Unsupervised Video Object Segmentation with Joint Hotspot Tracking

Lu Zhang[1], Jianming Zhang[2], Zhe Lin[2], Radomír Měch[2], Huchuan Lu[1⋆], and You He[3]

[1] Dalian University of Technology
[2] Adobe Research
[3] Naval Aviation University

Abstract. Object tracking is a well-studied problem in computer vision while identifying salient spots of objects in a video is a less explored direction in the literature. Video eye gaze estimation methods aim to tackle a related task but salient spots in those methods are not bounded by objects and tend to produce very scattered, unstable predictions due to the noisy ground truth data. We reformulate the problem of detecting and tracking of salient object spots as a new task called *object hotspot tracking*. In this paper, we propose to tackle this task jointly with unsupervised video object segmentation, in real-time, with a unified framework to exploit the synergy between the two. Specifically, we propose a Weighted Correlation Siamese Network (WCS-Net) which employs a Weighted Correlation Block (WCB) for encoding the pixel-wise correspondence between a template frame and the search frame. In addition, WCB takes the initial mask / hotspot as guidance to enhance the influence of salient regions for robust tracking. Our system can operate online during inference and jointly produce the object mask and hotspot track-lets at 33 FPS. Experimental results validate the effectiveness of our network design, and show the benefits of jointly solving the hotspot tracking and object segmentation problems. In particular, our method performs favorably against state-of-the-art video eye gaze models in object hotspot tracking, and outperforms existing methods on three benchmark datasets for unsupervised video object segmentation.

Keywords: Unsupervised video object segmentation; Hotspot tracking; Weighted correlation siamese network

1 Introduction

Unsupervised video object segmentation (UVOS) aims to generate the masks of the primary objects in the video sequence without any human annotation. It has attracted a lot of interests due to the wide application scenarios such as autonomous driving, video surveillance and video editing.

However, in some applications, such as video cropping, zooming and autofocus, tracking the full mask of a dominant object may not be sufficient. In video

⋆ Corresponding author

Fig. 1: The visual comparison of our object hotspot tracking results (top row) with the video eye gaze estimation results in [22] (middle row) and the eye gaze tracking ground truth in [48]. The peaks of hotspot map indicate the saliency distribution inside the object. Please zoom in for details.

zooming, for example, the deformable shape of object mask may not provide a stable focal point for the zooming window. Furthermore, the mask only can't support the users to zoom in to the attractive part of the object (*e.g.,* the head of a person or the hood of a car). In such scenario, the attention distribution inside the object region can be very useful. The peak of the attention distribution can be thought of as a virtual focal region of the object, which provides a clean saliency indication on the object part (see the highlighted regions in the first row of Fig. 1). We refer to this virtual focal region of an object as an Object Hotspot. In this work, we propose a new task of Hotspot Tracking, which aims to produce clean, stable and temporally consistent object hotspot distribution along video sequence (see the first row in Fig. 1). This task is different from the video eye gaze prediction task [22, 48] in which the eye gaze tracking ground truth is intrinsically unstable and noisy due to the saccadic eye movement [36, 13]. However, our hotspot tracking is a clean and stable intra-object focal saliency estimation and thus a more useful evidence to facilitate the video editing applications.

The UVOS and hotspot tracking are highly related tasks. The UVOS provides a sequence of object masks, which can provide a strong spatial guidance for the hotspot detection and tracking. Meanwhile, the hotspot of an object can also help the object mask tracking, as it tends to be more robust to object deformation, occlusion and appearance change. Jointly solving both tasks can leverage the synergy between them and provides a richer supervision for the model training.

In this paper, we show how to tackle the unsupervised video object segmentation and hotspot tracking in a simple unified framework. Our model consists of two modules, Target Object Initialization (TOI) and Object Segmentation and Hotspot Tracking (OSHT). The TOI module exploits the idea of the correlation between human eye gaze and UVOS [48] and builds an efficient method for automatically determining the target object in the video. Specifically, for an initial frame, an Eye Gaze Network (EG-Net) is first utilized to make a rough estimation on the location of the target object. Then a Gaze2Mask Network

(Gaze2Mask-Net) is proposed to predict the mask of the target object according to the distribution of eye gaze. By applying the automatic initialization process on the first frame, our model can perform online inference. Furthermore, this approach also allows interactive tracking, where the eye gaze distribution can be indicated by a real human via eye gaze sensors or gestures.

