Chinese Journal of Electronics
Vol.32, No.3, May 2023

Unsupervised Video Object Segmentation via Weak User Interaction and Temporal Modulation

FAN Jiaqing[1], ZHANG Kaihua[2,3], ZHAO Yaqian[4], and LIU Qingshan[2,3]

(1. *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*)

(2. *College of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China*)

(3. *Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing 210044, China*)

(4. *Inspur Suzhou Intelligent Technology Corporation, Suzhou 215000, China*)

Abstract — In unsupervised video object segmentation (UVOS), the whole video might segment the wrong target due to the lack of initial prior information. Also, in semi-supervised video object segmentation (SVOS), the initial video frame with a fine-grained pixel-level mask is essential to good segmentation accuracy. It is expensive and laborious to provide the accurate pixel-level masks for each training sequence. To address this issue, We present a weak user interactive UVOS approach guided by a simple human-made rectangle annotation in the initial frame. We first interactively draw the region of interest by a rectangle, and then we leverage the mask RCNN (region-based convolutional neural networks) method to generate a set of coarse reference labels for subsequent mask propagations. To establish the temporal correspondence between the coherent frames, we further design two novel temporal modulation modules to enhance the target representations. We compute the earth mover's distance (EMD)-based similarity between coherent frames to mine the co-occurrent objects in the two images, which is used to modulate the target representation to highlight the foreground target. We design a cross-squeeze temporal modulation module to emphasize the co-occurrent features across frames, which further helps to enhance the foreground target representation. We augment the temporally modulated representations with the original representation and obtain the compositive spatio-temporal information, producing a more accurate video object segmentation (VOS) model. The experimental results on both UVOS and SVOS datasets including Davis2016, FBMS, Youtube-VOS, and Davis2017, show that our method yields favorable accuracy and complexity. The related code is available.

Key words — Unsupervised video object segmentation, The earth mover's distance (EMD)-based modulation, Cross-squeeze modulation, Weak interaction, Region-based convolutional neural networks (RCNN).

I. Introduction

Video object segmentation (VOS) is designed to separate the foreground object from a single video sequence. It has wide applications in self-driving, surveillance camera, online video conference system, and other real-world applications [1]–[4]. The task normally is handled by either semi-supervised methods (the initial mask is labeled manually) [5], [6] or unsupervised approaches (no artificial labeling in the first frame) [7]–[9]. However, both semi-supervised and unsupervised algorithms have their inherent limitations [5], [6]. On the one hand, semi-supervised (also named one-shot) VOS significantly depends on the accurate pixel-level labels in the first frame, which automatically predicts the masks in following frames. Although the initial strong pixel-level mask contributes to the performance gains, the manual human annotations are expensive and time-consuming in the real-world VOS applications. On the other hand, unsupervised (also namely zero-shot) VOS, given no initial guidance, the model is easy to misunderstand user's intention and obtains inaccurate initial mask, yielding inferior subsequent VOS results (see the first two rows in Fig.1 [10]). As illustrated in Fig.1, the first and third rows are the initialization of FSNet and ours. The second and fourth rows are the feature maps and predicted results in the subsequent frame. It is

Manuscript Received May 18, 2022; Accepted Sept. 26, 2022. This work was supported by National Key Research and Development Program (2021ZD0112200) and National Natural Science Foundation of China (U21B2044).

DOI:10.23919/cje.2022.00.139

noted that the feature map (in 1st row) of no weak user interaction is noisy and progressively accumulates errors to the subsequent frame (in 2nd row), obtaining a suboptimal segmentation result. However, when assisted by the weak user interaction initialization, both the initial and subsequent frames are clear and consequently achieves the optimal pixel-wise predictions.

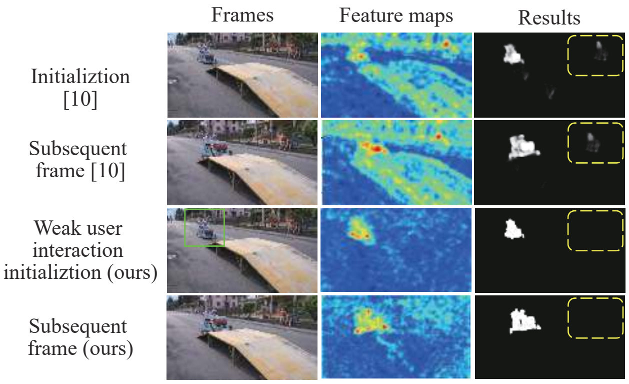


Fig. 1. Visualization of the initialization comparison between FSNet [10] and RectVOS (ours).

To alleviate the issues above, an intuitive solution is to leverage a user-friendly human interaction [11]–[13] (e.g. scribbles, strokes). These methods aim to find the target in the beginning of each video, using various types of interactions, e.g. scribbles, points, circles. Rotoscoping based interactive video cutout methods [14], [15] need the user to verify and update the object mask in each video frame, which requires much effort for human intervention. Recently, Nagaraja [16] *et al.* integrate motion and point trajectories, enforcing the consistency of the color distribution among successive frames, which yields favorable performance on ego motion videos. Then, Yang *et al.* [17] propose to find the corresponding patches between reference frame and target frame to enhance the interactive pixel-wise predictions, while maintaining the balance between the accuracy and performance. To establish human-in-the-loop interactive VOS system in the wild, Benard *et al.* [18] present to modify the existing semi-supervised approaches to adapt the interactive scenario. Very recently, given strokes or clicks, Oh *et al.* [19] jointly train two separate network modules to select the initial target and propagate the masks, respectively. The two networks are connected internally to solve the complex VOS task.

However, these methods above employ multiple human interactions, i.e. scribbles, strokes, clicks or cycles, leading to heavy user intervention burden. That is to say, it is essential to simplify the human interactions and keep the interactive VOS system to be more user-friendly. Furthermore, the aforementioned approaches learn a fixed offline VOS model to infer all the testing frames, which cannot take full advantage of the time-level contextual information. As the objects dramatically move in the video, these approaches cannot effectively adjust the model in following frames, which leads to the suboptimal results.

