Neurocomputing 453 (2021) 754–765

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weakly supervised video object segmentation initialized with referring expression ☆

XiaoQing Bu [a], YuKuan Sun [b], JianMing Wang [c,d,*], KunLiang Liu [e], JiaYu Liang [e], GuangHao Jin [e], Tae-Sun Chung [f]

[a] School of Electronic and Information Engineering, Tiangong University, Tianjin, China
[b] Center for Engineering Internship and Training, China
[c] Tianjin International Joint Research and Development Center of Autonomous Intelligence Technology and Systems, Tiangong University, Tianjin, China
[d] Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tiangong University, Tianjin, China
[e] School of Computer Science and Technology, Tiangong University, Tianjin, China
[f] Department of Software, Ajou University, Suwon, South Korea

ARTICLE INFO

Article history:
Received 11 February 2020
Revised 2 June 2020
Accepted 5 June 2020
Available online 9 September 2020
Communicated by Derui Ding

Keywords:
Video Object Segmentation
Referring Expression
Natural Language Processing

ABSTRACT

With the aid of one manually annotated frame, One-Shot Video Object Segmentation (OSVOS) uses a CNN architecture to tackle the problem of semi-supervised video object segmentation (VOS). However, annotating a pixel-level segmentation mask is expensive and time-consuming. To alleviate the problem, we explore a language interactive way of initializing semi-supervised VOS and run the semi-supervised methods into a weakly supervised mode. Our contributions are two folds: (i) we propose a variant of OSVOS initialized with referring expressions (REVOS), which locates a target object by maximizing the matching score between all the candidates and the referring expression; (ii) segmentation performance of semi-supervised VOS methods varies dramatically when selecting different frames for annotation. We present a strategy of the best annotation frame selection by using image similarity measurement. Meanwhile, we first to propose a multiple frame annotation selection strategy for initialization of semi-supervised VOS with more than one annotated frames. Finally we evaluate our method on DAVIS-2016 dataset, and experimental results show that REVOS achieves similar performance (79.94% measured by average IoU) compared with OSVOS (80.1%). Although current REVOS implementation is specific to the method of one-shot video object segmentation, it can be more widely applicable to other semi-supervised VOS methods.

1. Introduction

The rapid development of intelligent mobile terminals has led to an exponential increase in video data. In order to effectively analyze and use video big data, it is very urgent to automatically segment and track objects of interest in videos. Video object segmentation (VOS) and tracking are two basic and highly related tasks in the field of computer vision. Object segmentation divides pixels of video frames into two subsets (foreground target and the background region) and generates object segmentation masks. Object tracking is to determine the exact locations of targets in video images and generates object bounding boxes. These two topics are facing some common challenges, such as deformation, motion blur, and scale variation. Meanwhile, the former also has to deal with the problems of heterogeneous object, interacting object, edge ambiguity, and shape complexity. The latter suffers from difficulties in handling occlusion, fast motion, out-of-view and realtime processing.

Early non-learning methods typically address VOS task using handcrafted features. More recently, research of VOS has turned towards deep learning paradigms following the success of deep learning in many computer vision applications. One-Shot Video Object Segmentation (OSVOS) is a deep learning framework of semi-supervised video object segmentation, which processes each video frame independently for segmenting a particular object instance given a manually annotated video frame (one-shot)[1].

OSVOS formulates video object segmentation as a per-frame segmentation problem, and this stands in contrast to approaches

☆ This study is funded by The Tianjin Science and Technology Program (19PTZWHZ00020) and National Natural Science Foundation of China (Grant No. 61902281).

* Corresponding author at: No. 399, Bin-shui-xi Road, Xiqing District, Tianjin, China.
E-mail address: wangjianming@tiangong.edu.cn (J. Wang).

https://doi.org/10.1016/j.neucom.2020.06.129
0925-2312/

where temporal consistency plays the central role by assuming that objects do not change too much between one frame and the next (as in object tracking). The authors argue OSVOS has some advantages when processing each frame independently: (1) it is able to segment objects through occlusions; (2) it is not limited to certain ranges of motion; (3) it does not need to process frames sequentially, and errors are not temporally propagated.

As other semi-supervised VOS methods, OSVOS requires a pixel-level annotation of initialize the algorithm. However, annotating a precise segmentation mask is expensive and time-consuming, and this requirement often suffers from criticism when the semi-supervised methods are applied in real applications[2–4].

In this paper we consider the scenario where a user observes a video clip firstly and then specifies an object for segmentation. To alleviate the problem of object mask annotation, we propose a variant of OSVOS initialized with referring expressions (REVOS) and make the semi-supervised method working in a weakly supervised mode. Our contributions are concluded as:

(1) Generally speaking, people often select objects which draw their attention for segmentation. In the community of unsupervised VOS, [5] conducts a systematic study on the role of visual attention for video object segmentation task and shows a strong correlation between human attention and explicit primary object judgments during dynamic, task-driven viewing. Inspired by their observation, we suggest to interpret user's visual attention to VOS system with language interaction and initialize OSVOS with referring expressions. The idea is illustrated in Fig. 1.
(2) The current semi-supervised paradigms for VOS tasks always chooses the first frame as the user-annotated frame. However, [6] has proved that segmentation performance across the entire video varies dramatically when selecting different frames for annotation and the best frame for user annotation is seldom the first frame. The authors introduce a novel deep sorting network (BubbleNets) to select frames using a performance-based loss function. However, the loss function needs annotated frames to calculate performance labels, and this is not feasible in real applications. In the paper, we propose an annotation frame selection strategy by measuring the image similarity between video frames.
(3) Users can easily annotate frames with referring expressions, so REVOS needs a strategy to select multiple annotated frames. To the best of our knowledge, there has been no report on multiple-frame annotation methods in literature. In the paper, we first propose a strategy of multiple frame annotation selection which optimizes the label propagation from labeled data to unlabeled data.

Compared with the conference version, this paper makes the following extensions: (a) Three rules of referring expression generation are proposed, which normalize the way to generate the referring expression and reduce the difficulty of language analysis. (b) The problem how to select the best user-annotated frame is explored. To alleviate the limitation of BubbleNets [6], a novel method is proposed which is based on image similarity measurement. (c) An optimization strategy is proposed to carry out multiple user-annotated frame selection. (d) A variant of foreground branch loss function is derived and is utilized to train test network with multiple annotated frames. (e) We also conduct more comprehensive evaluations and analysis for the best user-annotated frame selection, multiple user annotated frame selection and rules of referring expressions generation.

The structure of the paper is organized as follows: Section 1 introduces the problems of OSVOS methods and proposes our method to solve the problem with language interaction. In Section 2, related work to REVOS is described. Section 3 deals with the framework of REVOS and how it works in details. Section 4 depicts the experimental results on DAVIS dataset. Section 5 presents limitations and future work of our method. Finally, we conclude our work in Section 6.

2. Related work

In the section, we provide a brief overview of recent work in three relevant fields: video object segmentation, annotation frame selection for semi-supervised VOS and referring expression comprehension.

2.1. Video object segmentation

According to the level of supervision, We categorize the video object segmentation methods into supervised, unsupervised, semi-supervised and weakly supervised methods. Recently, more research efforts have been devoted to tackling VOS task in deep learning frameworks. Generally, one-shot video object segmentation is understood as making use of a single annotated frame (often the first frame of the sequence) to estimate the remaining frames segmentation in the sequence. On the other hand, zero-shot video object segmentation is understood as building models that do not need any labeled data of video frames[7,8].

