

## Sketch-based Video Object Segmentation: Benchmark and Analysis

Ruolin Yang<sup>1</sup>, Da Li<sup>2</sup>, Conghui Hu<sup>3</sup>, Timothy Hospedales<sup>2</sup>, Honggang Zhang<sup>1</sup>, Yi-Zhe Song<sup>2</sup>

<sup>1</sup>PRIS, Beijing University of Posts and Telecommunications, <sup>2</sup>SketchX, CVSSP, University of Surrey, <sup>3</sup>Department of Computer Science, National University of Singapore



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.



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.

Reference	Method	Youtube-VOS			DAVIS17			DAVIS16		
		$\mathcal{J}\&\mathcal{F}$	${\mathcal J}$	${\mathcal F}$	$\mathcal{J}\&\mathcal{F}$	${\mathcal J}$	${\mathcal F}$	$\mathcal{J}\&\mathcal{F}$	${\mathcal J}$	
Text	VOSwL[26]	-	-	-	39.3	37.3	41.3	84.1	82.8	5
	URVOS[41]	46.5	44.2	48.8	51.7	47.3	56.0	-	-	
	HINet[52]	-	-	-	52.0	-	-	84.8	84.4	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	/0.0	69.5 74.4	67.0	72.0	-	-	
Sketch	Ours	79.6	73.4	82.1	74.4	66.9	73.4	81.6	80.2	
Sketch	Ours	75.4	73.4	11.5	70.2	00.9	73.4	81.0	80.2	
Table 1: Co	mparison with state	-of-the-	art met	hods o	n Youtuk	be-VOS	, DAVIS	517 and [	DAVIS1	6
datasets.										
		N		N		N				
					3		10			
I.S.	Local 1		1				-		-	1
			Paris	1			-			
			1		ANNER C		AL MIR	All in		4
\$ C	1 i and a	E alle	h		ALC	AL AND A	- M	F Ca		11
48		11 11-11-11-11-11-11-11-11-11-11-11-11-1							The state is a second	2
	A REAL PROPERTY OF					A AVE	AL		TREE	the state
	Les I	a stand				1	1		1 in	1
AN R	and a firmer									
			and the second second		The los					DADA
	D-A	1 Cade Sain	- States		A States				and the second	
	MARKEN STRATT		and the second		a market	123			and the second	
Figure 4. Vi	sual comparison wit	h langu	ane-hao	sed mo	del on th		Tube-V(	DS valida	ation se	t
rigure 4. vi.		in langa	uge bu					oo vanac		
							<b>^</b>			~
" A white duck										
walking behind a	nd			16 Dam	and the second		The second second		1	in a
passed a small								Z Z		
"no octor "	D	13	- Financial Statements	Contraction of		and the states	N		12	7
rooster. "							-	CONTRACTOR OF STREET		
rooster. "										
rooster. "										
rooster. "			Ţ			2				
rooster. "			ŢĿ							
rooster. "										
rooster. "	From									
rooster. "	From g									
rooster. "	From g									
" A zebra which is standing second the right is standi on the grass field	From Contraction of the second s									
rooster. "	From g									







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

## PRIS & SketchX Lab & NUS