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

Abstract

Reference-based video object segmentation is an emerging topic which aims to segment the corresponding target object in each video frame referred by a given reference, such as a language expression or a photo mask. However, language expressions can sometimes be vague in conveying an intended concept and ambiguous when similar objects in one frame are hard to distinguish by language. Meanwhile, photo masks are costly to annotate and less practical to provide in a real application. This paper introduces 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. We take advantage of STCN, a popular baseline of semi-supervised VOS task, and evaluate what the most effective design for incorporating a sketch reference is. Experimental results show sketch is more effective yet annotation-efficient than other references, such as photo masks, language and scribble. The datasets are released at https://github.com/YRlin-12/Sketch-VOS-datasets.

1 Introduction

Video object segmentation (VOS) aims to automatically identify and segment target objects in a given video and has witnessed considerable progress recently. Traditional VOS can be divided into three settings, unsupervised VOS [19, 36, 46, 47, 51], which segments the most salient object; semi-supervised VOS (Semi-VOS) [9, 35], which segments the target object

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

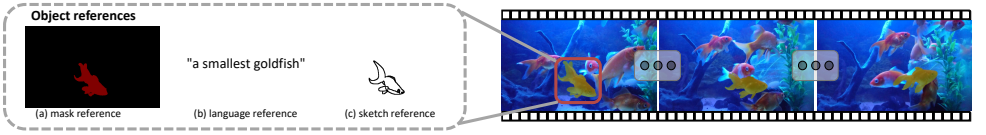

Figure 1: A comparison example between three different annotation types for the Semi-VOS task. **(a)** Mask reference. **(b)** Language reference. **(c)** Sketch reference (Ours).

given a photo mask reference for the first frame and supervised VOS [8, 23, 34, 55], requiring users to interact with the system to refine the output masks repeatedly until the result is satisfactory. However, unsupervised VOS lacks the flexibility to segment an object of interest, while supervised VOS requires intense user interactions. Thus, Semi-VOS is more useful and practical. Semi-VOS methods segment and propagate the object masks given the user annotated masks of the first frame and achieve strong results. However, annotating photo masks for even the first frame only of large-scale datasets like YouTube-VOS [50] for model training is time-consuming and not practical for users to provide in real applications. To overcome the annotation drawback of photo mask-based semi-VOS, referring VOS has been introduced recently [26, 41], which employs the language expressions as a new type of reference to guide object segmentation in VOS. However, despite its general efficacy, it is sometimes challenging to convey some concept effectively with words.

In recent years, sketch, as a complementary modality to text, has been investigated broadly due to the demand for interacting tools on popular touchscreen devices. The common view is that sketch is preferable when words are inconvenient to convey an intended concept. At the same time, the fact that the time necessary to create one sketch (54.84 secs) is less than that for one photo segmentation mask (109.01 secs) [25] demonstrates that drawing sketches is a more time-efficient user annotation. Sketches contain category and fine-grained level information verified by various sketch-based image retrieval works [1, 2, 9, 13, 15, 17, 40]. While sketch-based image editing works [45, 58] also demonstrate the expressiveness and editability of sketches. Sketches can be used as a query to classify, segment or locate novel objects in an input image, as shown in Sketch-a-Classifier [24], Sketch-a-Segmenter [25] and sketch guided localization (SGL) [42], respectively. However, the existing sketch-based benchmarks have only considered sketches for image-level tasks, giving less consideration to the video domain.

In this paper, we propose sketch-based video object segmentation and introduce a new task, Sketch-VOS, to predict photo masks in video frames by sketch references given for the first frames. Our three Sketch-VOS datasets are extended from DAVIS16 [37], DAVIS17 [38] and YouTube-VOS [50] datasets with free-hand sketches following the rules of Ref-DAVIS [26] and Ref-YouTube [41] for fair comparisons to the existing works. In Figure 1, we provide a comparison example between three different types of references, photo mask, language expression and sketch. As shown in Figure 1 (b), language expressions may be ambiguous in situations where objects are similar. Whereas fine-grained information like the shape and pose of a drawn sketch can easily overcome this limitation. As the first dataset that pairs sketches with video objects, our dataset will allow researchers to develop more annotation-efficient algorithms for video object segmentation and potentially other associated problems like video editing.