2.1.1. Supervised video object segmentation

Early non-learning methods typically address supervised VOS task using handcrafted [9], and more recent research has turned towards deep learning paradigms [10,11]. Typically, supervised

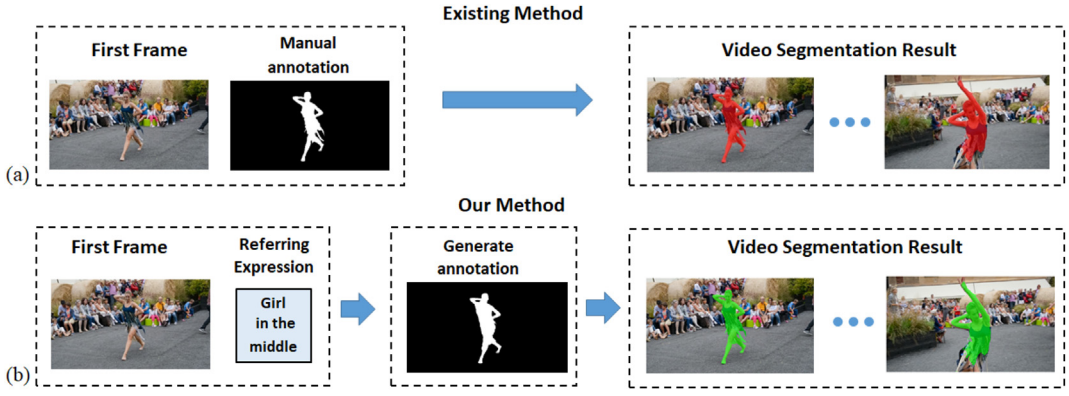

Fig. 1. Comparison between OSVOS and REVOS:(a) is the existing method (OSVOS), and (b) is our method (REVOS). By inputting a referring expression ("Girl in the middle"), our model automatically calculates the object mask and does object segmentation for the entire video clip.

methods require a mount of labeled data and model training is often costly and heavily affected by the availability of annotations from the target categories. Compared with unsupervised or semi-supervised methods, supervised approaches can produce more accurate partitions. However, the labor intensive process is unfeasible at large scale in applications.

2.1.2. Unsupervised video object segmentation

Unsupervised VOS models require neither category-specific training nor user's interactions and perform object segmentation with intrinsic cues, such as salient motion, object appearance, visual attention and etc. [12,13]. By taking video saliency as object-level cues for unsupervised VOS, [14] formulates the pixel-wise segmentation task as an energy minimization problem with a geodesic distance based technique that provides consistent saliency measurement of super-pixels as a priority for pixel-wise labeling. [13] proposes a fully end-to-end trainable recurrent network for multiple object VOS tasks. To model long-term temporal dependencies, [15] introduce a technique to establish dense correspondence between pixel embedding of a reference "anchor" frame and the current one. [16] proposes a DNN network called CO-attention Siamese Network to address the unsupervised video object segmentation task from a holistic view. The authors suggest a global co-attention mechanism which encodes useful information by processing multiple reference frames together, and their idea can be intuitively summarized as "see more, know more". [16] conducts a systematic study on the role of visual attention in the unsupervised video object segmentation task and it is the first attempt to collect human attention data on three public video segmentation datasets (DAVIS, Youtube-Objects and SegTrack). The authors quantitatively verified the high consistency of visual attention behavior among human observers and found a strong correlation between human attention and explicit primary object judgments during dynamic, task-driven viewing.

2.1.3. Semi-supervised video object segmentation

Semi-supervised VOS is a group of methods whose supervised level is between the supervised VOS and unsupervised VOS. With a few labeled data, semi-supervised methods leverage the inner structure of unlabeled data and propagate labels from labeled data to unlabeled data.

Traditionally, semi-supervised VOS lets user label the first frame or other key frames firstly and then performs object segmentation in the remaining frames. Most of the current literature on semi-supervised VOS enforce temporal consistency in video sequences to propagate the initial mask into the following frames. For example, in order to reduce the computational complexity, some work make use of super-trajectory [17] superpixels [18,19], patches [20,21], or even object proposals [22]. Moreover, an optimization using one of the previous aggregations of pixels is usually performed; which can consider the full video sequence [23], a subset of frames [19], or only the results in frame n to obtain the mask in $n+1$ [20,18,21]. As part of their pipeline, some of other methods include the computation of optical flow [19,20], which considerably reduces speed.

Unlike those methods, One-Shot Video Object Segmentation (OSVOS) [1] separates each frame independently without using temporal consistency. It is also state-of-the-art in semi-supervised VOS and has influenced other leading methods [6]. Given the manual annotation of the first frame, OSVOS makes the classification of all pixels of a video sequence into background and foreground. OSVOS adopts a CNN architecture and trains it in two stages: online training and offline training. In the offline training, base network is trained on ImageNet for image labeling. Then, the base network is further trained on the binary masks of the training set of DAVIS (parent network). In the online training, the test network is trained (fine-tuned) from the parent network on a segmentation example for the specific target object in a single frame (an annotated object mask), and this helps the network rapidly focus on that target object. One unique property of OSVOS is that it does not require temporal consistency, i.e., the order that OSVOS segments frames are inconsequential. Conversely, even when segmentation methods operate sequentially, segmentation can propagate forward and backward from annotated frames selected later in a video.

2.1.4. Weakly supervised video object segmentation

Semi-supervised VOS methods often suffer from criticisms because of their requirement for a pixel-level object mask. To further reduce the supervision cost, a few work has been found in literature (usually named weakly supervised methods). In scenarios where a mouse or a touch screen is available, clicks and scribbles are user friendly ways to do supervision. [24] explores the use of extreme points in an object(left-most, right-most, top, bottom pixels) as input to obtain precise object segmentation for images and videos. By taking one-shot video object segmentation[1] as the backbone, [25] proposes a human-in-the-loop video object segmentation method with a handful of clicks. [26] presents a deep learning method for the interactive video object segmentation which builds upon two core operations (interaction and propagation) using user scribble annotations. With the rapid growth of video sharing web sites, a massive amount of videos are associated with semantic tags and taken as weakly labeled at a video level (or image level)[27,2–4].

2.2. Annotation frame selection for semi-supervised video object segmentation

[6] is the only investigation on how to select the best user-annotated frame for semi-supervised video object segmentation. The authors found that segmentation performance across the entire video varies dramatically when selecting an alternative frame for annotation. This encourages them to address the problem and propose a deep sorting network (BubbleNets) that learns to select frames using a performance-based loss function. By using the performance-based loss function, BubbleNets is trained to predict the relative performance difference of two frames. In the testing stage, BubbleNets makes relative performance predictions, iteratively comparing and swapping adjacent frames until the frame with the greatest predicted relative performance is identified.

The BubbleNets method has two limitations: (i) BubbleNets require previously annotated video object segmentation datasets for training (calculate the performance-based loss function), and this requirement is expensive and time consuming in real applications. (ii) BubbleNets does not supply any solution for selecting multiple user-annotated frames.

2.3. Referring expression comprehension(REC)

The task of referring expression comprehension is to localize a region described by a given referring expression. To address this problem, some recent work [28,29] uses CNN-LSTM structure to model and looks for the object by maximizing the probability. Other recent work uses joint embedding model [30–33] to compute matching score directly. In a hybrid of both types of approaches, [34] proposed a joint speaker-listener-reinforcer model that combined CNN-LSTM (speaker) with embedding model (listener) to achieve state-of-the-art results. Most of the above treat comprehension as bounding box localization, but object segmentation from referring expression has also been studied in some recent work. Such as MAttNet [35] which takes a natural language

expression as input and softly decomposes it into three phrase embeddings. MAttNet learns to parse expressions automatically through a soft attention based mechanism, instead of relying on an external language parser [36,37].

3. Proposed method

REVOS is an user-friendly variant of OSVOS, because it obtains a pixel-level object mask with a language referring expression. The framework of REVOS is illustrated in Fig. 2, which includes three parts: referring expression analysis, object mask annotation and few-shot deep learning.

3.1. Referring expression analysis

Referring expressions are natural language utterances that indicate particular objects within a scene. Referring expression comprehension is the technique to locate an object in an image with a referring expression and is typically formulated as selecting the best match between an image region and a referring expression.

We use a referring expression to specify an object instance in a video frame. For any possible referring expression $r = \{u_t\}_{t=1}^{T}$, all the u_t forms a dictionary set $D = \{u_i\}_{i=1}^{N}$. $N = \{u_k\}_{k=1}^{K}$ is a subset of D. Each element u_k of N is a noun word (a name of an object). N actually is the set of objects (the word"object" in the paper means its name should be included in set N, otherwise it will be taken as background) and there are K object instance which can be potentially segmented by the system. To reduce the difficulty of calculating object mask with a referring expression, rules for generating the referring expression are stipulated.