To alleviate the problems above, in this paper, we design a weakly-interactive VOS method with a rectangle annotation in the first frame as prior (termed as RectVOS) that seamlessly integrates the earth mover's distance (EMD) temporal modulation module (ETM) and the cross-squeeze temporal modulation module (CTM) into the VOS framework for end-to-end learning. Specifically, the ETM is designed to learn the EMD-based associations between coherent frames, which co-operatively gives higher weight to the object that simultaneously exists in both frames. Furthermore, the CTM is inspired by the cross-reference operation in the field of co-segmentation [20] that is able to stress the invariant features in two images and generate reinforced representations. Different from the majority of VOS methods that directly employ local temporal information [21], [22] to assist the consistent segmentation, our ETM exploits the cross-reference information from coherent frames, which produces a temporally modulated representation. Both the ETM and CTM enhance the representations of co-existed objects between consecutive frames, which is essential to the precise VOS. Extensive valuations on both unsupervised and semi-supervised VOS benchmarks containing DAVIS-2016 [23], FBMS [24], Youtube-VOS [25], and DAVIS-2017 [26] demonstrate that the proposed RectVOS yield favourable performance against the other compared VOS methods. We conclude the main contributions as follows:

- To reduce the time-consuming pixel-level annotations in labeling the initial frame in semi-supervised VOS task and enhance the initial prior in un-supervised VOS framework, we leverage a weak user interactive rectangle annotation as guidance signal to initialize the VOS model.
- On the spatio-temporal dimension, we employ the EMD-based similarity maps to temporally modulate the current frame's representation, which is able to highlight the co-existing foreground objects in the coherent frames.
- On the channel dimension, we design a cross-squeeze temporal modulation module to emphasize the co-occurrent representations of the raw features. As a result, only the common features in the two branches will be assigned with a high importance.
- Our RectVOS keeps good balance between accuracy and complexity on both unsupervised and supervised benchmarks including DAVIS-2016 [23], FBMS [24],

Youtube-VOS [25], and DAVIS-2017 [26].

II. Related Work

1. Interactive VOS, UVOS, and SVOS

The VOS task is mainly divided into three sub directions: UVOS, Interactive VOS, and SVOS, which have their own characteristics.

UVOS Ji *et. al.* [10] studied a new network architecture to combine the features of appearance and movement. Wang *et. al.* [7] propose a real-time video object segmentation network. When the video frames input, it is performed by the encoder and then processed by the pixel-level memory matching module.

Interactive VOS Zhang *et al.* [27] present a joint learned self-paced fine-tuning network (SPFTN) to localize objects with weakly labelled training videos. Afterwards, Benard *et al.* [18] present a human-in-the-loop system for interactive VOS, which is able to accurately segment targets using only a handful of clicks (3.8 clicks on average). Recently, Miao *et al.* [28] integrate the interaction and the propagations into one network with the memory aggregation mechanism, which considerably improves the efficiency of VOS without leveraging the strategy of multi-round interactions.

SVOS Mao *et al.* [29] design a framework to integrate transductive and inductive learning into a unified one. Recently, Duke *et al.* [30] propose a transformer-based approach to extract per-pixel representations for semi-supervised VOS.

2. EMD learning

As the EMD is able to calculate the structured distance between the two image representations, a series of works utilize the metric to promote the few-shot image classification and matching tasks. Rubner *et al.* [31] combine EMD metric with a representation scheme that is based on vector quantization, successfully handling image retrieval task. Ling *et al.* [32] propose a fast and exact algorithm named the earth mover's L1 distance (EMD-L1) to compute the similarity between a pair of histograms, which simplifies the objective function of the linear program and reduces the number of constraints. To enhance the complicated inter-class relationships that always exist in the realistic applications such as the age classification, Hou *et al.* [33] propose to use the EMD^2 loss to penalize the miss-predictions, which yields competitive performance. Zhang *et al.* [34] attempt to directly classify image representations using the EMD that is able to be taken as a structured fully connected layer into the network in an end-to-end manner.

3. Tracking-based segmentation

Recently, video object tracking (VOT) community has attempted to develop an unified framework for both the VOT and VOS tasks. Yeo *et al.* [35] utilize Absorbing Markov chain on superpixel segmentation to develop a simple yet effective tracking-by-segmentation algorithm. Wang *et al.* [36] augment the fully-convolutional Siamese tracking framework with a binary segmentation branch. Motivated by the aforementioned methods, Voigtlaender *et al.* [37] combine detection and design a novel dynamic programming algorithm to utilize tracklets from both initial frame and previous frame.

III. Methodology

Fig.2 illustrates the pipeline of the presented RectVOS, which mainly contains three designed modules: the weak user interaction prior (Section III.1), the earth mover's distance temporal modulation module (Section III.2), and the cross-squeeze temporal modulation module (Section III.3).

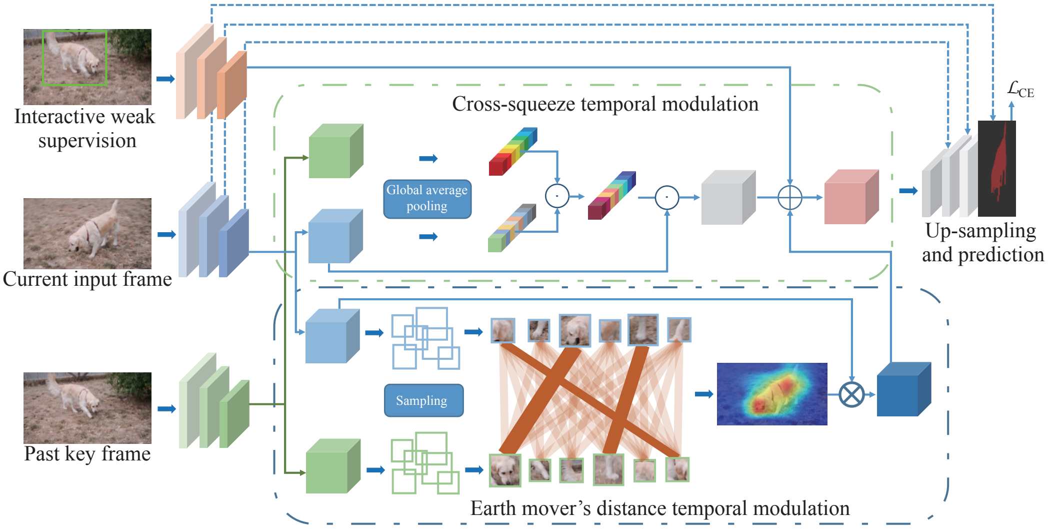

Interactive weak supervision
Current input frame
Past key frame
Cross-squeeze temporal modulation
Global average pooling
Sampling
Earth mover's distance temporal modulation
Up-sampling and prediction
$\mathcal{L}_{\mathrm{CE}}$

Fig. 2. Network pipeline of the proposed RectVOS including three components of interactive weak supervision, ETM and CTM.

Specifically, we first interactively draw the region of interest by a rectangle, and then we leverage the mask RCNN method to generate a set of reference labels for subsequent mask propagations. Then, the ETM is designed to highlight the co-existing foreground objects in the two frames, while the CTM is able to assign the higher weights to the informative channels. Furthermore, the multi-modal representations are integrated to yield the modulated spatio-temporal features. Finally, both the initial multi-stage features (transferred by the blue dotted line) and enhanced spatio-temporal representations are put into the up-sampling and prediction module (Section III.4) to obtain the pixel-level masks.

1. Weak user interaction prior

In the semi-supervised VOS community, it is time-comsuming to label the ground-truth of the first frame in the practical applications. To reduce the user effort, inspired by the interactive (e.g., scribbles and points) methods in co-segmentation community [19], [20], we propose an interactive weak supervision strategy to initialize the semi-supervised and unsupervised video segmentation tasks.