Our paper is structured to address three questions: (i) What is sketch-based video object segmentation? We will detail how we construct our three datasets for sketch-based VOS and how to use them for model evaluation. (iii) Why is sketch a better reference for VOS than photo mask, language, and scribble? We will compare our sketch based STCN with other

Figure 2: Left: Sketch reference examples of Sketch-DAVIS16 (row 1), Sketch-DAVIS17 (row 2) and Sketch-Youtube-VOS dataset (row 3). Right: examples of references in our Sketch-VOS benchmark.

SOTA VOS methods using other references.

2 Related Work

(Photo Mask-based) Semi-supervised Video Object Segmentation. Early semi-supervised video object segmentation focused on fine-tuning at test time [6, 10, 11, 32] or matching and propagating [9, 12, 33, 35, 44, 54] the pixel-level object(s) mask(s) of the first frame. The latter one is more efficient and usually trained end to end. STM [35] made remarkable progress by using space-time memory bank and became the backbone of many following state-of-art methods [9, 30, 31, 32]. STCN [9] improved the affinity and feature extraction of STM and reached a more effective and efficient version. In this work, we focus on modifying and extending STCN for experimentation and analysis due to its robustness of temporal coherence.

(Language-based) Referring Video Object Segmentation. Recently, referring video object segmentation (RVOS) has attracted great attention from researchers. This task aims to segment and track the mask of the target object in a video referred by a language expression. [26] released Ref-DAVIS dataset at first place with 90 videos and employed a complicated model which located object by bounding box first and then propagated to predict the mask. Then, URVOS [41] provided a large-scale referring video object segmentation dataset (Ref-YouTube-VOS) and introduced an STM-style model with cross-modal attention block to fuse frame feature and text feature. ReferFormer [49] achieved state-of-the-art results of RVOS by a top-down method with the support of the popular detector Deformable-DETR [59]. Language expression as the Transformer decoder input is the key component of this approach and Hungarian matching [27] is required for linking the instance tube. However, it relies on high computation resources due to its complexity.

Sketch as Queries. Sketch-based image retrieval is a fundamental task that aims to retrieve photos of the same category [15, 17, 39] or corresponding instance [1, 2, 4, 13, 40] given a query sketch. Sketch-a-Classifier [24] designed a model to generate a photo classifier by giving a sketch of an unseen category. Sketch-a-Segmenter [25] similarly were made to produce a novel pixel-level classifier by a sketch input. SGL [42] proposed to generate object proposals relevant to the sketch query. DIY [3] employed sketch queries to achieve the goal of few-shot class incremental learning. These methods mainly focused on image-level tasks and paid less attention to video-sketch correspondence. To the best of our knowledge, this is the first work to apply sketch to the video object segmentation task.

3 Sketch-based VOS Benchmark

We extend three popular VOS benchmarks including DAVIS16 [37], DAVIS17[38] and YouTube-VOS [50] with first-frame sketch annotations for segmenting target objects in video sequences. Examples are illustrated in Figure 2 (left). A detailed comparison of the datasets is given in the supplementary file.

Data Collection and Pre-Processing The sketch data is collected by a collection interface following FSCOCO [14] dataset. Similar to the language reference [26], we asked the participants to sketch for the target objects appearing in the first frame without seeing the full video. We provided each participant with a randomly selected object, as well as a blank canvas to sketch on. The participants had 60 seconds to remember as many details as they could about the pose, shape, and fine-grained characteristics, before the object is removed. The intention is to create a sketch that will make it possible for those who have never seen the video to identify the target object. To verify the quality of our sketch, we requested assistance from 20 volunteers to validate our dataset. Each of them would be provided a video and a sketch corresponding to one object. At the beginning, the first frame would last for a while and the video would play, volunteers then used the bounding box to label the object in the video. The final step is to re-draw references for cases where manual VOS above failed. To keep the diversity of sketch, we did not train any participants and asked them to draw in their own style. As shown in Figure 2 (right), the participants all drew the sketch in different styles, but salient visual properties (e.g., pose) of each object were uniformly depicted. On average each object has been annotated with three sketches and it takes the annotator around 30s to draw for a target object. One big challenge of VOS is there are many similar-looking instances in one video as shown in Figure 1. And according to Sketch-a-Segmenter [25], sketches with position and scale alignment will benefit segmenting instance-level objects. Therefore, we subject all sketches to this preprocessing strategy before our experimental evaluation.