To capture the temporal dependency among video frames, we formulate the task of UVOS and hotspot tracking as a template matching problem. In detail, we employ the eye gaze map and mask from TOI as a template for producing the corresponding hotspot and mask in the subsequent frames. We propose a novel Weighted Correlation Siamese Network (WCS-Net) for the joint tracking of object mask and hotspot. In the WCS-Net, the Weighted Correlation Block (WCB) is exploited to calculate the cross-pixel correspondence between a template frame and the current search frame. In WCB, a weighted pooling operation is conducted between the mask / hotspot and feature of template frame to emphasize the contribution of foreground pixels. The WCB is built on multi-level side-outputs to encode the correlation features, which will be used in the parallel decoder branches to produce the track-lets of mask and hotspot across time.

Considering the difficulty in annotating hotspot using eye tracker, we design a training strategy on multi-source data. The network learns to predict the intra-object hotspots from large-scale image-level eye gaze data [19] and the temporal correlation feature representation via limited video object segmentation data [34]. Once trained, our model can operate online and produce the object mask and hotspot at a real-time speed of 33 fps. To investigate the efficacy of our proposed model, we conduct thorough experiments including overall comparison and ablation study on three benchmark datasets [34, 29, 12]. The results show that our proposed method performs favorable against state-of-the-arts on UVOS. Meanwhile, our model could achieve promising results on hotspot tracking with only training on image-level data.

Our contributions can be summarized as follows:

- We introduce the problem of video object hotspot tracking, which tries to consistently track the focal saliency distribution of target object. This task is beneficial for applications like video editing and camera auto-focus.
- We present a multi-task systematic design for unsupervised video object segmentation and hotspot tracking, which supports both online real-time inference and user interaction.
- We conduct experiments on three benchmark datasets (DAVIS-2016 [34], SegTrackv2 [29] and Youtube-Objects [12]) to demonstrate that our model[1] performs favorable on both tasks while being able to run at 33 FPS.

2 Related Work

Unsupervised VOS. Different from SVOS, unsupervised VOS (UVOS) methods aim to automatically detect the target object without any human definition.

[1] Project Page: https://github.com/luzhangada/code-for-WCS-Net

Early approaches use the low-level cues like visual saliency [47, 18] or optical flow [23, 31, 21, 30] to segment the objects. With the recent success of CNNs on segmentation tasks [7, 49, 50, 16, 56, 54, 9, 8, 55, 57, 10], many methods [31, 5] develop various network architectures and have achieved promising performance. For example, some models [20, 42, 31] propose to build a bilateral network to combine the appearance feature and motion cues. In [20], two parallel streams are built to extract features from raw image and optical flow, which are further fused in the decoder for predicting the segmentation results. Except for the bilateral network, recurrent neural network (RNN) is also exploited in UVOS to capture temporal information [48, 39]. However, the RNN-based methods are not robust enough to handle the long-term videos. Inspired by the success of siamese network [27, 28] in video object tracking, several methods [3, 30–32, 53] propose to build siamese architecture to extract features for two frames to calculate their pixel-wise correspondence. [30, 31] construct a seed-propagation method to obtain segmentation mask, in which the foreground/background seeds are obtained through temporal link. Yang *et al.* [53] propose an anchor diffusion module which computes the similarity of each frame with an anchor frame. The anchor-diffusion is then combined with the intra-diffusion to predict the segmentation results. In this paper, we put forward a weighted correlation siamese network for UVOS. Our model is different from the aforementioned methods on three folds. First, our WCS-Net is designed as a multi-task architecture for jointly addressing UVOS and hotspot tracking. Second, a weighted correlation block is proposed, which exploits the mask and hotspot from the template frame to highlight the influence of foreground pixel in the correlation feature calculation. Third, our correlation block is performed on multiple side-outputs to capture the rich CNN feature.

