for the target frame by taking the maximum similarity at each pixel i, whose ith element $g(i)$ is given by

$$g(i) = \max_{j \in \mathbf{P}_a} s(i,j). \tag{S-1}$$

We then reshape the vector $\mathbf{g}_t$ to the global matching map $\mathbf{G}_t \in \mathbb{R}^{H_1 \times W_1}$ and concatenate it to $\mathbf{F}_t$, instead of the distribution $\mathbf{F}_g$.

For the local matching, at each pixel i, we compute the similarity between the local region $\mathcal{N}_i$ (in Eq. (4) in the paper) in the target frame and the segmentation

region $\mathbf{P}_p^{\mathrm{L}}$ in the previous frame. Similar to the global matching, we compute the local matching vector $\mathbf{l}_t$ through the maximization,

$$l(i) = \max_{j \in \mathbf{P}_p^{\mathrm{L}} \cap \mathcal{N}_i} s(i,j). \tag{S-2}$$

The local matching vector is also reshaped to the local matching map $\mathbf{L}_t$ and then $\mathbf{L}_t$ is fed into the mid-layer in the T-Net decoder, as shown in Fig. S-4.

S-4 Validation of Decoder

We carry out another ablation study to validate the proposed decoder architecture. As shown in Table S-2, without the ASPP module, the proposed algorithm experiences the performance degradation.

Table S-2: Ablation study on the decoder (J scores on the validation set in DAVIS2017).

	Round				
Method	1st	2nd	3rd	4th	5th
w/o ASPP	0.664	0.702	0.735	0.747	0.755
Proposed	**0.676**	**0.732**	**0.762**	**0.772**	**0.783**