Sketch-DAVIS-VOS DAVIS16 [37] is a dataset where only a single object is annotated per frame, which comprised of 30 training and 20 validation videos from four evenly dispersed classes (humans, animals, vehicles, objects) with all the frames annotated with pixel-level accuracy. Then DAVIS17 [38] extended DAVIS16 to 60 training and 30 validation videos and annotated multiple objects. Ref-DAVIS [26] extended DAVIS datasets with two referring expression based on first frame only as well as full video. The latter one is different from mask annotation where the annotator will describe the object after viewing the full video. We only provide first frame annotations same as the mask annotations. To the end, we collect 150 sketches for DAVIS16, and 615 sketches for DAVIS17.

Sketch-YouTube-VOS YouTube-VOS[50] is a large-scale multi-object VOS benchmark consisting of 3417 training videos of 65 categories and 507 validation videos of 65 training classes and 26 unseen categories. Ref-YouTube-VOS dataset [41] contains the language expressions for two type – first frame and full video. However, some objects can not be identified by words. Therefore, Ref-YouTube-VOS only takes partial objects from YouTube-VOS to annotate. Though sketch can be used to refer any objects in frames, we only draw sketches for objects appeared in Ref-YouTube-VOS for fair comparison. Note that only the ground-truth masks of YouTube-VOS training set is publicly available. The validation set results can be only evaluated on the competition server. And only the evaluation entry for the full-video language expressions is currently open on the server. Therefore, we draw sketches for the training set based on the language expressions of first frame and validation set based on the expressions of full video.

4 Sketch-based VOS Model

This section provides an overview of how sketch can be used as a new type of reference in a reference-based VOS model. We employ the popular Semi-VOS method STCN [9] as our baseline and explore the interaction between two modalities – sketch and video. We

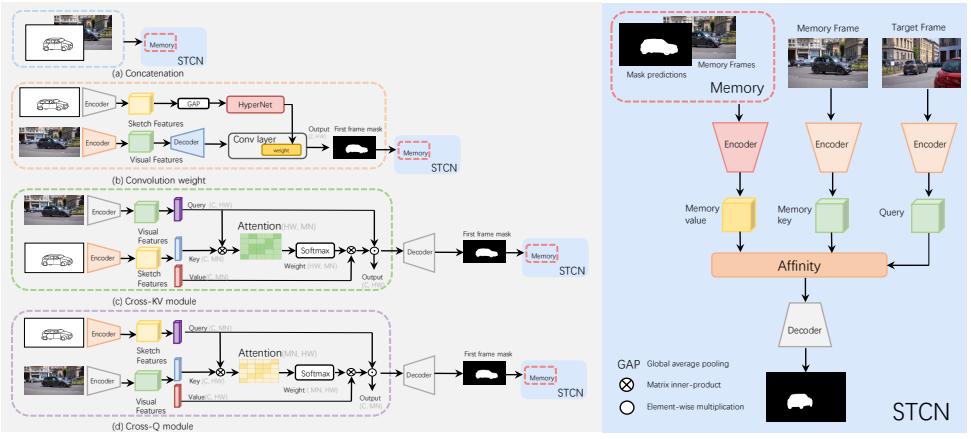

Figure 3: The Sketch-based VOS model with various designs: (a) Concatenation, (b) Convolution weight, (c) Cross-KV, and (d) Cross-Q.

extend STCN with various fusion designs, such as input fusion, latent fusion and sketch-based weight generation.

4.1 STCN

We first explain how STCN works before introducing our model. As shown in the right part of Figure 3, STCN is a memory-based method which encodes frames to keys and queries and encodes masks (concatenated with frames) to values. Every time the feature of a target frame is updated by an affinity between query features and memory key/value features in the memory bank. Then, features are gradually processed and upsampled by the decoder.

4.2 Design-Space of Sketch-based VOS

Concatenation The simplest way to combine sketch and video frames is by concatenating them at the input level. As shown in Figure 3 (a), we encode sketch and frame together as the memory value instead of mask in STCN.

Convolution Weight: Inspired by Sketch-a-Segmenter[25], we use a HyperNet[21] to generate an instance-level weight for the video segmentation head to predict object mask in the first frames as shown in Figure 3 (b). Then we store the prediction as memory and propagate it through the video. More clearly, Sketch-a-Segmenter predicts masks for all instances given a sketch, while our task is to segment one particular instance from all at a time.