Video Eye Gaze Tracking. Video eye gaze tracking aims to record human attention in the dynamic scenario. The CNN+RNN framework is widely used in this task to encode spatio-temporal information. Wang *et al.* [46] propose to learn spatial attention from static data and utilize LSTM to predict temporal eye tracking. In [22], the spatial and temporal cues are respectively produced from objectness sub-net and motion sub-net, which are further integrated to produce eye tracking by LSTM. The current eye tracking training data are annotated with eye tracker in a free-view manner, which would incur eye gaze shifting among several salient objects in the video. This limits the existing models to apply on the video editing tasks. To address this issue, Wang *et al.* [48] construct eye gaze data on existing video object segmentation datasets, where the eye fixation is consistent with the moving object. However, due to the eye tracker annotation, the proposed eye gaze data often flickers inside the object. In this paper, we propose a problem of object hotspot tracking, which aims to produce a clean and temporally consistent track-let for the salient region inside the target object.

3 Methodology

We present a neural network model for the task of unsupervised video object segmentation and hotspot tracking. Given a video sequence $I_i \in \{I_1, ..., I_N\}$, our

model aims to produce the corresponding object masks $\{M_i\}_{i=1}^N$ and hotspots $\{H_i\}_{i=1}^N$. Our model contains two sub-modules, Target Object Initialization (TOI) and Object Segmentation and Hotspot Tracking (OSHT). We formulate our task as a template matching problem, where the TOI module determines the target object of the video sequence and the OSHT module aims to predict the target object mask and its hotspot across time. In the subsequent sections, we will introduce the details of TOI, OSHT, and the training strategy, respectively.

3.1 Target Object Initialization

In the UVOS, it is very important to determine what the target object is to be segmented in a video sequence. Some recent CNN-based models [48, 32] propose to incorporate visual attention to localize the target object in a video, where the attention is temporally propagated with the RNN architecture. These works have demonstrated the correlation between human attention and video object determination. Inspired by this, we build a target object initialization module to identify the target object from the guidance of human eye gaze. Different from the previous methods [48, 32] where the human attention is encoded implicitly in the LSTM, we propose an Eye Gaze Network (EG-Net) to make an explicit eye gaze estimation on the initial frame and a Gaze to Mask Network (Gaze2Mask-Net) to generate the corresponding object mask. The advantages of our TOI module are two folds. First, instead of scanning multiple frames during object determination [3, 30, 31], using single frame could largely reduce the computational cost and meet the applications in real-time scenarios. Second, the initialization order from eye gaze to object mask makes it possible for our model to extend in the interactive applications where the eye gaze data can be easily acquired from user click or eye tracker.

Eye gaze network. Given a template frame I^t, the EG-Net aims to produce an eye gaze estimation E^t to indicate the location of the target object in the video. Considering the network efficiency, we build our EG-Net on the EfficientNet [40], which is a very small-size network with impressive performance. Specifically, we first exploit the encoder to extract features for the template frame I^t. Here, we use the last three level features, which are represented as $F^E = \{f_j^E\}_{j=3}^5$ (the feature size can be referred in Fig. 2). In the decoder, we stack three residual refinement blocks [6] to produce the eye gaze map in a coarse-to-fine manner, which can be represented as follows:

$$O_j^E = Conv^2(Cat(f_j^E, Up(O_{j+1}^E))) + Up(O_{j+1}^E) \tag{1}$$

where $Up()$ is the upsampling operation with stride 2. $Cat()$ is the channel-wise concatenation operation. $Conv^2()$ indicates the operation with two convolutional layers. O_j^E is the output of current residual block. Note that the term $Up(O_{j+1}^E)$ is ignored when $j = 5$. We take the final output from 3rd decoder block as the eye gaze prediction E^t.