As illustrated in Fig.2, instead of labeling the complex ground-truth, we directly employ the user interaction interface to give a weak supervision (e.g., rectangle) to initialize the video sequence. Then, we utilize the image segment method (mask RCNN [38]) to normalize the foreground and background likelihoods to achieve a 2-class reference mask. It is aggregated with the initial representations and viewed as the supplementary reference feature.

2. Earth mover's distance temporal modulation

Inspired by the earth mover's distance based image matching methods [31], [34], [39], we propose the temporal modulation module to enhance the representations of current input frame. The earth mover's distance measures two sets of weighted distributions or objects, which can be formalized to the optimal transportation problem [31]. Specifically, a set of suppliers $S=\{s_i|i=1,2,\dots,m\}$ need to be transported to a set of destination $D=\{d_j|j=1,2,\dots,n\}$, where s_i indicates the unit of supplier i and d_i denotes the destination of the j-th demander. The price of each unit that is taken from the supplier i to destination j is represented by c_{ij}.

In the similar manner, the number of units transported is indicated by π_{ij}. The aim of the transportation problem is finding a optimal flow $\pi^*=\{\pi^*_{ij}|i=1,2,\dots,m,\ \ j=1,2,\dots,n\}$ from suppliers to destinations:

$$\begin{aligned}
\min_{\pi_{ij}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\pi_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n}\pi_{ij}=s_i, \quad i=1,\dots,m \\
& \sum_{i=1}^{m}\pi_{ij}=d_j, \quad j=1,\dots,n \\
& \pi_{ij}\geq 0, \quad i=1,2,\dots,m, \ \ j=1,2,\dots,n
\end{aligned} \tag{1}$$

The d_j and s_i are named as weights of the nodes, which takes control of the total matching flows that are generated by each node. In order to properly constrain the total matching flows $\sum_{j=1}^{n}\pi_{ij}=s_i$ and $\sum_{i=1}^{m}\pi_{ij}=d_i$, the weight of the node i is obtained by the node representation and the average node representation in the other structure:

$$s_i=\max\left\{0,\mathbf{u}_i^{\top}\cdot\frac{1}{WH}\sum_{j=1}^{WH}\mathbf{v}_j\right\} \tag{2}$$

where $\boldsymbol{v}_j$ and $\boldsymbol{u}_i$ indicate the vectors from the two feature maps, respectively, the function $\max(\cdot)$ makes sure that the generated weights are non-negative. For the weights of d_j, we can achieve the value using the same operations. Furthermore, we normalize the weights in the structure by

$$\bar{s}_i=s_i\left(WH\bigg/\sum_{j=1}^{WH}s_j\right) \tag{3}$$

Back to equation (1), in order to obtain c_{ij}, we employ a fully convolutional network [40] to obtain the current-frame image embedding $\boldsymbol{U}\in\mathbb{R}^{H\times W\times C}$, and previous-frame image embedding $\boldsymbol{V}\in\mathbb{R}^{H\times W\times C}$, respectively, where H, W indicate the spatial positions and $\{\boldsymbol{u}_1,\boldsymbol{u}_2,\dots,\boldsymbol{u}_{HW}\}$ is the local representation vectors, while C denotes the channel number. Known the embedding nodes $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ of two coherent frames, the cost of each unit c_{ij} is computed via

$$c_{ij}=1-\frac{\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j}{\|\boldsymbol{u}_i\|\,\|\boldsymbol{v}_j\|} \tag{4}$$

where nodes with similar representations produce less matching cost. As in equation (3), we achieve the node weights $\bar{s_i}$ and $\bar{d_j}$ that gets the balance between two coherent image frames. Then, we obtain the approximate optimal matching flows $\pi^*=\{\pi^*_{ij}|i=1,2,\dots,m,\ \ j=1,2,\dots,n\}$ by the solution of linear programming [34]. Afterwards, we compute the i-th position of the EMD similarity map $\boldsymbol{M}_i$ via

$$\boldsymbol{M}_i = \sum_{j=1}^{HW} (1 - c_{ij})\pi_{ij}^* \tag{5}$$

which is able to assign the large weight to the co-occurrent objects and the background regions have the smaller weights. We weight the current representations $\boldsymbol{U} \in \mathbb{R}^{W \times H \times C}$ according to their degrees similar to the previous frame, yielding the modulated object representation:

$$\bar{\boldsymbol{U}} = \text{softmax}(\boldsymbol{M}) \odot \boldsymbol{U} \tag{6}$$

where $\odot$ refers to the hadamard product, and we leverage the function $\text{softmax}(\cdot)$ to normalize the weights $\boldsymbol{M}$. Based on the theorem of implicit function [34], [41], [42] on the optimality (KKT) conditions, we are able to yield a closed-form gradient of the optimal match with respect to the linear programming parameters, ensuring the back-propagation through the end-to-end training phase.

3. Cross-squeeze temporal modulation

To restrain the noisy and highlight the informative representation for video segmentation, we propose the cross-squeeze temporal modulation module. Motivated by the squeeze and excitation neural network [43] and the cross reference operations in the fields of co-saliency and co-segmentation [20], [44], [45], our CTM squeezes both the current input representations $\boldsymbol{U}$ and the previous representations $\boldsymbol{V}$ to highlight the co-occurrent objects in the frame pair (see Fig.2).

Cross squeeze To mine the jointly owned objects in the pair of images, we simultaneously obtain the global statistics in the two images via cross squeeze operations. Specifically, we firstly utilize the global average pooling operation to produce two channel descriptors that capture the channel-wise contextual cues. Then, the c-th unit of the current weight vector $\boldsymbol{g} \in \mathbb{R}^{C \times 1}$ is acquired by

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{u}_c(i,j) \tag{7}$$

where $\boldsymbol{u}_c(i,j)$ indicates the ij-th unit(position) of the input feature $\boldsymbol{U} \in \mathbb{R}^{H \times W \times C}$ in c-th channel. Then, the global statistic vector $\boldsymbol{f} \in \mathbb{R}^{C \times 1}$ of the previous frame can be obtained in the similar manner. The only difference is that it is applied to the previous frame $\boldsymbol{V} \in \mathbb{R}^{H \times W \times C}$ via

$$f_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{v}_c(i,j) \tag{8}$$

where $\boldsymbol{u}_c(i,j)$ means the ij-th position of the input tensor in c-th channel. Based on the aforementioned global statistic vectors, we obtain the cross squeezed weights $\boldsymbol{h} \in \mathbb{R}^{C \times 1}$ by

$$h_c = f_c \cdot g_c \tag{9}$$

where c represents the index of channel, and $.$ denotes the ordinary scalar multiplication.