Cross Attention Given the first frame of a video input and a sketch reference, their features are generated by a visual encoder and a sketch encoder separately as follows: $\mathcal{F} \in \mathbb{R}^{C_v \times H \times W}$, $\mathcal{S} \in \mathbb{R}^{C_s \times M \times N}$, where H, W, M, N are the spatial dimensions, C_v and C_s are the channel dimensions. We construct two cross-modal attention design strategies to fuse the features. Figure 3 (c) illustrates our first attention module, motivated by ReferFormer [49] and LAVT [53], we encode sketch as Key and Value by two 1×1 convolution mappings. And to match the dimension of sketch feature, we encode visual feature as Query by another 1×1 convolution filter. The outputs are Key $\mathcal{K}_s \in \mathbb{R}^{C \times MN}$, Value $\mathcal{V}_s \in \mathbb{R}^{C \times MN}$ and Query $\mathcal{Q}_f \in \mathbb{R}^{C \times HW}$. Then, a dot-product attention [43] is computed between Query and Key. The attention map stores the correspondences between the visual feature and the sketch reference.

Then Value will be transformed by the attention map. We called this module Cross-KV. The output features $\mathcal{O} \in \mathbb{R}^{C \times HW}$ can be obtained as follows:

$$\mathrm{W} = \mathrm{Softmax}(\frac{{Q_f}^{\mathrm{T}} K_s}{\sqrt{C}}), \tag{1}$$

$$\mathrm{O} = V_s W^{\mathrm{T}} \odot Q_f, \tag{2}$$

where $\odot$ denotes element-wise multiplication, which was introduced by [53] and can be explored with other options. We also provide a novel cross-modal attention module as shown in Figure 3 (d). Since sketch is the reference condition that leads a model to segment a target mask, we use a sketch feature as the Query in cross-modal attention. Similar to the above, we gain the new Key $\mathcal{K}_f \in \mathbb{R}^{C \times HW}$, Value $\mathcal{V}_f \in \mathbb{R}^{C \times HW}$ and Query $\mathcal{Q}_s \in \mathbb{R}^{C \times MN}$. This time we compute the attention map between Query from the sketch feature and Key from the visual feature. This Attention map is simply the transposed matrix of the attention map mentioned above, but the Value now is from the visual feature. We called this design Cross-Q. The output features $\mathcal{O} \in \mathbb{R}^{C \times MN}$ can be obtained as follows:

$$\mathrm{W} = \mathrm{Softmax}(\frac{{Q_s}^{\mathrm{T}} K_f}{\sqrt{C}}), \tag{3}$$

$$\mathrm{O} = V_f W^{\mathrm{T}} \odot Q_s, \tag{4}$$

The output features are fed into the STCN decoder which generates a binary mask of the first frame and then propagates this mask by the memory bank to segment the remaining target object masks in the video.

5 Experiments

We evaluate the performance of our Sketch-based VOS model and compare sketch with other references, such as photo mask, language expression and scribble.

5.1 Experimental Setup

Evaluation Metrics: We evaluate our reuslts by the standard evaluation metrics [37] for VOS tasks, i.e., region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and the average of $\mathcal{J}$ and $\mathcal{F}$ ($\mathcal{J}\&\mathcal{F}$). For DAVIS dataset, we evaluate by the official evaluation code [1]. All experiments on YouTube datasets are evaluated on the competition server [2] same as Ref-YouTube-VOS method [49]. Note that only 202 videos in validation set can be evaluated on the server. All the following results are based on these 202 videos.

Implementation Details: Following STCN [9], every video frame and corresponding sketch are downscaled to 384p. We train our model using the Adam optimizer with initial learning rate of 1e-5. The frame encoder is initialized with classification weights pre-trained on ImageNet[16] while sketch encoder is initialized with classification weights pre-trained on QuickDraw [20]. Different from STCN, we pick first frame as a default frame and randomly sample other two temporal frames to form a training clip. We use BCE Loss for all experiments.

[1] https://github.com/davisvideochallenge/davis2017-evaluation
[2] https://youtube-vos.org/dataset/rvos/

Table 1: Sketch-based VOS Model evaluated on Sketch-YouTube-VOS validation set. GAP indicates global average pooling. Decoder means the convolution layer of segmentation head of Decoder. HyperNet means the weight generated by HyperNet.