Rules for generating referring expression:

(1) If there is only one object instance in the video frame, then the referring expression is "*subject*", e.g. "dog".
(2) If there is more than one object instance and the one to be segmented does not overlap another object in the video frame, then the referring expression is "*subject* + *location*", e.g. "girl in the middle".
(3) If there is more than one object instance and the one to be segmented does overlap another object in the video frame, then the referring expression is "*subject* + *relationship*", e.g. "man riding on a horse".

Given a referring expression $r = \{u_t\}_{t=1}^{T}$ under the rules above, the words u_t are grouped into three phrases categories $(\{sub\}, \{loc\}, \{rel\})$ by using the strategy in Fig. 3. Then we embed each word u_t into a vector e_t using a one-hot word embedding. Correspondingly, e_t can also be grouped into three categories$(\{e_{sub}\}, \{e_{loc}\}, \{e_{rel}\})$. Three phrase embeddings are calculated by:

$$q_{sub} = \sum e_i, e_i \in \{e_{sub}\} \tag{1}$$

$$q_{loc} = \sum e_i, e_i \in \{e_{loc}\} \tag{2}$$

$$q_{rel} = \sum e_i, e_i \in \{e_{rel}\} \tag{3}$$

3.2. Object mask annotation

In the paper, we utilize Mask R-CNN [38] as the backbone net for faster implementation and predicting pixle-level object masks.

Corresponding to three phrases embedding$(q_{sub}, q_{loc}, q_{rel})$, we design three modules ("subject", "location", "relationship") to locate target objects. Given a video frame X and a referring expression, we run Mask R-CNN extended from ResNet[39] and get a set of candidates o_i. According to the type of the referring expression label ("*subject*", "*subject* + *location*" or "*subject* + *relationship*"), three combinations of the modules are adopted to do object localization and output a bounding box. And then, the binary object mask is calculated with the mask branch network in Mask R-CNN [38] (see Fig. 4).

Subject Module: The visual feature of o_i is denoted as v_{sub}^{i}. The subject module is formulated as:

$$S(o_i|q_{sub}) = \mathcal{F}(v_{sub}^{i}, q_{sub}) \tag{4}$$

where $\mathcal{F}(.)$ is the matching function to measure the similarity between o_i representation v_{sub}^{i} and phrase embedding q_{sub}. As is shown in Fig. 5, the matching function consists of two MLPs (multi-layer perceptions) and two L2 normalization layers. Each MLP is composed of two fully connected layers with ReLU activations, serving to transform the visual feature and phrase embedding into a common embedding space. The inner product of the two L2-normalized representations is computed as their similarity score[35].

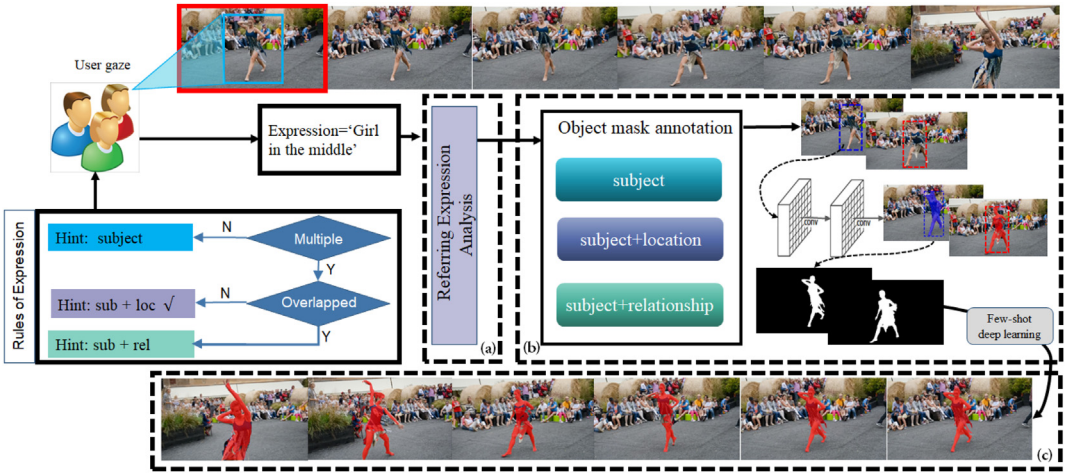

Fig. 2. The framework of our method:(a): Referring expression analysis. (b): Object mask annotation. (c): Few-shot deep learning. By observing a video frame, users specify an object instance by a referring expression; our model calculates the best match between image candidates and the referring expression and generate the object mask; few shot deep learning is utilized to do object segmentation on the remaining frames.

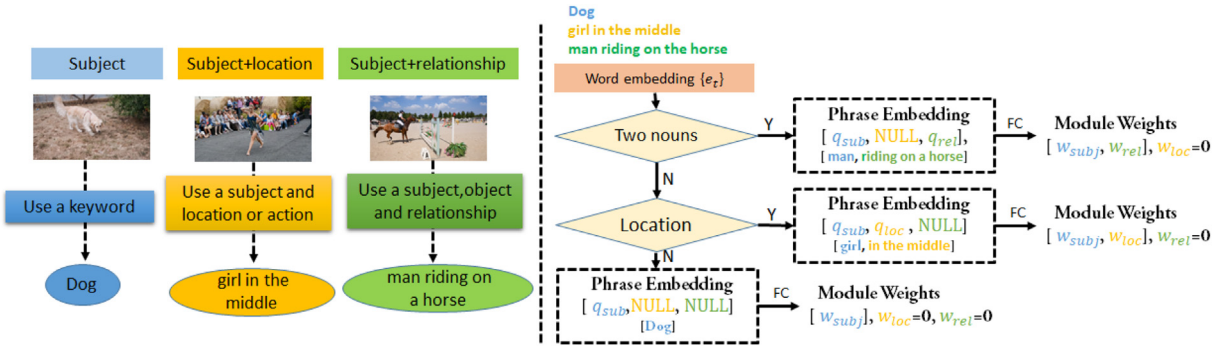

Fig. 3. The rules of generating referring expressions and the flow chart of expression analysis module.

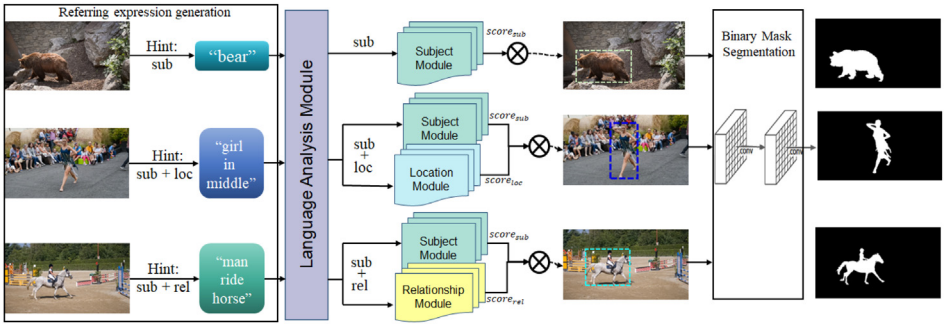

Fig. 4. Pipeline of object mask annotation.

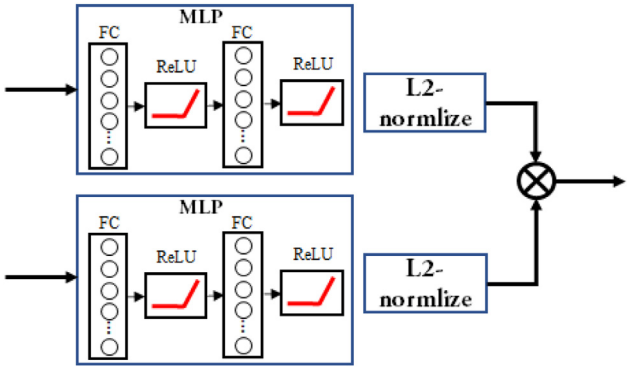

Fig. 5. Matching function (MLP is a two full connected layers with ReLu activations).

