Supplementary Materials: Sketch-based Video Object Segmentation: Benchmark and Analysis

Ruolin Yang[1]
yangruolin@bupt.edu.cn
Da Li[2]
dali.academic@gmail.com
Conghui Hu[3]
conghui@nus.edu.sg
Timothy Hospedales[2]
t.hospedales@ed.ac.uk
Honggang Zhang[1]
zhhg@bupt.edu.cn
Yi-Zhe Song[2]
y.song@surrey.ac.uk

[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] SketchX, CVSSP University of Surrey, UK
[3] Department of Computer Science, National University of Singapore

A Experiments

A.1 More Implementation Details

During training, we use random affine transformation, cropping and horizontal flip for data augmentation. Also, we set the size of the memory bank as 3 to save computation resources. During inference, frames and sketches are resized to 480*480. We update memory bank every 5 frames and propagate masks temporally without any post-processing. For fair comparisons, text encoder is frozen at all the time. As for the scribbles, we first generate a skeleton tree according to the binary mask of the target object, and then select the longest path in the tree as the final scribble. We conduct all experiments with a batch size of 8 and weight decay of 1e-7. All experiments are trained end-to-end for 150,000 iterations on a single NVIDIA RTX 3090 GPU.

A.2 Further Experimental Studies

Effect of joint training. In addition to directly testing the pre-trained model on YouTube-VOS, we further jointly train the model using both Sketch-DAVIS and Sketch-YouTube-VOS. As shown in Table 1, the joint training can further boost the performances on Sketch-DAVIS16 and Sketch-DAVIS17 by a large margin.

	Sketch-DAVIS16			Sketch-DAVIS17		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
#1	81.6	80.2	83.1	70.2	66.9	73.4
#2	**84.2**	**83.2**	**85.2**	**72.9**	**69.8**	**76.1**

Table 1: Performance on Sketch-DAVIS16 and Sketch-DAVIS17 validation datasets. #1 indicates that testing the pre-trained model on YouTube-VOS directly. #2 means joint training Sketch-YouTube-VOS and Sketch-DAVIS.

Further studies of Cross-Q. We also investigate the alternatives to the function of Eq.4 in the main paper to further study the ways of fusing sketch and frame features of our best designs: Cross-Q. From the Table 2, we can see that all options work well in fusing sketch and visual features. While concatenating query and weighted value leads to an increase and achieves the best performance of $\mathcal{J}\&\mathcal{F}$ to 75.44 using visual and sketch features of res5.

Function	Level	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Add	res5	74.75	72.67	76.84
	multi-level	75.11	73.18	77.05
Mul	res5	74.80	72.78	76.83
	multi-level	74.94	72.90	77.00
Concat	res5	**75.44**	**73.38**	**77.50**
	multi-level	75.10	73.04	77.16

Table 2: Ablation study on design of function (Eq 4 in the main paper) for Cross-Q on Sketch-YouTube-VOS validation set. Add indicates element-wise add, Mul indicates element-wise multiplication, and Concat indicates concatenate. Level of res5 means using res5-sketch+res5-visual features while multi-level means using multi-level-visual+res5-sketch features.

Subject bias. Considering that sketches and drawings are very subjective in general and to explore the effect of subject bias, we present the results by using sketches drawn by different people in Tab. 3. Remarkably, the variance in segmentation results among different sketches appears to be very small. And when sketches of different styles are provided, Fig. 1 demonstrates that the segmentation results remain similar. These results serve as verification that our proposed model can be utilized by any user without noticeable bias.

Person	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
#1	74.10	72.11	76.08
#2	74.02	72.03	76.02
#3	73.89	71.99	75.79

Table 3: The segmentation results by using sketches drawn by different people on YouTube-VOS dataset.

Comparison of different references. Figure 2 presents the comparison on the performance of different references, *i.e.* text, scribble, sketch and mask. As can be seen, the presented videos contain many similar objects. In such cases, text-based model failed to detect and track the object, while scribble-based model loses the track of object and can not distinguish the object in the following frames. In contrast, sketch-based and mask-based model can track

and segment objects accurately. It is also worth noting that sketches are much cheaper to collect than pixel-level masks, which demonstrates the superiority of sketch-based approaches.

More qualitative results. Figure 3 and 4 provide visual results on Sketch-DAVIS validation set. We can observe that our model is robust even in videos with multiple similar-looking objects, appearance changes or fast motion. Figure 5 shows more visual results on Sketch-YouTube-VOS.

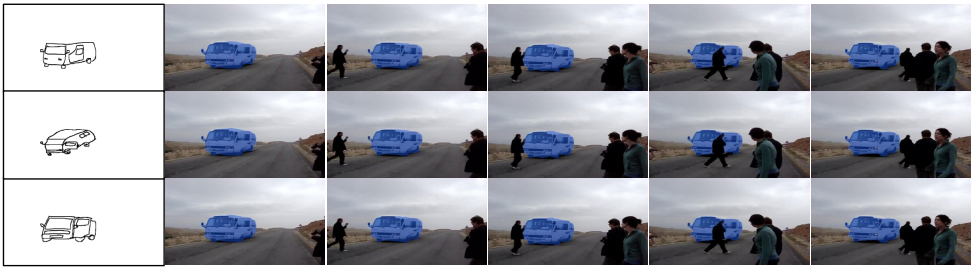

Figure 1: The results of different style sketches.