Cross excitation Furthermore, the obtained vector $\boldsymbol{h}$ is introduced into the excitation function, getting

$$\boldsymbol{e} = \sigma\left(\boldsymbol{W}_2 \delta\left(\boldsymbol{W}_1 \boldsymbol{h}\right)\right) \tag{10}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ is a parameter matrix (fully connected layer) that has the dimensionality reduction ratio γ, $\boldsymbol{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ has the same ratio γ for dimensionality changing, $\sigma(\cdot)$ denotes a sigmoid function, and a ReLU function $\delta(\cdot)$ is leveraged here. Then, the enhanced feature tensor $\tilde{\boldsymbol{U}} \in \mathbb{R}^{H \times W \times C}$ is computed by re-weighting the input feature $\boldsymbol{U}$:

$$\tilde{\boldsymbol{u}}_c = e_c \cdot \boldsymbol{u}_c \tag{11}$$

where $\tilde{\boldsymbol{u}}_c \in \mathbb{R}^{H \times W}$ indicates the c-th feature map of the tensor $\tilde{\boldsymbol{U}} \in \mathbb{R}^{H \times W \times C}$, and $(\cdot)$ is channel-wise multiplication operation between the feature map $\boldsymbol{u}_c$ and the scalar e_c.

4. Augmentation and decoder sub-network

We augment the earth mover's distance based temporal modulated (ETM) representation $\bar{\boldsymbol{U}}$ and the cross-squeeze temporal modulated (CTM) representation $\tilde{\boldsymbol{U}}$ via

$$\boldsymbol{X} = (1 - \lambda)\,\bar{\boldsymbol{U}} \oplus \lambda \tilde{\boldsymbol{U}} \tag{12}$$

where the λ controls the balance of the two representations, $\oplus$ denotes the add operation, and $\boldsymbol{X}$ indicates the enhanced representation. Also, as a common practice in VOS [46], we extract the optical flow between current frame and previous frame to supplement the representation of current frame. Then, the supplemented representations $\boldsymbol{X}$ (with flow-based representation $\boldsymbol{F}$) are transported to the up-sampling and decoder part, and the whole model is supervised by the binary cross-entropy (BCE) loss [47]. Specifically, the BCE loss is calculated on the ground truth mask and the output predicted map.

Similar to the U-shape in MATNet [46], we leverage a multi-stage decoder to up-sample and predict the target mask. The ETM and CTM modules are simultaneously trained in an end-to-end manner, which utilizes stochastic gradient descent (SGD) as the optimizer.

IV. Experiments

In this part, we briefly describe the experimental settings in Section IV.1. Section IV.2 presents the corresponding UVOS and SVOS datasets and their metrics. Secstion IV.3–IV.4 express the evaluations on UVOS and SVOS benchmarks that includes DAVIS-2016 [23], FBMS [24], DAVIS-2017 [26], and Youtube-VOS [25]. Moreover, Section IV.5, IV.6 and IV.7 show the ablation study, speed analysis and qualitative examples, respectively. Finally, Section IV.8 has relevant discussions of the temporal modulation and the subsistent failure cases.

1. Setup

We implement the proposed RectVOS in Pytorch 1.6 [48] on the platform with a core i7-4790 CPU, single Nvidia RTX 2080Ti GPU, and 16.0 GB RAM.

We utilize the ShuffleNet [49] as backbone, which is pre-trained in several datasets and we freeze the 1–4 layers. Following MATNet [46], the multi-level encoding features are propagated to the decoding phase, which aims to preserve the local conditions in different receptive fields. Specifically, we train the SVOS model using the DAVIS-2017 [26] dataset that contains 150 training videos, and Youtube-VOS [25] dataset that has 4,453 training sequences. For UVOS model, the DAVIS-2016 [23] and FBMS [24] sets are utilized to train the RectVOS. To supervise the UVOS and SVOS model, we employ the cross-entropy(CE) loss to monitor the segmentation.

2. Datasets and evaluation schemes

We evaluate the method on DAVIS-2016 [23], FBMS [24], Youtube-VOS [25], and DAVIS-2017 [26].

Un-supervised VOS datasets We leverage DAVIS-2016 and FBMS benchmarks to evaluate the performance of the unsupervised VOS task. It has 50 full high-resolution videos with thousands of pixel-wise annotations. Following the standard evaluation schemes, based on the contour accuracy, we report the intersection over union and F-measure, respectively.

Semi-supervised VOS datasets For DAVIS-2017 dataset, we leverage the intersection-over-union (IOU) between the ground truth and the predicted mask, which is termed as the Jaccard index $\mathcal{J}$.

For Youtube-VOS benchmark, it contains 4,453 video sequences (including 3,471 training videos, 474 validation videos, and 508 testing videos). Following the metrics in [23], we employ the the contour precision $\mathcal{F}$ and region similarity $\mathcal{J}$ to calculate scores $\mathcal{J}$ [25].

3. Overall quantitative performance on UVOS datasets

Results on the DAVIS-2016 and FBMS datasets To further demonstrate the effectiveness of our RectVOS in the unsupervised situations, we conduct adequate evaluations on the DAVIS-2016 and FBMS unsupervised validation datasets, which includes realistic videos with no initial binary pixel-level labels. Table 1 and Table 2 illustrates the results on the DAVIS-2016 and FBMS datasets.

Our RectVOS obtains the state-of-the-art scores of 84.5% (mean $\mathcal{J}\&\mathcal{F}$) and 77.6% ($\mathcal{J}$) in DAVIS-2016 and FBMS, respectively. We can observe that the proposed method outperforms other competitors over all the metrics across the dataset. This brilliant performance mainly thanks to the weak rectangle supervision that we interactively obtain, which has more instructional information than a completely unsupervised approach. Owing to the full-duplex strategy, the very recent method FSNet [10] obtains the state-of-the-art performance of 83.3%, which is lower than the proposed RectVOS by a margin of 1.2%, which demonstrates the effectiveness of the user interaction proposed by the method. Also, the learned EMD and cross-squeeze based temporal modulation plays significant role in the object representations, especially in the fast-moving video scenarios.

Although F2Net [50] utilizes both the motion and appearance cues to model the spatio-temporal relation and gets a competitive score of 77.5%(−0.1%) than the proposed RectVOS in FBMS [24] benchmark, our approach is capable to model long-range spatio-temporal using the EMD-based temporal modulation, surpassing the MATNet by 2.9% in the mainstream dataset DAVIS-2016. The evaluation results significantly verify the effectiveness of the proposed EMD temporal modulation mechanism in RectVOS that contributes to a robust representation when meeting the challenging video scenarios.