Location Module: Location is modeled as a 5-d vector encoding the x and y locations of the top left and bottom right corners of the target object bounding box, as well as the bounding box size with respect to the image[40].

$$l^i_{loc} = \left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}\right] \tag{5}$$

The location module calculates the matching score by:

$$S(o_i|q_{loc}) = \mathcal{F}(l^i_{loc}, q_{loc}) \tag{6}$$

Relationship Module: The relationship module deal with another object out of bounding box o_i. Given a candidate object o_i we first look for its closest object o_{ij}. We denote the visual representation of o_{ij} as v_{ij}. The offsets from o_i to o_{ij} the candidate object via is encoded by:

$$\delta m_{ij} = [\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w \cdot h}{W \cdot H}] \tag{7}$$

Then the visual representation of relationship is modeled by:

$$v^{ij}_{rel} = W_r[v_{ij}; \delta m_{ij}] + b_r \tag{8}$$

And the matching score for o_{ij}, and q_{rel} is:

$$S(o_{ij}|q_{rel}) = \mathcal{F}(v^{ij}_{rel}, q_{rel}) \tag{9}$$

Loss Function: The overall weighted matching score for the candidate object o_i and referring expression r is:

$$S(o_i|r) = w_{sub}S(o_i|q_{sub}) + w_{loc}S(o_i|q_{loc}) + w_{rel}S(o_{ij}|q_{rel}) \tag{10}$$

Where w_{sub}, w_{loc} and w_{rel} are weight coefficients, which are determined by the strategy in Fig. 3.

During training, for each given positive pair of (o_i, r_i), we randomly sample two negative pairs (o_i, r_j) and (o_k, r_i), where r_j is the expression describing some other object and o_k is some other object in the same image, to calculate a combined hinge loss,

$$\mathcal{L}_r = \sum_i [\lambda_1 max(0, \Delta + S(o_i|r_j) - S(o_i|r_i)) + \lambda_2 max(0, \Delta + S(o_k|r_i) - S(o_i|r_i))] \tag{11}$$

The overall loss incorporates both ranking loss and mask loss $\mathcal{L} = \mathcal{L}_r + \mathcal{L}_{mask}$

L_{mask} is defined as the average binary cross-entropy loss[41]. The architecture of M R-CNN has a mask branch to get binary masks. The mask branch has an output b^K_i for each o_i, which encodes K binary masks (one for each of the K objects in $N = \{u_i\}^K_{k=1}$). For a given o_i associated with ground-truth class k, L_{mask} is only defined on the k-th mask(other mask outputs do not contribute to the loss). The

binary mask B of video frame X can be calculated by b_i^K, and input to few-shot deep learning as an annotated object mask.

3.3. Annotation frame selection

Semi-supervised learning utilizes a few labeled data and inner structure of unlabeled data to propagate labels from labeled data to unlabeled data. So the locations of labeled data are of critical importance to label propagation, as is shown in Fig. 6.

Inspired by this observation, we propose a strategy of annotation frame selection, and the intuition is that the first user-annotated frame should be the one with the highest similarity to other frames; the second user annotated frame should be the one with the biggest difference with the first annotated frame, and so on.

Image Similarity: we use Eq. (12) to measure the similarity between two images.

$$S(I_i, I_j) = exp(-\frac{\left\|\hat{I}_i - \hat{I}_j\right\|_1}{HW}) \tag{12}$$

For two images I_i and I_j, $\hat{I}_i$ and $\hat{I}_j$ are their down-sampling counterparts with size $H \times W$; $\|\cdot\|_1$ is L1 norm operation.

The Best Annotation Frame Selection: given a video with L frames, we select the best annotated frame $I_{i_0^*}$ by Eq. (13).

$$i_0^* = \min_i \left[\frac{1}{L-1}\sum_{j\neq i} S(I_i, I_j)\right] \tag{13}$$

Multiple Annotation Frame Selection: the user-annotated frames after the best annotated frame are selected by Eq. (14).

$$i_l^* = \arg\min_{i\neq i_0^*,\ldots,i_{l-1}^*} \left\{\lambda_1\left[\frac{1}{L-1}\sum_{j\neq i} S(I_i, I_j)\right] + \lambda_2 \sum_{j=i_0^*}^{i_{l-1}^*} exp\left(-\frac{|j-i|}{c_0}\right)\right\}, l = 1, 2, \ldots \tag{14}$$

3.4. Few-shot deep learning

Following the work of OSVOS[1], we train the Fully Convolutional Neural Network (FCN) on two stages. In the offline training stage, the FCN is pre-trained on ImageNet and DAVIS training sets for image labeling and this helps to construct a model that is able to discriminate the general notion of a foreground object; in the online stage, we fine-tune the network with multiple annotated object masks. So the one-shot framework is extended to a few-shot framework. In order to deal with K annotated frames, a variant of foreground branch loss function is developed and utilized to train the test network[1].

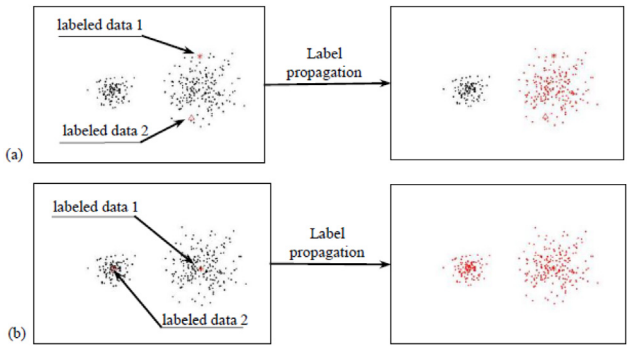

Fig. 6. Illustration of label propagation with different labeled data.

$$\mathcal{L}_{mod}(W) = \sum_{k=1}^{K}\left\{-\beta^{(k)}\left[\sum_{j\in Y_+^{(k)}} logP(y_j^{(k)} = 1|X^{(k)}; W)\right] - (1-\beta^{(k)})\left[\sum_{j\in Y_-^{(k)}} logP(y_j^{(k)} = 0|X^{(k)}; W)\right]\right\}$$

where W are the standard trainable parameters of a CNN; Given K binary annotated frames $Y^{(k)}, k = 1, 2, \ldots, K$, and the original images $X^{(k)}, k = 1, 2, \ldots, K$, we train the test network with the foreground branch loss function. Two binary mask $Y_+^{(k)}$ and $Y_-^{(k)}$ are derived from $Y^{(k)}$, which are positive and negative labeled pixels respectively. $y_j^{(k)} \in 0, 1, j = 1, .., |X^{(k)}|$ is the pixelwise binary label of $X^{(k)}$. $P(\cdot)$ is obtained by applying a sigmoid to the activation of the final layer. $\beta^{(k)} = |Y_-^{(k)}|/|Y^{(k)}|$. Eq. 15 allows training for imbalanced binary tasks[1].

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

The main part of our experiment is done on DAVIS-2016 validation sets. In the experiment of ablation study, two other datasets (GyGo and Youtube-VOS) are also utilized to train our model.

DAVIS-2016 dataset: DAVIS dataset consists of 50 full HD video sequences, including 30 training sets and 20 validation sets. It has a total of 3455 labeled frames, a video frame rate is 24fps, and resolution ratio is 1080p. We divide the samples into three categories. The first category has 30 video clips with only a single object; the second category includes 8 video samples containing more than one object instances; in the third category (12 video clips), each video sample has multiple object instance meanwhile the one to be segmented overlapping with other kinds of object. By applying the rules of generating referring expression, we use "*subject*" (a single noun word) to describe the object instance in the first category, "*subject* + *location*" for the second category and "*subject* + *relationship*" for the third category.

GyGo dataset: GyGo dataset consists of approximately 150 short videos. The sequence of the video is very simple, with almost no deformation, motion blur, and scale variation, or other attributes that increase video complexity. It has more categories than the DAVIS-2016 dataset, many of which contain known semantic categories (such as people, car, etc.). GyGO specializes in collecting videos taken by smart phones, so the frames are sparse (the video frame rate is only about 5 fps).

Yotube-VOS dataset: The YouTube-VOS dataset contains 4,453 YouTube video clips and 94 object categories, including humans, common animals, vehicles, and attachments. Each video clip is about 36 s long and usually contains multiple objects. This is by far the largest video object segmentation dataset we know.

4.1.2. Implementation details

Our code is based on Pytorch and Tensorflow. We use Mask R-CNN as the backbone, and the three visual modules are based on the code of MattNet model [35]. The few-shot deep learning is derived from the code of OSVOS [1].