Fusion Designs	Level		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
	visual	sketch						
(a) Concatenation	raw pixel	raw pixel	74.1	72.1	76.0	-	-	-
(b) Convolution weight	Decoder	HyperNet	19.6	20.1	19.0	-	-	-
			Cross-KV			Cross-Q		
(c) Cross-KV/Q	res5	GAP	19.6	16.2	23.1	55.1	54.1	56.1
	res4	res4	56.6	55.5	57.6	74.3	72.3	76.3
	res5	res5	67.1	65.4	68.7	74.8	72.8	76.8
	multi-level	res4	57.3	55.9	58.6	74.7	72.7	76.7
	multi-level	res5	55.8	54.4	57.1	**74.9**	**72.9**	**77.0**

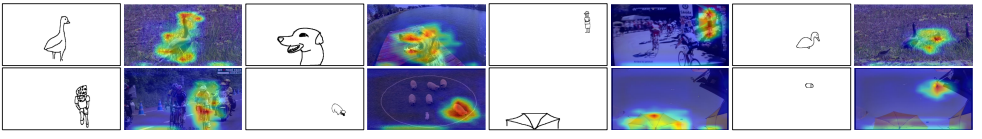

Figure 4: Visualized sketch queries and corresponding feature maps weighted by attention on Sketch-YouTube-VOS validation set.

5.2 Sketch-based VOS models Results

We encode sketches and video frames by two separate ResNet50 [22]. We first conduct the ablation study of various designs proposed in 4.2.

For Cross-KV and Cross-Q, we also tried different levels of features and different variants of interacting sketch reference and video frames as follows: *GAP sketch features* – Given the output from layer4 of the sketch encoder, we aggregate global information of a sketch using global average pooling (GAP). This gives a reference to verify the importance of the spatial information embedded in sketch features. *Spatial sketch features* – Namely, we use the spatial feature maps generated from the sketch encoder, which retains the spatial information. *Multi-level visual features* – i.e. features from multiple layers of a video encoder.

Table 1 reports the results for various sketch-based VOS models. We can see that designs using cross attention and input concatenation all give reasonable results and the best one boosts the performance to $\mathcal{J}\&\mathcal{F}$ of 74.9, $\mathcal{J}$ of 72.9 and $\mathcal{F}$ of 77.0, whereas sketch-based weights generation does not work in this task. This may be the reason that, in this case, sketch reference is interacting less with video frames and the reference information is hard to propagate among video frames. Among cross-attention variants, Cross-Q works much better than Cross-KV and improves the $\mathcal{J}\&\mathcal{F}$ by more than around 7 absolute points, which is interestingly different from the existing wisdom from [49, 53], indicating sketch as a reference may require different designs than language. Without a surprise, using GAP features works badly and indicates the spatial information provided by sketch is crucial for the VOS task. Using multi-level visual features improves the performance marginally, which may not be favoured considering the extra computational cost it brings.

We also visualise the generated attention maps by querying the first frame from various videos of Sketch-YouTube-VOS validation set using sketch references. As illustrated in Figure 4, we can see that the VOS model can attend to the indicated areas by the sketch references precisely regardless the categories and scales of the target objects. Even an object

Table 2: Comparison with state-of-the-art methods on Youtube-VOS, DAVIS17 and DAVIS16 datasets.

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[6]	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

which is too small to be described by language can be localized precisely by sketch.

In summary, using sketch features as Query in the cross-modal attention with a spatial sketch feature works effectively on the video object segmentation task.

5.3 Sketch v.s. Other References

We compare our Sketch-VOS with the existing VOS works incorporated with different references that appeared in the literature.Furthermore, we conduct comprehensive comparisons between several references: language, sketch, mask, and interactive annotations including cross, circle, scribble and object contour.