Table 1. Quantitative comparisons against state-of-the-arts on DAVIS-2016 [23] dataset that is widely used in UVOS. The other results with no citations are borrowed from [44], [46]

Method	NLC	CUT	FSEG	LMP	SFL	LVO	UOVOS	ARP	PDB	MotAdapt	AGNN
$\mathcal{F}$	52.3	55.2	65.3	65.9	66.7	72.1	68	65.3	72.1	70.6	79.1
$\mathcal{J}$	55.1	55.2	70.7	70	64.7	75.9	73.9	76.2	77.2	77.2	80.7
$\mathcal{J}\&\mathcal{F}$	53.7	55.2	68	68	65.7	74	71	70.8	74.7	73.9	79.9
Method	LSMO	AGS	AnDiff [51]	COSNet [52]	MATNet [46]	WCS [53]	DFNet [54]	F2Net [50]	RTNet [3]	FSNet [10]	RectVOS(ours)
$\mathcal{F}$	74.5	77.4	80.5	79.4	80.7	80.7	81.8	84.4	83.5	83.1	84.8
$\mathcal{J}$	78.2	79.7	81.7	80.5	82.4	82.2	83.4	83.1	84.8	83.4	84.2
Mean $\mathcal{J}\&\mathcal{F}$	76.4	78.6	81.1	80	81.6	81.5	82.6	83.8	84.2	83.3	84.5

Table 2. Comparison with recent UVOS approaches on the FBMS [24]

Method	NLC	FST	ARP	MSTP	FSEG	IET	OBN
Mean $\mathcal{J}$	44.5	55.5	59.8	60.8	68.4	71.9	73.9
Method	PDB[55]	SFL[56]	COSNet[52]	MATNet[46]	F2Net[50]	IMP[3]	RectVOS(ours)
Mean $\mathcal{J}$	74.0	56.0	75.6	76.1	77.5	77.5	77.6

4. Overall quantitative performance on SVOS datasets

General quantitative evaluations on DAVIS-2017 dataset The general quantitative comparison on DAVIS-2017 [26] validation dataset is listed in the Table 3 using the metrics of Jaccard-based score $\mathcal{J}$, boundary-based score $\mathcal{F}$, and the speed (fps). The table contains several SVOS methods including Premvos [57], RGMP [58], VideoMatch [59], FEELVOS [60], OSVOS [61], AGAME [5], OnAVOS [62], SiamMask [36], FRTM [63], and the proposed RectVOS. We can note that, for mean $\mathcal{J}$, our RectVOS achieves the second place comparing with the other solutions. It is because that the generative VOS model in AGAME has the superior ability on localizing body of the target, but ours has better boundary precision than the compared method AGAME. Hence, our method has a mean $\mathcal{J}\&\mathcal{F}$ score of 70.9%, which is significantly higher than AGAME (70.0%). Moreover, the proposed RectVOS obtains the best performance on the metrics of mean $\mathcal{J}\&\mathcal{F}$ and mean $\mathcal{F}$ with scores of 70.9% and 72.4%. Although the FRTM [63] utilizes fast optimization techniques to predict the target segmentation and achieves mean $\mathcal{J}\&\mathcal{F}$ of 68.8%, the presented RectVOS yields higher performance than FRTM by a significant margin of 2.1%. This is owing to that our method employs the Earth Mover's Distance based similarity maps to temporally modulate the current frame's representation, which is able to highlight the co-existing foreground objects in the coherent frames. As a result, our VOS model has superior segmentation accuracy on consecutive frames.

Table 3. Comparison results of SVOS models on DAVIS-2017 [26] validation dataset. The common metrics are utilized in evaluations, containing mean $\mathcal{J}\&\mathcal{F}$, $\mathcal{J}$, $\mathcal{F}$, and speed (fps)

Method	Mean-$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	OF	Speed
OSVOS [61]	60.3	61.0	66.1	✓	<1
OnAVOS [62]	63.6	61.6	69.1	✓	<1
RGMP [58]	66.7	64.8	68.8	×	4
SiamMask [36]	56.4	54.3	58.8	×	55
VideoMatch [59]	62.4	56.5	68.2	×	4
FEELVOS [60]	69.1	65.9	72.3	×	2
AGSS-VOS [64]	66.6	63.4	69.8	×	10
AGAME [5]	70.0	72.7	67.8	×	30
FRTM [63]	68.8	66.4	71.2	×	8
RectVOS(ours)	70.9	69.6	72.4	×	22

Results on Youtube-VOS dataset Youtube-VOS [25] is a large-scale dataset for semi-supervised VOS, which has both seen and unseen (not exist in the training set) categories. We use the region similarity ($\mathcal{J}$_seen and $\mathcal{J}$_unseen) and contour accuracy ($\mathcal{F}$_seen and $\mathcal{F}$_unseen), respectively. As listed in Table 4, we compare RectVOS with recent SVOS methods including S2S [25], Premvos [65], AGAME [5], OnAVOS [62], RGMP [58], Rvos [65], OSVOS [61], and TVOS [22].

Table 4. Performance comparisons on the SVOS dataset (Youtube_vos validation set). The evaluated metrics are consists of overall performance

Method	Overall	$\mathcal{J}$_seen	$\mathcal{F}$_seen	$\mathcal{J}$_unseen	$\mathcal{F}$_unseen
OnAVOS [62]	0.552	0.601	0.627	0.466	0.514
OSVOS [61]	0.588	0.598	0.605	0.542	0.607
RGMP [58]	0.538	0.595	–	0.452	–
S2S [25]	0.644	0.710	0.700	0.555	0.612
Premvos [57]	0.669	0.714	0.759	0.565	0.637
Rvos [65]	0.568	0.636	0.672	0.455	0.510
AGAME [5]	0.660	0.678	0.695	0.612	0.662
SiamMask [36]	0.528	0.602	0.582	0.451	0.477
TVOS [22]	0.678	0.671	0.694	0.630	0.716
RectVOS(ours)	0.688	0.704	0.696	0.686	0.667

We can note that our RectVOS surpasses the competitors in majority of the metrics. Specifically, in terms of overall performance and $\mathcal{J}$_unseen, we obtain the best score of 0.688 and 0.686, which demonstrates the effectiveness of the proposed user interaction strategy in first frame for SVOS. It has the ability of rough localization in the first frame and even without the target prior, the model is able to find the accurate position of the unseen objects in the training sets. Consequently, our method has superior performance than the competitors in the unseen categories. While in terms of $\mathcal{F}$_seen, $\mathcal{J}$_seen, we has the top 3 performance with scores of 0.704 and 0.696. This is owing to the earth mover's distance temporal modulation, which dynamically handles the drastic appearance variations of the target. The Premvos and S2S has superior performance than ours with tiny margins of ±0.01 and ±0.006, while relying on the large-scale statical image training datasets (borrowing from image semantic segmentation, instance segmentation or saliency detection). It is noted that the heavy training phase makes the model overfitting and hard to generalize to the other unseen objects, which naturally leads to the poorer performance in unseen categories.

Moreover, without extra pre-training or online fine-tuning, our RectVOS achieves the relatively good score of 0.704 and 0.696 in unseen categories, which shows that the proposed method RectVOS has better generalization capability and efficiency than competitors. For the rectangle initialization based method SiamMask, our RectVOS significantly exceeds SiamMask in both seen and unseen categories performance by large margins.