Gaze to mask network. With the EG-Net, we can obtain an eye gaze estimation on the template frame. The highlighted region is capable of indicating

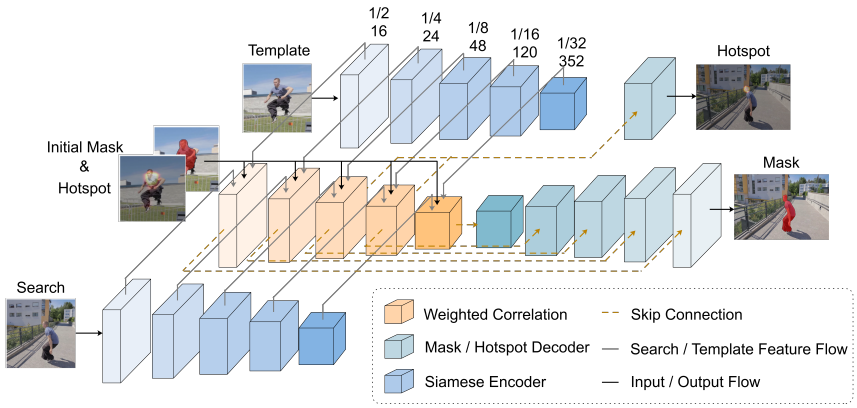

Fig. 2: The framework of Weighted Correlation Siamese Network (WCS-Net). The WCS-Net is a siamese architecture for jointly video object segmentation and hotspot tracking. Given the template frame and search frame, we first use siamese encoder to extract their multi-level features. The weighted correlation block, which takes the initial mask and eye gaze map (see Sec.3.2) as guidance, is exploited to calculate the correspondence between template and search features. Then we build two parallel decoder branches to generate the hotspot and mask for the search frame.

the location of the target region. In order to generate the mask of the target object, we propose a Gaze to Mask Network (Gaze2Mask-Net). Given the template frame I^t and its eye gaze map E^t, the Gaze2Mask-Net aims to segment out the object mask M^t according to the highlight region in eye gaze map. In the Gaze2Mask-Net, we also exploit EfficientNet [40] as encoder. To emphasize the guidance of eye gaze, the encoder takes the concatenation of both template frame and eye gaze map as the input. We extract five level features, which are represented as $F^{G2M} = \{f_j^{G2M}\}_{j=1}^{5}$. In Gaze2Mask-Net, we utilize a revised residual block, which adds the eye gaze map E^t into architecture.

$$O_j^{G2M} = Conv^2(Cat(f_j^{G2M}, Up(O_{j+1}^{G2M}), E^t)) + Up(O_{j+1}^{G2M}) \tag{2}$$

O_j^{G2M} is the output of the revised residual block. Note that the eye gaze map E^t should be resized according to the corresponding feature resolution. We stack five residual blocks as decoder and exploit the output from the first block O_1^{G2M} as object mask M^t.

3.2 Object Segmentation and Hotspot Tracking

The previous UVOS methods [48, 32, 17] usually utilize LSTM or mask propagation method to capture the temporal consistency between adjacent frames. While these methods are not effective enough to handle the object in long-term video. Recent tracking approaches [27, 28, 44] propose a siamese architecture,

which uses the cross correlation operation to capture the dependence between template and search region. Their works have demonstrated the robustness of such siamese network in long-term videos. Inspired by this, we formulate the task of VOS and hotspot tracking as template matching problem. We propose a Weighted Correlation Siamese Network (WCS-Net) for joint unsupervised video object segmentation and hotspot tracking. The overall architecture of WCS-Net is shown in Fig. 2. Given the template frame I^t and the search region of current frame I^i, we first build a siamese encoder network to extract their multi-level features, which are represented as $F^t = \{f_j^t\}_{j=1}^5$ and $F^i = \{f_j^i\}_{j=1}^5$. Then the weighted correlation block is constructed among multiple side-outputs to calculate the multi-level correlation features. Taken the eye gaze map E^t and mask M^t from EG-Net and Gaze2Mask-Net as template guidance, we implement two parallel decoder branches to generate the corresponding hotspot H_i and mask M_i for each frame in the video sequence.