Baselines: (i) State-of-the-art methods: VOSwL [26] predicts masks by localising and segmenting in two stages. URVOS [41] is similar to our design but they concatenate visual and linguistic features before feeding into the cross-modal attention. YOFO [28] transfers object information by meta-learning. HINet[52] employs a hierarchical fusion of language and frame. MLRL [48] fuses linguistic features with video, frame and object features in different levels. LBDT [18] transfers spatial and temporal visual features by language. MTTR [6] and ReferFormer [49] use transformer-based detectors to localise and segment masks. STM [35] designs a spatial-temporal memory bank, then STCN [9] improves it with a more effective affinity module. (ii) Fair comparison baselines: All experiments use STCN as the backbone and follow the same implementation setting in Sec. 5. The fusion methods vary in different modalities. Since YouTube-VOS dataset does not collect scribbles as annotations, we extend YouTube-VOS dataset with scribbles in the first frame by following [7]. We simply concatenate it with the first frame before feeding it into STCN following [8]. As for language expression, we encode first-frame expressions by the popular language encoder BERT [43] initialized by the official pre-trained weights. We utilize Cross-KV module to fuse the linguistic features and frame figures following language-based methods [49, 53]. We did not pre-train the mask-based STCN on extra static image datasets for a fair comparison. (iii)Ablation Study of different references: To verify the effectiveness of sketch, we further compared sketch with text+click, text+bounding box, cross, circle and contour. We extend YouTube-VOS dataset in the first frame with theses interactive annotations. Specifically, the cross and the click are drawn on the center point of the ground-truth mask; the circle and the box are obtained by fitting an outer circle/box to the ground-truth mask; the contour is acquired by computing the convex hull of the sketch. All experiments use STCN as the backbone and follow the same implementation setting in Sec. 5.

Table 3: Fair comparison with different references on YouTube-VOS validation dataset. The rightmost column shows the average time of annotating one object.

Reference	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	Annotating time
Text	44.4	42.7	46.1	5.0s[26]
Scribble	69.1	67.3	71.0	1.25s [29]
Sketch	75.4	73.4	77.5	30.6s
Mask	79.6	77.1	82.1	109.0s[25]

Table 4: Ablation study of different references on YouTube-VOS validation dataset.

Reference	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Text	44.4	42.7	46.1
Text+Click	57.0	54.4	59.6
Text+Box	58.0	55.3	60.6
Cross	56.1	53.2	59.1
Circle	58.6	56.6	60.6
Contour	71.8	69.6	74.1
Sketch	75.4	73.4	77.5

YouTube-VOS. Table 2 shows the results of the recent state-of-the-art video segmentation methods on the YouTube-VOS validation set. Our best model achieves a competitive $\mathcal{J}\&\mathcal{F}$ among all competitors. Nevertheless, our model outperforms all language-based methods under all metrics and by significant margins. For a fair comparison with mask-based methods, we retrain the STM [35] and STCN [9] without any extended image datasets. Our model demonstrates better performance compared to STM, albeit slightly trailing STCN.

DAVIS. In Table 2, we evaluate our model on the DAVIS17 validation set. Due to its small scale, we directly evaluate this dataset using models trained on YouTube-VOS. However, DAVIS has longer videos and is annotated more strictly than YouTube-VOS which is more challenging for VOS. The results show that our model can easily generalize to another dataset, and again our method outperforms all language-based methods without using extra image datasets. Compared to the mask-based methods, we outperform STM by 0.7 points and underperform STCN by 4.2 points in terms of $\mathcal{J}\&\mathcal{F}$ metrics. Table 2 shows the results on DAVIS16. Competitors are limited in this case, including only VOSwL [26] and HINet [52], nevertheless our sketch-VOS has a close performance to these two mask-based VOS methods. Please see more fine-tuning results on DAVIS datasets in the supplementary.

Visualizations. Figure 5 visualizes results on Sketch-YouTube-VOS dataset. Our model can successfully segment the object mask in each frame given a sketch reference. Our model is robust even in situations with multiple similar objects, appearance changing, fast motion and outside the frame. More visualization of our sketch-VOS results on Sketch-DAVIS datasets can be found in the supplementary. In Figure 6, we visualize the comparison between the state-of-art language referring VOS method ReferFormer and our method. ReferFormer loses track of the target object as there are many similar objects in the first video. However, our model can track and segment the referred duck consistently. In the more difficult second video, ReferFormer can not distinguish the target object at all, while our sketch-VOS conducts the perfect segmentation due to the embedded distinctive fine-grained information.

Fair comparison between different references. Table 3 shows the fair comparison of different references. We can see in such case, sketch outperforms text by more than 20 $\mathcal{J}\&\mathcal{F}$ points. Scribble also works well in this setting but still does not achieve comparable performance to sketch, with a margin of 6.3 $\mathcal{J}\&\mathcal{F}$. Without pre-training on huge static image datasets, mask reference only achieves 79.6 of $\mathcal{J}\&\mathcal{F}$. All these results suggest that sketch can be effectively used as an alternative and cheaper reference for video object segmentation. Additional visual comparisons are provided in the supplementary.