Mask R-CNN model is pre-trained on COCO dataset and we did not use extra datasets to train it, so it can only predict 80 categories of objects [38]. In DAVIS −2016, there are 26 categories of target objects and 10 of them do not have an accurate label in the 80 categories. Among the 10 categories, 7 of them can be classified to one more extensive class, for instances, "gril" and "man" are grouped into "person" and "swan" is into "bird". Other 3 categories are classified to a reasonable nearby class("camel" to "horse", "rhinoceros"

and "baby carriage" to "chair"). In our experiment, we choose the nouns included in the 80 categories for the referring expression labels. For example, we actually give the label "horse" to the image containing one or more camels. GyGo and Yotube-VOS are used in the ablation study experiment, and only the video samples with target objects in the 80 categories are adopted to train our model.

4.1.3. Evaluation metrics

We evaluate the effectiveness of our method on the DAVIS dataset with two evaluation metrics: intersection-over-union metric for measuring the region-based segmentation similarity, and F-measure for measuring the contour accuracy.

Region Similarity is measured by Intersection over Union (IoU), which is an evaluation metric frequently used to measure the accuracy of an object detector. The definition of IoU is given by $IoU = \frac{M \cap G}{M \cup G}$, where G is the ground truth and M is the calculated object mask.

Contour Accuracy is evaluated by F-measure, a combination of accuracy $P = \frac{M \cap G}{M}$ and recall $R = \frac{M \cap G}{G}$. So F is taken as the weighted harmonic average of accuracy and reall and is calculated by $F = \frac{a^2+1}{a^2} \times \frac{P \times R}{P+R}$. We make a = 1 in the paper, then the F metric becomes the common $F_1 = \frac{2 \times P \times R}{P+R}$, where recall and accuracy share the same contribution to the evaluation result. (The weight a can be adjusted according to specific needs by users).

4.2. Comparison with the state-of-art

We compare our method in one-shot mode with other five methods, OSVOS [1], OFL [42], BVS [23], HVS [19], SEA [20]. In the experiment, both OSVOS and REVOS use the first frame of each video as labeled data. OSVOS takes the annotated frames in DAVIS dataset to fine-tune its network. REVOS takes the original images and corresponding referring expressions to calculate the object masks, and then the binary masks are utilized to initialize the object segmentation on other frames (see Fig. 7).

The final statistical results are shown in Table 1 and Fig. 8. From the experimental results, we observed that the recall mean of REVOS is 93.7% which is a slightly higher than that of OSVOS. Three other values are slightly less than that of OSVSO. So we can conclude that our method has the similar or a litter bit lower performance than OSVOS. However, in Section 4.5 we will see that, the accuracy performance of REVOS is improved with multiple annotated frames since annotating video frames with a referring expression do not require much user workload.

4.3. Evaluation of referring expression generation

REVOS requires users to input referring expression under three rules and referring expression can be grouped into three categories: "*subject*", "*subject* + *location*", "*subject* + *relationship*". In the section, we design an experiment to verify the effectiveness of the rules and the experimental results are listed in Fig. 9 and Table 2. We used "*subject*" and "*subject* + *location*" to generate object masks on 30 single-object videos, and finally we got the similar IoU value. For 8 multi-object videos, we also used "*subject*" and "*subject* + *location*" to generate object masks, the results show that the accuracy of "*subject* + *location*" rule is best. For 12 overlaped videos, we evaluated all kinds of referring expression and concluded that "*subject* + *relationship*" rule is best.

By looking at the experimental results, we observed that:

(1) The first category of referring expressions ("subject") can achieve satisfactory results on the single object cases (e.g. "bear" in Fig. 9(a)), and adding more information ("location" or "relationship") dose not improve any segmentation performance.
(2) To deal with the multiple instances, the second category of referring expressions ("*subject* + *location*") are necessary and ignoring location information leads to a drop of the accuracy from 79% to 20.7%. In Fig. 9 (b), the referring expression "man" dose not help to locate the man in the middle probably because the face of the person at the most right is visible.
(3) The benefit of "*subject* + *relationship*" are two folds: Firstly, it helps to locate the target object and improve the segmentation accuracy; secondly, it can represent more sophisticated object masks, e.g. "person + horse" in DAVIS-2016 dataset.

4.4. Evaluation of object mask precision

OSVOS requires manually annotated object masks because the precision of object masks have obvious influence on its performance. To verify this, we lower the image quality of the first frame by crystallization operation (Photoshop Tools). We process the ground truth of the first video frame with different crystallization parameters (20%, 40%, 60%, 80%) and get object masks with deferent precision measured by IoU (81%, 73%, 67%, 55%) shown in Fig. 10.

We initialize OSVOS with deferent object masks and list the experimental results in Table 3. By observing the table, we conclude from the observation that lower object mask precession leads to an obvious decline of OSVOS performance.

4.5. Evaluation of annotation frame selection

To compare our method with BubbleNets, we select user annotated frames with the two strategies and feed the ground truth of the frames in DAVIS dataset to REVOS. We determine the effective-

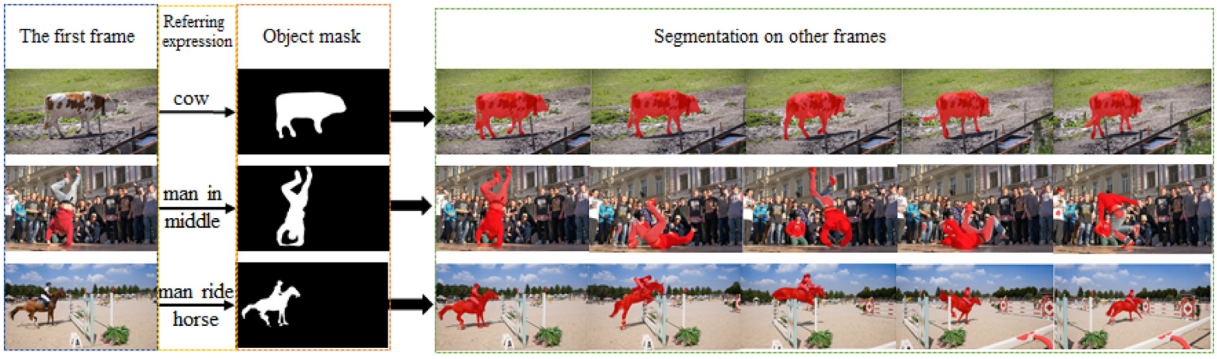

Fig. 7. Examples with three kinds of referring expressions (the first row: subject; the second row: subject + location; the third row: subject + relationship).

Table 1
DAVIS Validation: REVOS versus the state of the art, and practical quality.

	REVOS	OSVOS	OFL	BVS	HVS	SEA	
Region Similarity(%)	Mean	79.6	79.8	68.0	60.0	54.6	50.4
	Recall	93.7	93.6	75.6	66.9	61.4	53.1
Contour Accuracy(%)	Mean	80.3	80.6	63.4	58.8	52.9	48.0
	Recall	92.5	92.6	70.4	67.9	61.0	46.3

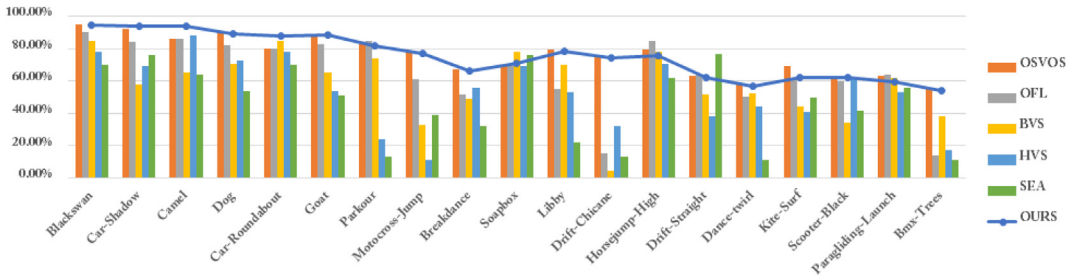

Fig. 8. DAVIS Validation: Per-sequence results of region similarity.