5. Ablation study of various components

To analyze the effectiveness of each variations in RectVOS, we exhibit the RectVOS variations including without the earth mover's distance temporal modulation (ETM), cross-squeeze temporal modulation (CTM), ResNet50, weak user interaction, and optical flow. In detail, we apply FBMS dataset to analyze different variations, which are reported in Table 5 in the term of mean $\mathcal{J}$.

Table 5. By using different part models, we conduct ablative experiments on the UVOS dataset (DAVIS2016). Comparing with the variants

Variations	Mean $\mathcal{J}$	Gain
RectVOS w/o ETM	76.5	±1.1
RectVOS w/o CTM	76.9	±0.7
RectVOS w/o ResNet50	72.4	±5.2
RectVOS w/o weak user interaction	75.5	±2.1
RectVOS w/o optical flow	74.6	±3.0
RectVOS	77.6	±0.0

Without ETM, the model dramatically gains by ±1.1% in mean $\mathcal{J}$, showing the effectiveness of the proposed earth mover's distance temporal modulation module. Then, the RectVOS (without CTM) performs superior than the RectVOS-ETM principally owing to the reason that the RectVOS only with CTM fails to segment accurately when the object suffering from heavy occlusion, which drops from 77.6% to 76.9%. Afterwards, if the backbone changes from ResNet101 to ResNet50, as reported in the Table 5, the performance of the RectVOS (without ResNet50) drops from 77.6% to 72.4%, which has a large gain of 5.2%. Moreover, without weak user interaction in first frame, the performance has a decline of 2.1% because of the absence of unseen categories. Also, if dropping the optical flow between successive frames, the proposed method reduces by 3.0% in mean $\mathcal{J}$.

6. Speed and complexity analysis

We can note that in Table 3 our method runs at 22 fps in the phase of inference. Specifically, we implement the speed analysis in the platform with an Intel Core 4790 CPU, 16GB RAM, and a Nvidia RTX 2080Ti GPU on DAVIS-2017 dataset. As in Table 3, OnAVOS, OSVOS, and Premvos nearly run at the speed of lower than 1 fps, which is mainly because of the time-consuming online finetuning in each frame. In addition, the inference speeds of FEELVOS, RGMP, VideoMatch, and FRTM are less than 10 *fps*, and runs faster than these online fine-tuning solutions. The proposed RectVOS obtains the third place with a fps of 22, which is lower than SiamMask and AGAME. Although the SiamMask and AGAME obtain better speed performance than ours, our RectVOS yields much better performance in mean $\mathcal{J}\&\mathcal{F}$ by the improvements of 14.9% and 0.9%, which demonstrates the balance between accuracy and speed in the proposed RectVOS. Furthermore, we have investigated the complexity of each module in the proposed RectVOS. The time complexity of weak user interaction module is depended on human factors, which is not a fixed time. Moreover, the CTM takes 0.006 seconds to handle one frame. Similar to DeepEMD [34], the EMD layer in ETM can be accelerated by the QPTH libraty, which takes 0.021 seconds. In a word, without the online fine-tuning and post processing, our method RectVOS achieves favorable balance between accuracy and speed against state-of-the-art approaches.

7. Qualitative analysis

Fig.3 qualitatively shows remarkable instances of the proposed RectVOS on Youtube dataset and DAVIS-2017 dataset. We select several frames at various time on 6 sequences, from top to bottom where are *parrots*, *sea snake*, *owls*, *bus*, *dancer*, and *dog*. It is noted that, the presented RectVOS favourably handles the challenging scenarios (e.g. heavy occlusion and drastic appearance variation) in sequences *parrots* and *sea snake*. In the 1st row, although the parrot is occluded by the similar parrot in the background, the proposed EMD based temporal modulation module accurately distinguishes foreground and background areas. In sequences *sea snake* and *owls*, owing to the weak user interaction of the initialization, our method successfully localizes the position of the uncommon categories e.g. snake and owl. While in the forth row, the presented model has a superior performance in the challenging scene in video *bus*, which suffers from occlusion and fast moving. Moreover, the last two sequences *dancer* and *dog* have the fast-changing appearances as time goes by, which leads to the inaccurate contour segmentation. However, with the long-range temporal modulation, our model results in the robust representations and obtains the accurate pixel-wise predictions.

8. Discussions

Temporal Modulation To enhance the temporal consistency between the current frame and the key frame, we propose the two different temporal modulation strategy to enhance the target representations. In this work, we directly set a keyframe every ten frames (e.g., 1, 11, 21,...). As illustrated in Fig.2 , the cross-squeeze temporal modulation and the earth mover's distance temporal modulation both play vital role in the modulation phase. Furthermore, the modulation module is able to transmit the temporal slow-changing information from the past video frames to the current frame. The purpose of temporal modulation is to highlight the co-occurrent objects in the sequence. When the target suffers from fast moving, the learned representations are polluted by the low-quality frames. However, thanks to the EMD-based (ETM) and cross temporal modulation (CTM), our model is capable of establishing the temporal consistency and relieves the inaccurate contour brought by fast motion. In summary, it alleviates the segmentation of target contour when meeting the fast-moving scenarios.

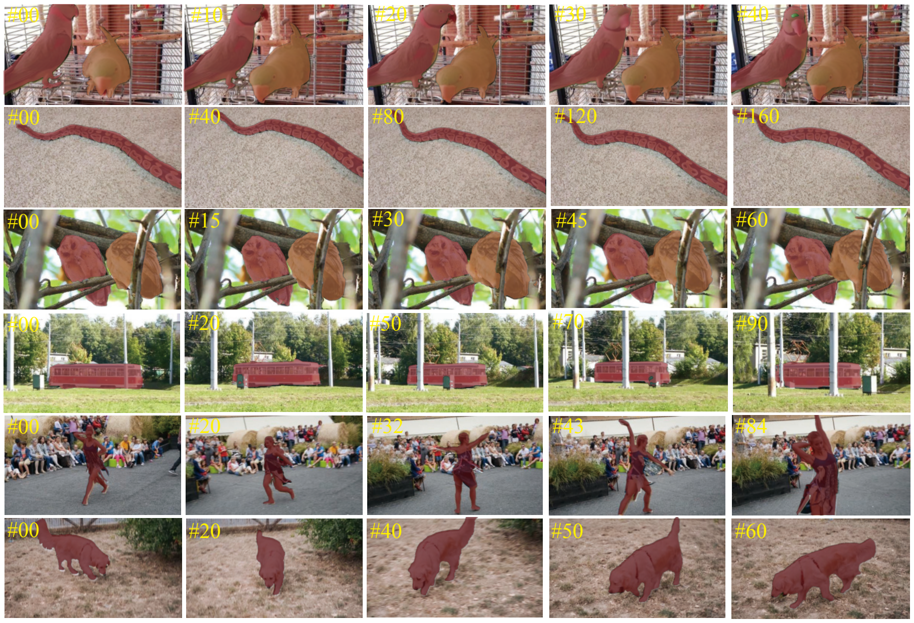


Fig. 3. Qualitative examples of the proposed RectVOS on Youtube-Objects dataset [25] and DAVIS-2017 benchmark [26].