Ablation Study of different references From the results in Tab. 4, sketch can brings more benefits than simple indicator like cross, circle, box or click. This is because they cannot provide additional information beyond object location, such as semantic context or pose. We can also see that only keeping the contour performs worse than the whole sketch. Our

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

Figure 6: Visual comparison with ReferFormer [49] on the YouTube-VOS validation set. Best viewed in color.

speculation is that the inclusion of fine-grained details in the sketch aids in effectively representing and segmenting the object, whereas relying solely on contours may cause confusion in subsequent video frames.

6 Limitations and Future Directions

Despite the fact that Sketch-VOS datasets are the largest public datasets for sketch-based video object segmentation to date, they are still smaller than the standard large-scale benchmarks. However, a sketch is much cheaper to collect than a photo mask and can be manipulated with relative ease to generate variants for data augmentation [56, 57].

In the future, we will investigate ways for combining motion information with sketches to refer to dynamic object activity. We also plan to increase dataset diversity by generative models and other data augmentation techniques.

7 Conclusion

We introduced three instance-level datasets for sketch-based video object segmentation. We evaluate our datasets by extending the popular VOS method STCN and explore various fusion designs for better aggregating sketch and visual features. The experimental results show that our method can easily beat language-referring VOS methods and is comparable to mask-based VOS methods. We hope our proposed datasets will drive future research in the field and inspire people to see the potential of sketch for solving complex video tasks.

References

[1] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*, 2020.

[2] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021.

[3] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Doodle it yourself: Class incremental learning by drawing a few sketches. In *CVPR*, 2022.

[4] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *CVPR*, 2022.

[5] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, 2022.

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, 2017.

[7] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018.

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, 2021.

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 2021.

[10] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, 2017.

[11] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *CVPR*, 2018.

[12] Suhwan Cho, Heansung Lee, Minhyeok Lee, Chaewon Park, Sungjun Jang, Minjung Kim, and Sangyoun Lee. Tackling background distraction in video object segmentation. In *ECCV*, 2022.

[13] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially does it: Towards scene-level fg-sbir with partial input. In *CVPR*, 2022.

[14] Pinaki Nath Chowdhury, Aneeshan Sain, Ayan Kumar Bhunia, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Fs-coco: Towards understanding of freehand sketches of common objects in context. In *ECCV*, 2022.

[15] John Collomosse, Tu Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[17] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[18] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *CVPR*, 2022.

[19] Shreyank N Gowda, Panagiotis Eustratiadis, Timothy Hospedales, and Laura Sevilla-Lara. Alba: Reinforcement learning for video object segmentation. *BMVC*, 2020.

[20] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[21] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[23] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *CVPR*, 2021.

[24] Conghui Hu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-classifier: Sketch-based photo classifier generation. In *CVPR*, 2018.

[25] Conghui Hu, Da Li, Yongxin Yang, Timothy M Hospedales, and Yi-Zhe Song. Sketch-a-segmenter: Sketch-based photo segmenter generation. *IEEE transactions on image processing*, 2020.

[26] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018.

[27] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955.

[28] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *AAAI*, 2022.

[29] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[30] Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *CVPR*, 2022.

[31] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, 2022.

[32] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *ECCV*, 2022.

[33] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *ICCV*, 2021.

[34] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *CVPR*, 2020.

[35] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019.

[36] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *ECCV*, 2022.

[37] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[38] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[39] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *CVPR*, 2020.

[40] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021.

[41] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*.

[42] Aditay Tripathi, Rajath R Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *ECCV*, 2020.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[44] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.

[45] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own gan. In *ICCV*, 2021.

[46] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019.

[47] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, 2019.

[48] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. In *CVPR*, 2022.

[49] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022.

[50] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 2018.

[51] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019.

[52] Zhao Yang, Yansong Tang, Luca Bertinetto, Hengshuang Zhao, and Philip HS Torr. Hierarchical interaction network for video object segmentation from referring expressions. In *BMVC*, 2021.

[53] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022.

[54] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[55] Zhaoyuan Yin, Jia Zheng, Weixin Luo, Shenhan Qian, Hanling Zhang, and Shenghua Gao. Learning to recommend frame for interactive video object segmentation in the wild. In *CVPR*, 2021.

[56] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Sketch-a-net that beats humans. 2015.

[57] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[58] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. In *CVPR*, 2022.

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.