Sketch-based Video Object Segmentation: Benchmark and Analysis

Ruolin Yang[1], Da Li[2], Conghui Hu[3], Timothy Hospedales[2], Honggang Zhang[1], Yi-Zhe Song[2]

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

Introduction

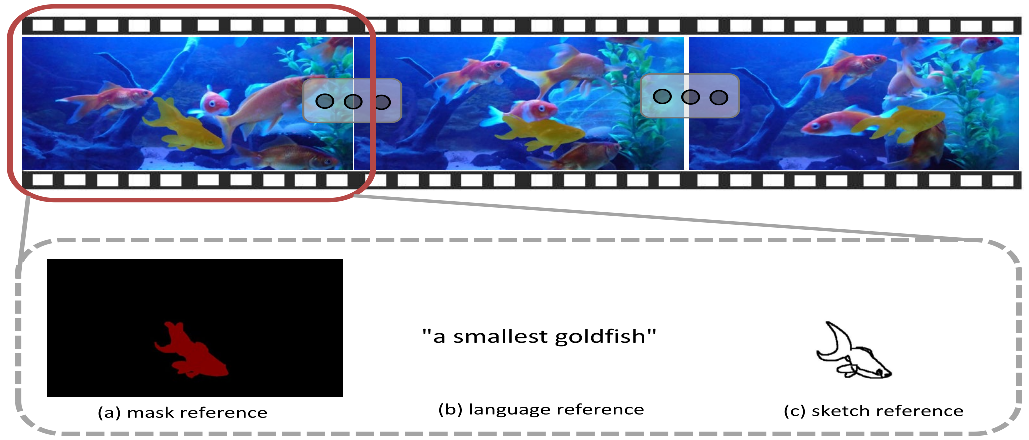

Figure 1: A comparison example between three different annotation types for the Semi-VOS task. (a) Mask reference. (b) Language reference. (c) Sketch reference (Ours).

- **Motivation**: Reference-based video object segmentation, whether language-based or mask-based, faces these challenges:
  - language expressions can sometimes be vague in conveying an intended concept and ambiguous when similar objects in one frame.
  - photo masks are costly to annotate and less practical to provide in a real application.
- **Contribution**: We present the following three key points:
  - A new task of sketch-based video object segmentation, an associated benchmark, and a strong baseline.
  - Our benchmark includes three datasets, Sketch-DAVIS16, Sketch-DAVIS17 and Sketch-YouTube-VOS, which exploit human-drawn sketches as an informative yet low-cost reference for video object segmentation.
  - Experimental results show sketch is more effective yet annotation-efficient than other references, such as photo masks, language and scribble.

Datasets

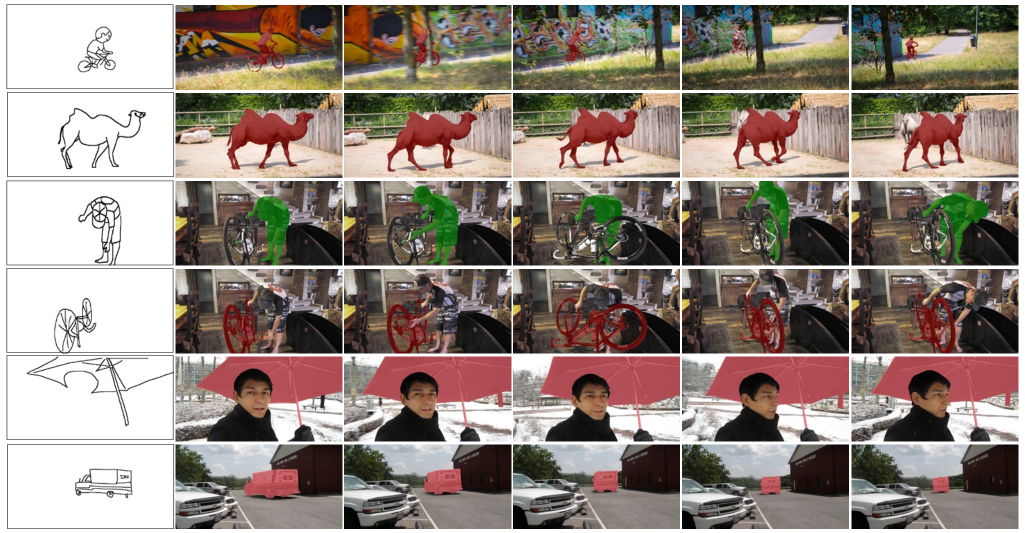

Figure 2: Sketch Reference examples of Sketch-DAVIS16 (row 1&2), Sketch-DAVIS17 (row 3&4) and Sketch-Youtube-VOS dataset (row5&6).