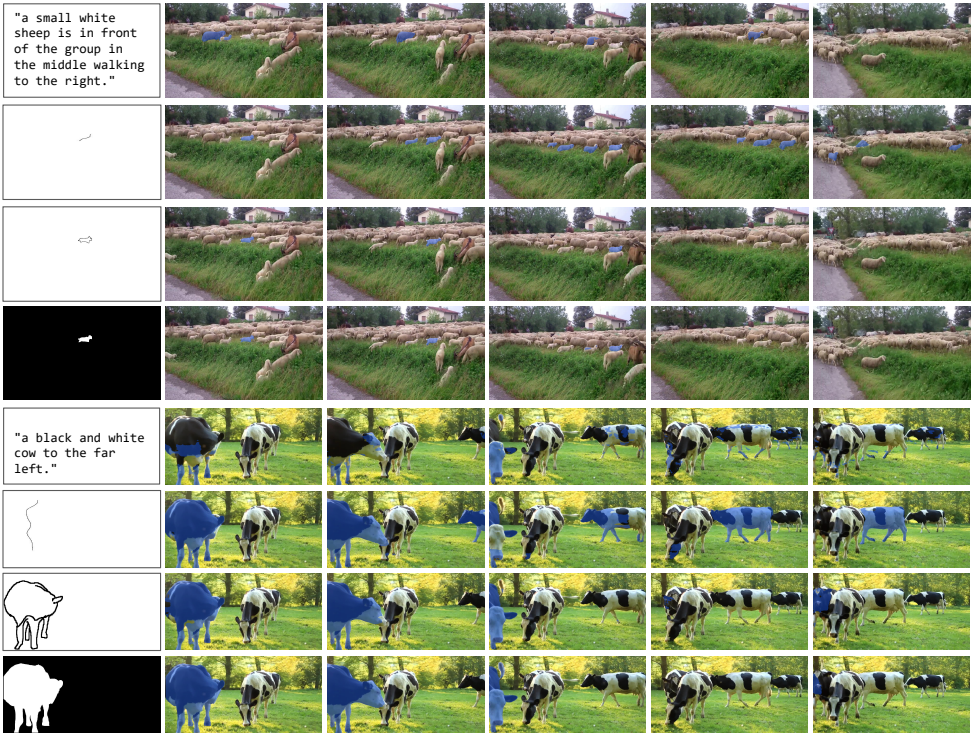

Figure 2: Visual comparison between different references. Best viewed in color.

Figure 3: Qualitative results on the Sketch-DAVIS16 validation set. Best viewed in color.

Figure 4: Qualitative results on the Sketch-DAVIS17 validation set. Best viewed in color.

Figure 5: Qualitative results on the Sketch-YouTube-VOS validation set. Best viewed in color.

B Dataset

B.1 Dataset statistics

Fifure 6 shows the distribution of object number per category in our Sketch-DAVIS-VOS dataset and Sketch-YouTube-VOS dataset. Figure 7, 8 and 9 present more examples of our datasets. We can see that the objects can be easily identified by sketches based on the salient visual details (*e.g.*, pose). Also, it can be observed that sketch helps to represent object parts in row 4 of figure 9. Table 4 shows the detailed comparison of the train and validation split for different datasets.

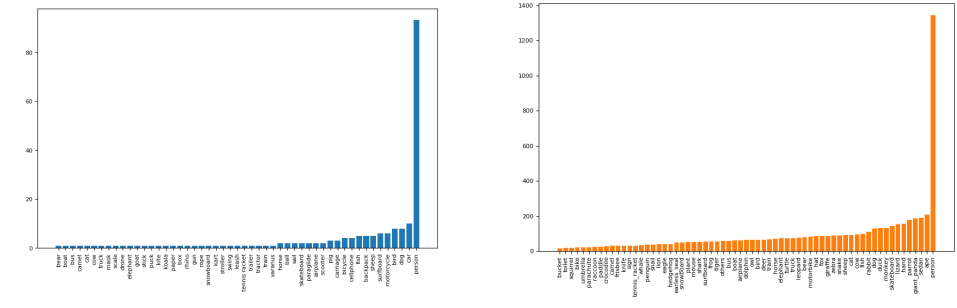

Figure 6: The distributions of object number per category in Sketch-DAVIS-VOS dataset (left) and Sketch-Youtube-VOS dataset (right).

Figure 7: Examples of Sketch and corresponding video frames in our Sketch-DAVIS16 dataset.