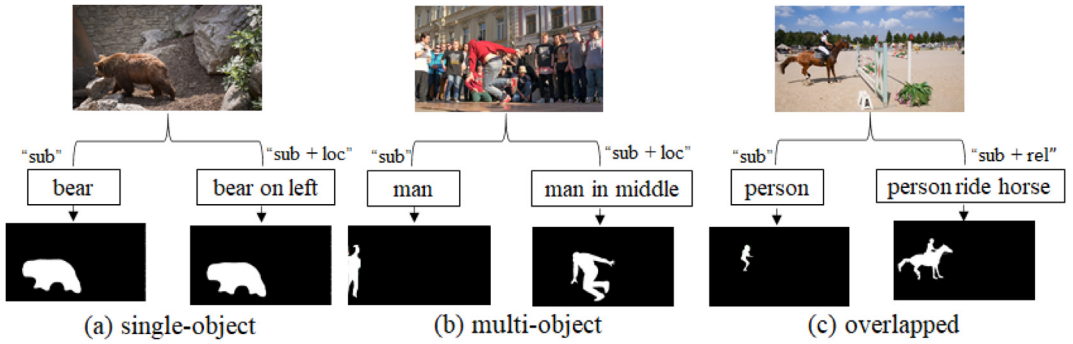

Fig. 9. Some correct and failed cases for evaluation of referring expression generation.

ness of each frame selection strategy by calculating the mean of Region Similarity (IoU) for the resulting segmentation.

We also take the current semi-supervised standard strategy (select the first frame) as the baseline method.

4.5.1. The best annotation frame selection

Complete the best annotation frame selection results for the 20 video samples in validation sets are provided in Fig. 11. By observing the results, we find that our method outperforms BubbleNets on 15 videos and BubbleNets has better segmentation performances on other 5 videos.

4.5.2. Multiple annotation frame selection

We use our multiple annotation frame selection strategy to select one more frame and feed the two annotated frames to REVOS, and experimental results are provided in Fig. 12. By using two annotation frames, we found that our method outperforms BubbleNets on all the videos. We can conclude that multiple annotation frame strategy can help achieve higher segmentation performance than the single frame selection strategies on the DAVIS-2016 validation datasets (Fig. 13).

To make a closer observation, we choose the video samples whose IoU is less than 80% by using our best annotation frame selection and list in Table 4 (8 video samples).

For the baseline method (the first frame selection), it is worth acknowledging that DAVIS dataset intends for annotation to take place on the first frame, which guarantees that objects are visible for annotation (in some videos, objects become occluded or leave the view). In the 8 video samples, we observed that BubbleNets get worse performance than the baseline method on 3 videos and our strategy(single frame) has two.

Although most of the experimental results prove that more annotated frames help to get better segmentation performance, we still find an exception that more annotated frames make the segmentation performance worse. By looking at the last row of Table 4, we see that the IoU of two annotated frames is less than that of the baseline method. To figure out the reason, we run an additional experiment and the result is shown in Fig. 14. We compare the result taking one annotated frame("frame 1") with that of two annotated frames("frame 1 + frame 41"). We observed that adding more annotated frame ("frame 41") causes more segmentation noise in the region of the ropes and makes the IoU value lower. So we also boldly conclude that unsuitable annotated object masks can do harm to segmentation performance of semi-supervised VOS.

Table 2
Evaluation of referring expression generation.

Referring expression	Single-object (%)	Multi-object (%)	overlapped (%)
Subject	83.6 ✓	20.7	51.3
Subject + location	83.5	79.2 ✓	60.9
Subject + relationship			78.4 ✓

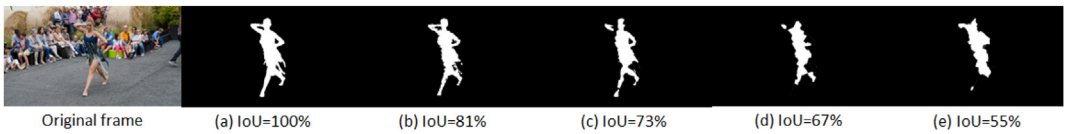

Fig. 10. Evaluation of object mask precision. To get an object mask with different accuracy ratio, we apply crystallization operation tool of Photoshop with different crystallization parameters (20%, 40%, 60%, 80%) and get different IoUs (81%, 73%, 67%, 55%) respectively.

Table 3
Performance on object mask with different accuracy ratio.

	IoU = 100%	IoU = 81%	IoU = 73%	IoU = 67%	IoU = 55%
Dance-Twirl (%)	58.9	56.1	54.4	49.8	47.5
Goat (%)	86.3	80.5	72.4	64.2	59.8
Parkour (%)	83.0	81.7	79.2	61.3	57.3
Breakdance (%)	67.9	62.5	59.7	56.3	52.6
Camel (%)	85.9	82.6	78.4	63.4	55.1

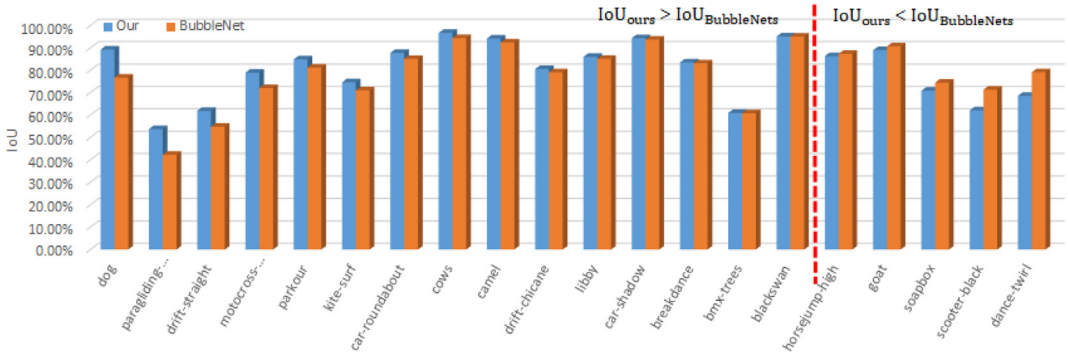

Fig. 11. Comparison between our strategy and BubbleNets on DAVIS-2016.

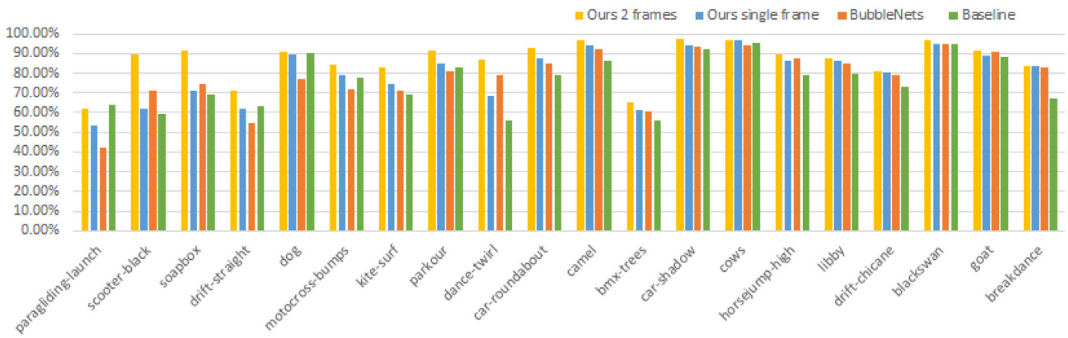

Fig. 12. Evaluation of multiple annotation frame selection strategy on DAVIS-2016.

4.6. Ablation study

To study the effect of different training dataset on experimental results, we performed an ablation study. Our complete model is first pre-trained on ImageNet dataset and then fine-tuned it by using video data. To verify the effectiveness of pre-training, we compared different models that are just trained on the video data without the pre-training. In addition, to further examine the impact of the amount of video training data, we evaluated variants using only DAVIS-2016 training videos for fine-tuning. Table 5 summarizes the results obtained from the variant models that we trained with different combination of training datasets.

In this experiment, we adopt GyGo dataset and Youtube-VOS dataset for the ablation study. However, our Mask R-cnn is trained on 80 object categories and not all of the categories on GyGo and Youtube-VOS are included in the 80 categories. To handle the problem, we only use the video samples with foreground objects which are included the 80 categories. Youtube-VOS has 94 object categories and 33 of them are out of the 80 categories (117 video clips). GyGo has 57 object categories and 10 of them are not included (17 video clips).