Methodology

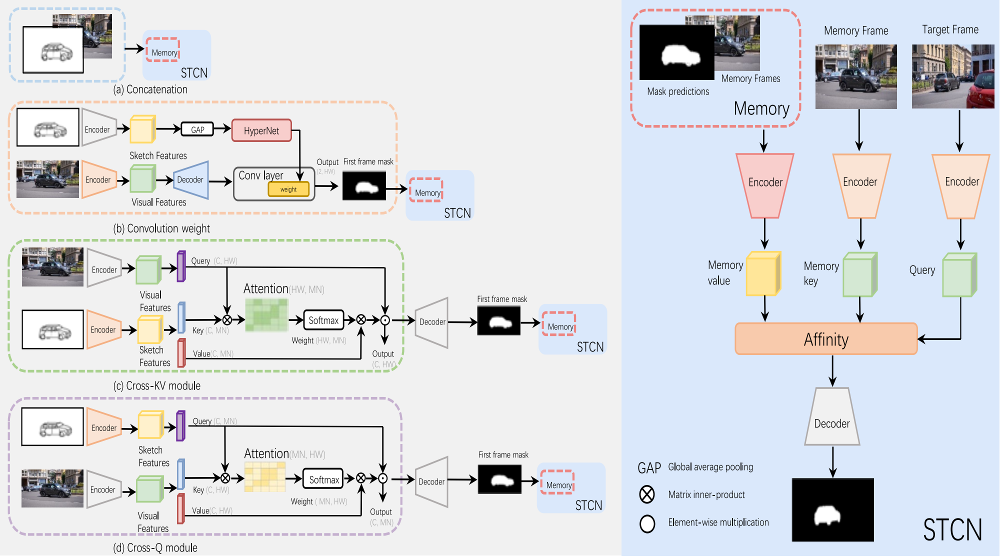

Figure 3: The Sketch-based VOS model with various designs: (a) Concatenation, (b) Convolution weight, (c) Cross-KV, and (d) Cross-Q.

Experiments

Reference	Method	Youtube-VOS			DAVIS17			DAVIS16		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Text	VOSwL[26]	-	-	-	39.3	37.3	41.3	84.1	82.8	85.4
	URVOS[41]	46.5	44.2	48.8	51.7	47.3	56.0	-	-	-
	HINet[52]	-	-	-	52.0	-	-	84.8	84.4	85.3
	YOFO[28]	48.6	47.5	50.0	55.4	50.1	58.7	-	-	-
	MLRL[48]	49.7	48.4	51.0	57.9	53.9	62.0	-	-	-
	LBDT[18]	49.4	48.2	50.6	54.1	-	-	-	-	-
	MTTR[5]	55.3	54.0	56.6	-	-	-	-	-	-
	ReferFormer[49]	64.9	62.8	67.0	61.1	58.1	64.1	-	-	-
Mask	STM[35]	74.7	72.8	76.6	69.5	67.0	72.0	-	-	-
	STCN[9]	79.6	77.1	82.1	74.4	71.5	77.2	-	-	-
Sketch	Ours	75.4	73.4	77.5	70.2	66.9	73.4	81.6	80.2	83.1

Table 1: Comparison with state-of-the-art methods on Youtube-VOS, DAVIS17 and DAVIS16 datasets.

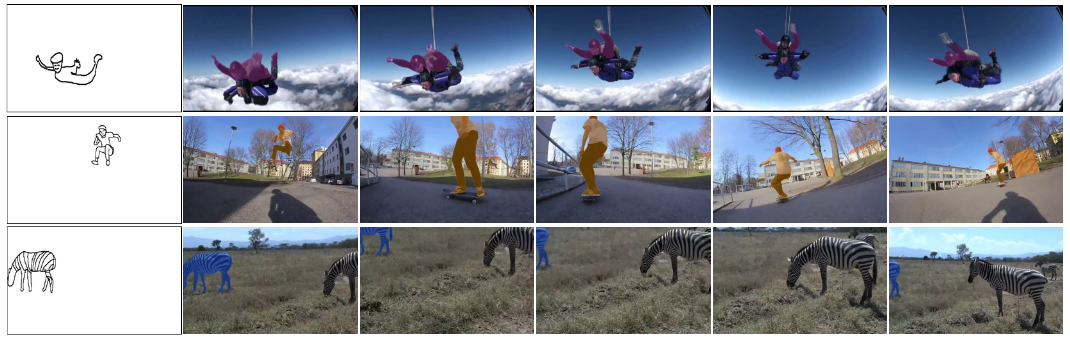

Figure 4: Visual comparison with language-based model on the YouTube-VOS validation set.

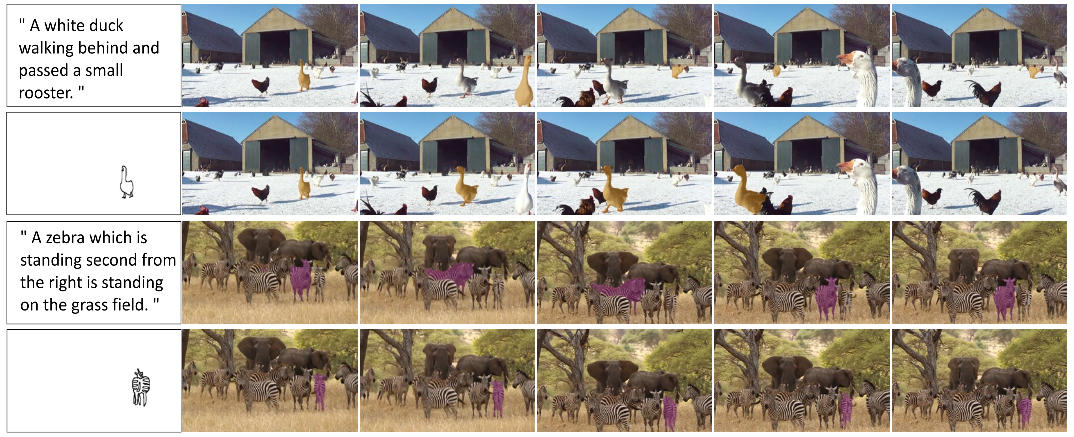
" A white duck walking behind and passed a small rooster. "
" A zebra which is standing second from the right is standing on the grass field. "

Figure 5: Visual comparison with language-based model on the YouTube-VOS validation set.