Failure Cases It illustrates some failure cases of the proposed RectVOS in Fig.4.

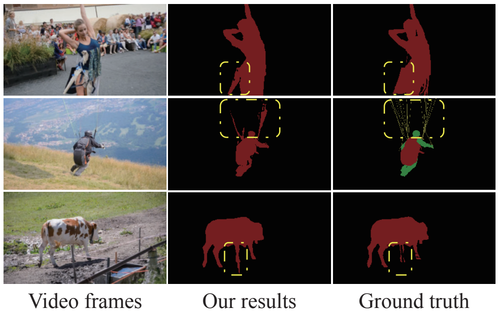


Fig. 4. Several failure cases of our RectVOS.

In the first row of Fig.4, the RectVOS cannot distinguish the contour of tiny components in sequence *dancer*. This is because that the dancer moves in a very high speed, which limits the modulation ability of ETM and CTM. In the second case, our method is capable of discriminating the player, his backpack, and his arm, but the model fails to find the ropes in the air. This phenomenon is caused by the limited learning ability of the model, the EMD based modulation model has less discriminative ability on background area. Moreover, in the last sequence, our RectVOS inaccurately distinguishes the poles as the foreground. Our method is not able to segment the pole as it has very similar appearance with the leg of *cow*. Also, the wood poles have the challenging of low resolution, the RectVOS model cannot accurately segment the contour of the target. This is because of that the CNNs is easy to limited to local area. In the future work, we will relieve the issue by using the Transformer to establish the long-range dependencies.

V. Conclusions and Future Work

In this work, we have presented a weakly-interactive VOS architecture with rectangle annotation as prior for accurate VOS. We first manually draw the weak user interaction (rectangle) of the target area in the initial frame. Second, we obtain the EMD relation map between the current and key frame and subsequently highlight the current representations using the learned similarity to temporally modulate the object representations in the current frame. Then, to emphasize the salient representations of the raw representations, we present a novel CTM module to enhance the robustness of the learned features against the drastic appearance variations. Finally, both ETM and CTM features are augmented with the raw representation to produce the refined spatio-temporal representation, which is propagated into the decoder and predict the segmentation mask. Extensive evaluations have showed the benefits of the proposed RectVOS approach against the state-of-the-art approaches.

The related code will be released at: https://github.com/liyuwang2016/RectVOS.

References

[1] Z. Zhang, B. L. Wang, Z. Z. Yu, *et al.*, "Dilated convolutional pixels affinity network for weakly supervised semantic segmentation," *Chinese Journal of Electronics*, vol.30, no.6, pp.1120–1130, 2021.

[2] W. L. Qiu, X. B. Gao, and B. Han, "Video saliency detection via pairwise interaction," *Chinese Journal of Electronics*, vol.29, no.3, pp.427–436, 2020.

[3] S. C. Ren, W. X. Liu, Y. T. Liu, *et al.*, "Reciprocal transformations for unsupervised video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp.15430–15439, 2021.

[4] P. L. Huang, J. W. Han, N. Liu, *et al.*, "Scribble-supervised video object segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol.9, no.2, pp.339–353, 2021.

[5] J. Johnander, M. Danelljan, E. Brissman, *et al.*, "A generative appearance model for end-to-end video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.8945–8954, 2019.

[6] P. S. Wen, R. L. Yang, Q. Q. Xu, *et al.*, "DMVOS: Discriminative matching for real-time video object segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, pp.2048–2056, 2020.

[7] H. C. Wang, X. L. Jiang, H. B. Ren, *et al.*, "SwiftNet: Real-time video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp.1296–1305, 2021.

[8] L. Wang, G. Hua, R. Sukthankar, *et al.*, "Video object discovery and co-segmentation with extremely weak supervision," in *Proceedings of 13th European Conference on Computer Vision*, Zurich, Switzerland, pp.640–655, 2014.

[9] J. W. Han, L. Yang, D. W. Zhang, *et al.*, "Reinforcement cutting-agent learning for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.9080–9089, 2018.

[10] G. P. Ji, K. R. Fu, Z. Wu, *et al.*, "Full-duplex strategy for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp.4902–4913, 2021.

[11] A. Azulay, T. Halperin, O. Vantzos, *et al.*, "Temporally stable video segmentation without video annotations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, pp.1919–1928, 2022.

[12] B. Luo, H. L. Li, F. M. Meng, *et al.*, "Video object segmentation via global consistency aware query strategy," *IEEE Transactions on Multimedia*, vol.19, no.7, pp.1482–1493, 2017.

[13] Z. Y. Yin, J. Zheng, W. X. Luo, *et al.*, "Learning to recommend frame for interactive video object segmentation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp.15440–15449, 2021.

[14] A. Agarwala, A. Hertzmann, D. H. Salesin, *et al.*, "Keyframe-based tracking for rotoscoping and animation," *ACM Transactions on Graphics*, vol.23, no.3, pp.584–591, 2004.

[15] W. B. Li, F. Viola, J. Starck, *et al.*, "Roto++ accelerating professional rotoscoping using shape manifolds," *ACM Transactions on Graphics*, vol.35, no.4, article no.62, 2016.

[16] N. S. Nagaraja, F. R. Schmidt, and T. Brox, "Video segmentation with just a few strokes," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 3235–3243, 2015.

[17] L. J. Yang, Y. R. Wang, X. H. Xiong, *et al.*, "Efficient video object segmentation via network modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.6499–6507, 2018.

[18] A. Benard and M. Gygli, "Interactive video object segmentation in the wild," *arXiv preprint*, arXiv: 1801.00269, 2017.

[19] S. W. Oh, J. Y. Lee, N. Xu, *et al.*, "Fast user-guided video object segmentation by interaction-and-propagation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.5242–5251, 2019.

[20] D. Batra, A. Kowdle, D. Parikh, *et al.*, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp.3169–3176, 2010.

[21] K. Xu, L. Y. Wen, G. R. Li, *et al.*, "Spatiotemporal CNN for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.1379–1388, 2019.

[22] Y. Z. Zhang, Z. R. Wu, H. W. Peng, *et al.*, "A transductive approach for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.6947–6956, 2020.

[23] F. Perazzi, J. Pont-Tuset, B. McWilliams, *et al.*, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.724–732, 2016.

[24] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.36, no.6, pp.1187–1200, 2014.

[25] N. Xu, L. J. Yang, Y. C. Fan, *et al.*, "YouTube-VOS: sequence-to-sequence video object segmentation," in *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, Munich, Germany, pp.603–619, 2018.