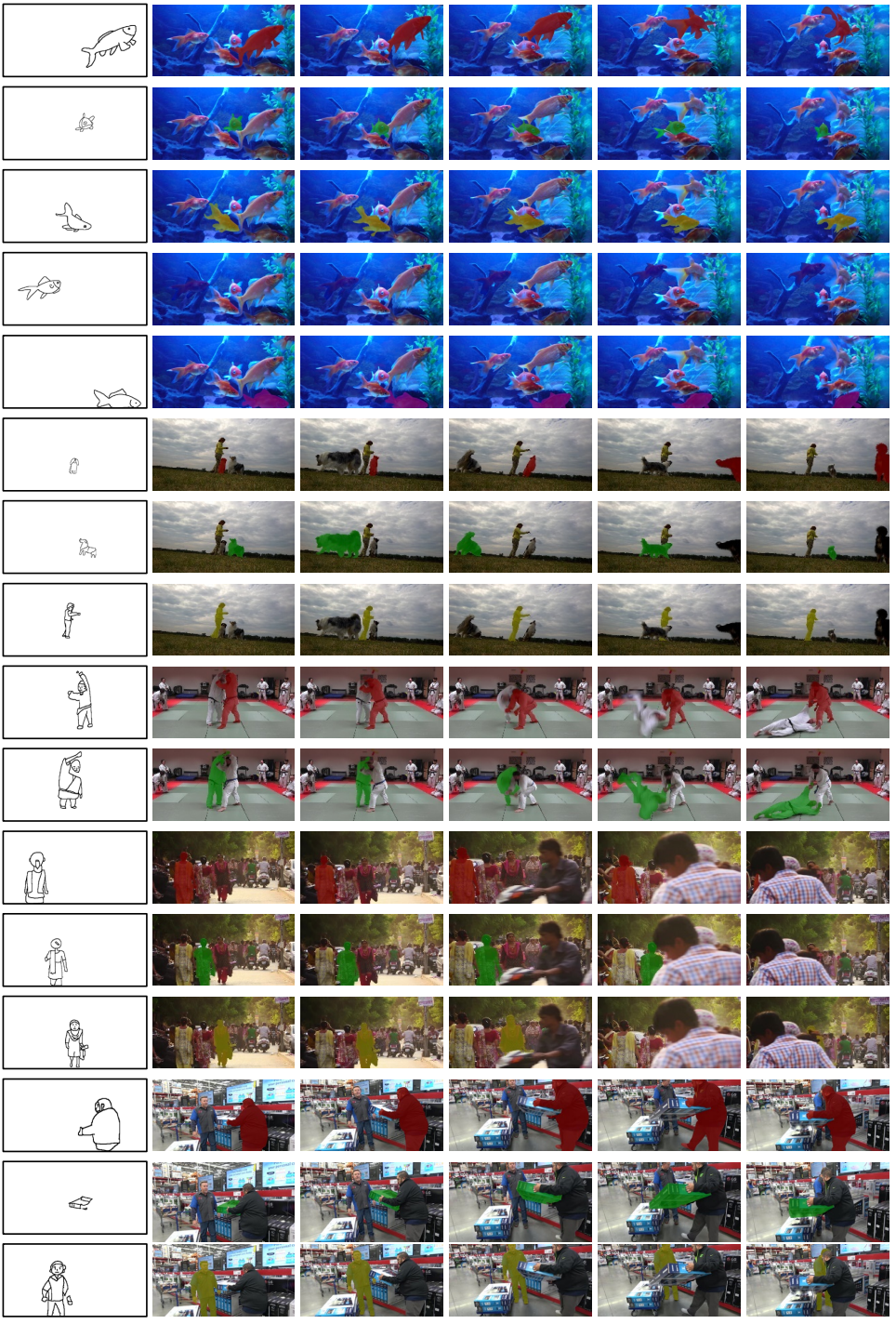

Figure 8: Examples of Sketch and corresponding video frames in our Sketch-DAVIS17 dataset.

Figure 9: Examples of Sketch and corresponding video frames in our Sketch-YouTube-VOS datasset.

	Reference	Train set			Validation set		
		Videos	Objects	#A	Videos	Objects	#A
DAVIS16	Mask	30	30	30	20	20	20
	Text-1st	30	30	60	20	20	40
	Sketch (Ours)	30	30	90	20	20	60
DAVIS17	Mask	60	144	144	30	61	61
	Text-full	60	144	288	30	61	122
	Text-1st	60	144	288	30	61	122
	Sketch (Ours)	60	144	432	30	61	183
YouTube-VOS	Mask	3471	6459	6459	507	1063	1063
	Text-full	3471	6388	12913	507	1063	2096
	Text-1st	3412	6006	10897	507	1030	1993
	Sketch (Ours)	3412	6006	18018	507	1063	3189

Table 4: Dataset statistics: reference type, video numbers, object numbers, and annotation numbers (#A). For mask reference, only annotations in the first frame are taken into account. Text-1st means the language expressions of the first frame. Text-full indicates the language expressions of the full video.