PT: pre-training on static images. DV, GG and YV: the use of DAVIS, GyGo, and Youtube-VOS for fine-tuning. AM: annotate more 5 frames. By experiments we can conclude that without

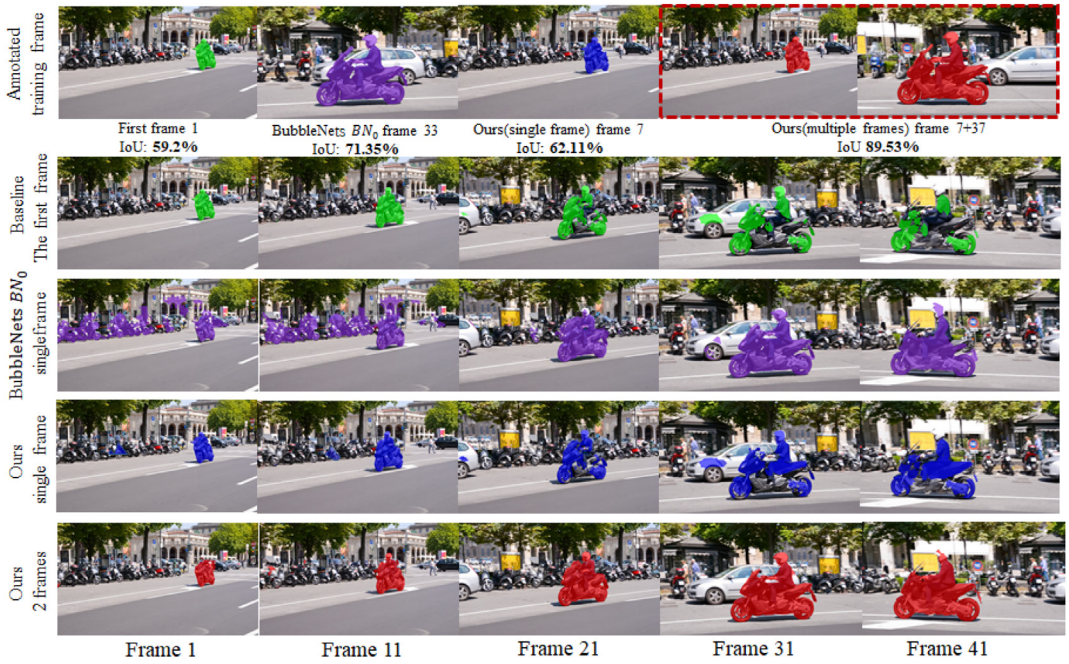

Fig. 13. Qualitative Comparison on DAVIS 2016 Validation Set: Segmentations from different annotated frame selection strategies.

Table 4
Evaluation of multiple annotation frame selection strategy. The video samples whose IoU is less than 80% by using our best annotation frame selection are listed in the table.

Video Name	Baseline IoU (%)	BubbleNets BN_0 Frame No./ IoU (%)	Ours (single frame) Frame No./ IoU (%)	Ours (two frames) Frame No./ IoU (%)
Dance-twirl	55.81	90/ 79.10 ↑ 27.42	46/ 68.60 ↑ 12.79	46 & 3/ 87.10 ↑ 31.29
Scooter-black	59.20	33/ 71.35 ↑ 12.15	7/ 62.11 ↑ 2.91	7 & 37/ 89.53 ↑ 30.33
Soapbox	69.26	99/ 74.45 ↑ 23.29	90/ 70.82 ↑ 1.56	90 & 46/ 91.63 ↑ 22.37
Kite-surf	69.25	50/ 70.98 ↑ 1.73	19/ 74.61 ↑ 5.36	19 & 44/ 82.89 ↑ 13.64
BMX-trees	55.78	80/ 60.80 ↑ 5.02	36/ 60.97 ↑ 5.19	36 & 1/ 64.95 ↑ 9.17
Drift-straight	63.49	50/ 54.78 ↓ 8.71	12/ 61.88 ↓ 1.61	12 & 34/ 70.88 ↑ 7.39
Motocross-bumps	77.42	40/ 72.06 ↓ 5.36	10/ 78.94 ↑ 1.52	10 & 40/ 84.56 ↑ 7.14
Paragliding-launch	63.64	80/ 42.26 ↓ 21.38	41/ 53.70 ↓ 9.94	41 & 2/ 62.00 ↓ 1.64

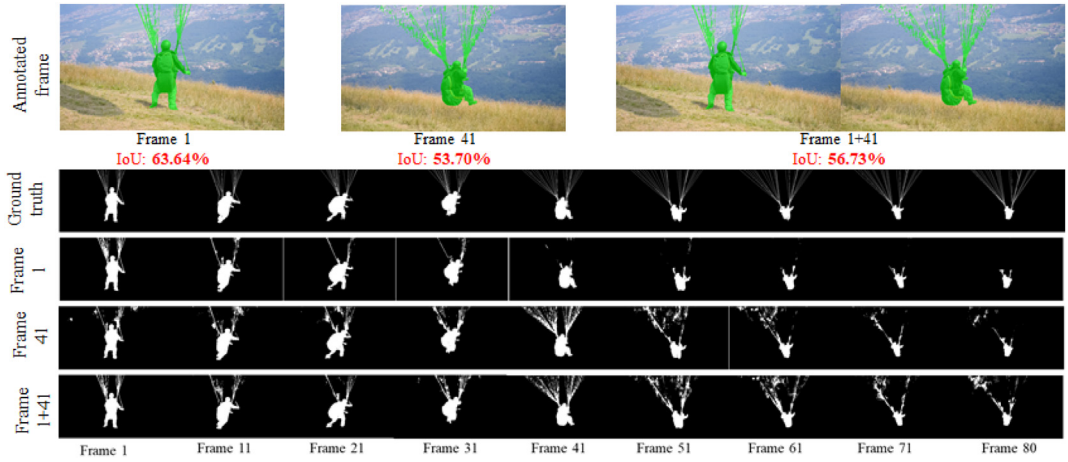

Fig. 14. An exception of multiple annotation frame selection strategy (video "paragliding-launch").

pre-training, our performance would drop significantly. Using additional training video dataset further improves the accuracy of our model.

5. Limitations and future work

Most state-of-the-art semi-supervised video object segmentation methods rely on a costly initialization with a pixel-level mask

Table 5
Ablation study.

PT	DA	GG + YV	AM	Region Similarity	Contour Accuracy
	✓			41.9	54.6
	✓	✓		62.2	63.2
✓	✓			79.6	80.3
✓	✓		✓	85.8	86.5
✓	✓	✓	✓	**88.1**	**91.9**

for the first frame of a video. To overcome the limitation we explore the problem of how to initialize the semi-supervised methods with referring expression. In this section, we discuss further improvements and several potential directions for future work. (1) In the paper, we propose three rules to generate a referring expression to specify an object in a video frame. Thus, the referring expression is represented by three forms (subject, subject + location or subject + relationship). This simplification does help to reduce the difficulty of language analysis algorithm, but It could also lead to the limitation that the target object can not be clearly described in some multiple instances scenarios (e.g. several girls in a line). (2) We utilize Mask R-CNN as the backbone of our method. Without training Mask R-CNN on extra image datasets, only 80 categories of objects can be predicted. Objects out of the 80 categories have to been treated as background. Further improvements have to be made to calculate a pixel-level mask of an arbitrary object. (3) We propose a strategy of annotated frame selection with image similarity measurement, and the strategy does not require any labeled object data or segmentation results. By observing the experimental results, our strategy shows good performance on OSVOS based methods. However, performance on other semi-supervised video object segmentation methods still need to be verified.

6. Conclusion

We propose a referring expression based variant (REVOS) of one-shot video object segmentation (OSVOS), which mainly solve the problem of manually annotated object mask required by OSVOS. To simplify referring expression analysis, we seek out three rules to generate a referring expression and select the target from all candidate objects in a video frame by finding the highest matching score with the referring expression. We also explore the issues of user annotation frame selection. By measuring image similarity between video frames, we propose two strategies, the best annotation frame selection and multiple annotation frame selection.