[26] J. Pont-Tuset, F. Perazzi, S. Caelles, *et al.*, The 2017 DAVIS challenge on video object segmentation, *arXiv preprint*, arXiv: 1704.00675, 2017.

[27] D. W. Zhang, J. W. Han, L. Yang, *et al.*, "SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.2, pp.475–489, 2020.

[28] J. X. Miao, Y. C. Wei, and Y. Yang, "Memory aggregation networks for efficient interactive video object

segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.10363–10372, 2020.

[29] Y. Y. Mao, N. Wang, W. G. Zhou, *et al.*, "Joint inductive and transductive learning for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp.9650–9659, 2021.

[30] B. Duke, A. Ahmed, C. Wolf, *et al.*, "SSTVOS: Sparse spatiotemporal transformers for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp.5908–5917, 2021.

[31] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover' s distance as a metric for image retrieval," *International Journal of Computer Vision*, vol.40, no.2, pp.99–121, 2000.

[32] H. B. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.5, pp.840–853, 2007.

[33] L. Hou, C. P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint*, arXiv: 1611.05916, 2016.

[34] C. Zhang, Y. J. Cai, G. S. Lin, *et al.*, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.12200–12210, 2020.

[35] D. Yeo, J. Son, B. Han, *et al.*, "Superpixel-based tracking-by-segmentation using Markov chains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.511–520, 2017.

[36] Q. Wang, L. Zhang, L. Bertinetto, *et al.*, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.1328–1338, 2019.

[37] P. Voigtlaender, J. Luiten, P. H. S. Torr, *et al.*, "Siam R-CNN: Visual tracking by re-detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.6577–6587, 2020.

[38] K. M. He, G. Gkioxari, P. Dollár, *et al.*, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp.2980–2988, 2017.

[39] S. T. Liu, Z. M. Li, and J. Sun, "Self-EMD: Self-supervised object detection without ImageNet," *arXiv preprint*, arXiv: 2011.13677, 2020.

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 3431–3440, 2015.

[41] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem: History, Theory, and Applications*. Birkhäuser, New York, NY, USA, doi: 10.1007/978-1-4614-5981-1, 2013.

[42] S. Barratt, "On the differentiability of the solution to convex optimization problems," *arXiv preprint*, arXiv: 1804.05098, 2018.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.7132–7141, 2018.

[44] H. Z. Fu, X. C. Cao, and Z. W. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol.22, no.10, pp.3766–3778, 2013.

[45] W. D. Liu, C. Zhang, G. S. Lin, *et al.*, "CRNet: cross-reference networks for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.4164–4172, 2020.

[46] T. F. Zhou, J. W. Li, S. Z. Wang, *et al.*, "MATNet: Motion-attentive transition network for zero-shot video object segmentation," *IEEE Transactions on Image Processing*, vol.29, pp.8326–8338, 2020.

[47] S. Yang, L. Zhang, J. Q. Qi, *et al.*, "Learning motion-appearance co-attention for zero-shot video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp.1544–1553, 2021.

[48] A. Paszke, S. Gross, S. Chintala, *et al.*, "Automatic differentiation in PyTorch," in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

[49] N. N. Ma, X. Y. Zhang, H. T. Zheng, *et al.*, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp.122–138, 2018.

[50] D. Z. Liu, D. D. Yu, C. H. Wang, *et al.*, "F2Net: Learning to focus on the foreground for unsupervised video object segmentation," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Vancouver, British Columbia, Canada, pp.2109–2117, 2021.

[51] Y. Q. Wang, Z. L. Xu, H. Shen, *et al.*, "CenterMask: Single shot instance segmentation with point representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.9310–9318, 2020.

[52] X. K. Lu, W. G. Wang, C. Ma, *et al.*, "See more, know more: Unsupervised video object segmentation with co-attention Siamese networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.3618–3627, 2019.

[53] L. Zhang, J. M. Zhang, Z. Lin, *et al.*, "Unsupervised video object segmentation with joint hotspot tracking," *in 16th European Conference on Computer Vision*, Glasgow, UK, pp. 490–506, 2020.

[54] M. M. Zhen, S. W. Li, L. Zhou, *et al.*, "Learning discriminative feature with CRF for unsupervised video object segmentation," in *16th European Conference on Computer Vision*, Glasgow, UK, pp.445–462, 2020.

[55] H. M. Song, W. G. Wang, S. Y. Zhao, *et al.*, "Pyramid dilated deeper ConvLSTM for video salient object detection," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp.744–760, 2018.

[56] J. C. Cheng, Y. H. Tsai, S. J. Wang, *et al.*, "SegFlow: Joint learning for video object segmentation and optical flow," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp.686–695, 2017.

[57] J. Luiten, P. Voigtlaender, and B. Leibe, "PReMVOS: Proposal-generation, refinement and merging for video object segmentation," in *14th Asian Conference on Computer Vision*, Perth, Australia, pp.565–580, 2019.

[58] S. W. Oh, J. Y. Lee, K. Sunkavalli, *et al.*, "Fast video object segmentation by reference-guided mask propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.7376–7385, 2018.

[59] Y. T. Hu, J. B. Huang, and A. G. Schwing, "VideoMatch: Matching based video object segmentation," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp.56–73, 2018.

[60] P. Voigtlaender, Y. N. Chai, F. Schroff, *et al.*, "FEELVOS: Fast end-to-end embedding learning for video object seg-

mentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.9473–9482, 2019.

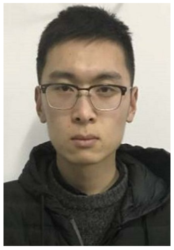
[61] S. Caelles, K. K. Maninis, J. Pont-Tuset, *et al.*, "One-shot video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.5320–5329, 2017.

[62] P. Voigtlaender and B. Leibe, "Online adaptation of convolutional neural networks for video object segmentation," in *British Machine Vision Conference*, London, UK, 2017.

[63] A. Robinson, F. J. Lawin, M. Danelljan, *et al.*, "Learning fast and robust target models for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.7404–7413, 2020.

[64] H. J. Lin, X. J. Qi, and J. Y. Jia, "AGSS-VOS: Attention guided single-shot video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp.3948–3956, 2019.

[65] C. Ventura, M. Bellver, A. Girbau, *et al.*, "RVOS: End-to-end recurrent network for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.5272–5281, 2019.



FAN Jiaqing received the M.S. degree from the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China, in 2019. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His research interests include video object segmentation.
(Email: jqfan@nuaa.edu.cn)

LIU Qingshan (corresponding author) is a Professor with Nanjing University of Information Science and Technology, Nanjing, China. He received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, Beijing, China, in 2003. He was an Assistant Research Professor with the Department of Computer Science, Computational Biomedicine Imaging and Modeling Center, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA, from 2010 to 2011. His current research interests are image and vision analysis, including face image analysis, graph and hypergraph-based image and video understanding.
(Email: qsliu@nuist.edu.cn)