Finally, while the current REVOS implementation is specific to the method of One-shot video object segmentation, it is more widely applicable to other semi-supervised VOS methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by The Tianjin Science and Technology Program under grant 19PTZWHZ00020 and the National Natural Science Foundation of China under grant 61902281.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, One-shot video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 221–230.
[2] F. Saleh, S. Aliakbarian, M. Salzmann, L. Petersson, J.M. Alvarez, Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation, 2017, pp. 2125–2135. doi:10.1109/ICCV.2017.232.
[3] D. Zhang, J. Han, L. Yang, D. Xu, Spftn: a joint learning framework for localizing and segmenting objects in weakly labeled videos, IEEE Trans. Pattern Anal. Mach. Intell. PP (2018) 1–1. doi:10.1109/TPAMI.2018.2881114.
[4] Y. Zhang, X. Chen, J. Li, W. Teng, H. Song, Exploring weakly labeled images for video object segmentation with submodular proposal selection, IEEE Transactions on Image Processing PP (2018) 1–1. doi:10.1109/TIP.2018.2806995.
[5] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. Hoi, H. Ling, Learning unsupervised video object segmentation through visual attention, Comput. Vision Pattern Recogn. (2019), https://doi.org/10.1109/CVPR.2019.00318.
[6] B.A. Griffin, J.J. Corso, Bubblenets: learning to select the guidance frame in video object segmentation by deep sorting frames, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, https://doi.org/10.1109/CVPR.2019.00912.
[7] T. Zhou, S. Wang, Y. Zhou, Y. Yao, J. Li, L. Shao, Motion-attentive transition for zero-shot video object segmentation (2020). arXiv:2003.04253.
[8] W. Wang, X. Lu, J. Shen, D. Crandall, L. Shao, Zero-shot video object segmentation via attentive graph neural networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Oct 2019, https://doi.org/10.1109/iccv.2019.00933.
[9] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, B. Chen, Jumpcut: non-successive mask transfer and interpolation for video cutout, ACM Trans. Graphics 34 (6) (2015) 1–10.
[10] P. Lei, S. Todorovic, Recurrent temporal deep field for semantic video labeling, in: European Conference on Computer Vision, 2016, pp. 302–317.
[11] A. Kundu, V. Vineet, V. Koltun, Feature space optimization for semantic video segmentation, Comput. Vision Pattern Recogn. (2016).
[12] H. Song, W. Wang, S. Zhao, J. Shen, K.-M. Lam, Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XI, 2018, pp. 744–760.
[13] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, X. Giro-i Nieto, Rvos: end-to-end recurrent network for video object segmentation, Comput. Vision Pattern Recogn. (2019) 5272–5281.
[14] W. Wang, J. Shen, R. Yang, F. Porikli, Saliency-aware video object segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) 20–33.
[15] Z. Yang, Q. Wang, L. Bertinetto, W. Hu, S. Bai, P.H.S. Torr, Anchor diffusion for unsupervised video object segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[16] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[17] W. Wang, J. Shen, F. Porikli, R. Yang, Semi-supervised video object segmentation with super-trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2019) 985–998.
[18] J. Chang, D. Wei, J.W. Fisher, A video representation using temporal superpixels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2051–2058.
[19] M. Grundmann, V. Kwatra, M. Han, I. Essa, Efficient hierarchical graph-based video segmentation, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2141–2148.
[20] S. Avinash Ramakanth, R. Venkatesh Babu, Seamseg: video object segmentation using patch seams, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 376–383.
[21] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, B. Chen, Jumpcut: non-successive mask transfer and interpolation for video cutout, ACM Trans. Graph. 34 (6) (2015), 195–1.
[22] F. Perazzi, O. Wang, M. Gross, A. Sorkine-Hornung, Fully connected object proposals for video segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3227–3234.
[23] N. Märki, F. Perazzi, O. Wang, A. Sorkine-Hornung, Bilateral space video segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 743–751.

[24] K.-K. Maninis, S. Caelles, J. Pont-Tuset, L. Van Gool, Deep extreme cut: from extreme points to object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 616–625.
[25] A. Benard, M. Gygli, Interactive video object segmentation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[26] S. Oh, J.-Y. Lee, N. Xu, S. Kim, Fast user-guided video object segmentation by interaction-and-propagation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5242–5251, https://doi.org/10.1109/CVPR.2019.00539.
[27] M. Siam, N. Doraiswamy, B. Oreshkin, H. Yao, M. Jagersand, One-shot weakly supervised video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[28] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[29] L. Yu, P. Poirson, Y. Shan, A.C. Berg, T.L. Berg, Modeling context in referring expressions, in: European Conference on Computer Vision, 2016.
[30] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, B. Schiele, Grounding of textual phrases in images by reconstruction, in: European Conference on Computer Vision, 2016.
[31] L. Wang, L. Yin, S. Lazebnik, Learning deep structure-preserving image-text embeddings, Comput. Vision Pattern Recogn. (2016).
[32] J. Liu, W. Liang, M.H. Yang, Referring expression generation and comprehension via attributes, in: IEEE International Conference on Computer Vision, 2017.
[33] C. Kan, R. Kovvuri, R. Nevatia, Query-guided regression network with context policy for phrase grounding, in: IEEE International Conference on Computer Vision (ICCV), 2017.
[34] L. Yu, T. Hao, M. Bansal, T.L. Berg, A joint speaker-listener-reinforcer model for referring expressions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[35] L. Yu, L. Zhe, X. Shen, J. Yang, T.L. Berg, Mattnet: modular attention network for referring expression comprehension, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[36] J. Li, L. Mu, H. Zan, K. Zhang, Research on chinese parsing based on the improved compositional vector grammar, Chin. Lexical Semant. (2015) 649–658.
[37] I. Paraboni, M.R. Galindo, D. Iacovelli, Stars2: a corpus of object descriptions in a visual domain, Lang. Resour. Eval. 51 (2) (2016) 1–24.
[38] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 2980–2988.
[39] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, Arxiv (03 2016).
[40] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, K. Saenko, Modeling relationships in referential expressions with compositional modular networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[41] S. Xie, Z. Tu, Holistically-nested edge detection, Int. J. Comput. Vision (2017).
[42] Y.H. Tsai, M.H. Yang, M.J. Black, Video segmentation via object flow, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Xiaoqing Bu was born in JiangSu, China, in 1995. She is a master student at the School of Computer Science and Technology, Tianjin Polytechnic University. Her major is computer vision.

Yukuan Sun received his B.E and M.A.Eng degrees in 2009 and 2014, respectively, from Tiangong University, China. From 2019 to present, he is studying for a Ph.D. degree at Ajou University, South Korea. From 2014, he works as an engineer at Tiangong University. He focuses on research domains, e.g., Model compression and pruning, distributed deep neural network, Meta-learning, and swarm intelligent robot. He had published over five papers in journals and international conferences.

Jianming Wang is a professor at the School of Computer Science and Technology, Tianjin Polytechnic University. He received his PhD from Tianjin University, 2003. His main research interests are computer vision, neural networks, and pattern recognition.

KunLiang Liu is an Assistant Professor in the department of computer science and technology, TianGong University. He received the M.S degree from China University of Geoscience, WuHan, China in 2004. His research interests include virtual reality, scientific computing visualization and computer vision.

Jiayu Liang received her B.Sc. and M.Sc. degree in 2011 and 2014 respectively from University of Science and Technology Beijing, China. She got her Ph.D degree in Computer Science from Victoria University of Wellington, New Zealand. From 2018, she works as a lecturer in Tiangong University, China. Dr. Liang focuses on research areas, e.g. Evolutionary Computation, Multi-objective Optimization and Image Processing. She has published over 10 papers in top journals and international conferences in those areas. Moreover, she is a regular reviewer of journals, e.g. ‘‘Applied Soft Computing’’, ‘‘Soft Computing’’ and ‘‘Engineering Applications of Artificial Intelligence’’.

GuangHao Jin was born in Jilin, China, in 1979. He received the B.S. degree in Peking Uni- versity, Beijing, China, in 2002. He received the Ph.D. degree in Tokyo Institute of Technology, Tokyo, Japan, in 2014. Since 2016, has been an Lecturer in Tianjin Polytechnic University.He has published about 20 research papers in international SCI/EI journals and conferences. His research in- terests include computer vision, AI, deep learning, heterogeneous/reconfiguration computing.

Tae-Sun Chung received his B.S. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 1995, and M.S. and Ph.D. degrees in Computer Science from Seoul National University, in 1997 and 2002, respectively. He is currently a professor at Department of Software, Ajou University, Suwon, Republic of Korea. His current research interests include flash memory storages, database systems, and machine learning.