Weighted correlation block. We propose a weighted correlation block to calculate the correspondence between template and search features. Taken the template feature f_j^t, search feature f_j^i and the template guidance G as input, the weighted correlation block produces the corresponding correlation feature by:

$$C_j^i = W(f_j^t, R_j(G)) \star f_j^i \tag{3}$$

where $\star$ denotes the depth-wise cross correlation layer [27, 44]. C_j^i is the correlation feature between template and search frames. $R_j()$ is the resize operation. $W()$ indicates the weighted pooling operation, which transfers the feature map $(h \times w \times c)$ into feature vector $(1 \times 1 \times c)$ by weighted summation with $R_j(G)$. $G = M^t$ when constructing the correlation block for the video object segmentation, or $G = H^t$ for the hotspot tracking (see Fig. 3 for more details). Compared with the original cross correlation [27, 44] (formulated as $f_j^t \star f_j^i$), our weighted correlation block is more effective at our pixel-wise prediction problem. On one hand, the weighted pooling with template guidance is able to highlight the contribution of foreground and reduce the noise from background. On the other hand, the correlation between template vector and search feature would not decrease the resolution of search feature and thus helps to remain the details of target object. We conduct the comparison experiment in Sec. 4.4 to demonstrate the effectiveness of our weighted correlation on UVOS.

Mask decoder branch. Different from previous methods [48, 32] in which only the single level feature is used to generate object mask, we exploit multi-level feature representations. Specifically, we build the weighted correlation blocks among five side-outputs to produce the multi-level correlation feature between template and search frames. Similarly with the decoder in EG-Net, we stack five residual refinement blocks to produce the object mask in a coarse to fine manner.

$$O_j^{i,M} = Conv^2(Cat(C_j^{i,M}, Up(O_{j+1}^{i,M}))) + Up(O_{j+1}^{i,M}) \tag{4}$$

where the $O_j^{i,M}$ represents the the output in the j-th decoder residual block. $C_j^{i,M}$ is the correlation feature calculated by weighted correlation block with

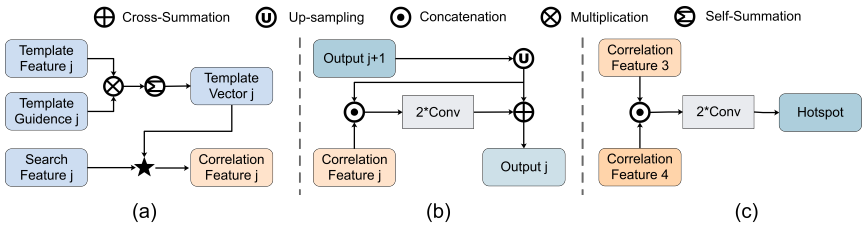

Fig. 3: The details of components in weighted correlation siamese network. (a) The details of weighted correlation block at j-th level. (b) The details of mask decoder block at j-th level. (c) The details of hotspot decoder block.

initial mask M^t as template guidance (*i.e.,* $G = M^t$ in Eq. 3). We take the output from the 1st residual decoder block as the final object mask M_i for i-th frame. The architecture of the decoder block for mask branch is illustrated in Fig. 3 (b).

Hotspot decoder branch. The function of hotspot decoder is to generate the track-let of hotspots $\{H^i\}_{i=1}^N$ for the video sequence according to the initial eye gaze. We visiualize the framework of hotspot decoder branch in Fig. 3 (c). In the hotspot decoder branch, the correlation features on both 3-rd and 4-th side-outputs are used for generating the hotspot map,

$$O^{i,H} = Conv^2(Cat(C_3^{i,H}, Up(C_4^{i,H})) \tag{5}$$

where $O^{i,H}$ is the hotspot map H^i for i-th frame. $C_3^{i,H}$ and $C_4^{i,H}$ are the correlation features by Eq. 3 with G set as initial eye gaze map E^t.

3.3 Network Training

Implementation details. The input size of EG-Net and Gaze2Mask-Net is set to 320×320. In the WCS-Net, the template and search frames fed into the siamese encoder network share the same resolution as 320×320. Similar with the object tracking methods [27, 44], we take the sub-region centered on the target object as template frame. The current search region is a local crop centered on the last estimated position of the target object. We use the EfficientNetv2 [40] as the backbone of siamese encoder. For all the residual refinement blocks in the decoders of EG-Net, Gaze2Mask-Net and WCS-Net, the two convolutional layers are set with kernel size 3×3 and 1×1, respectively.

Training of EG-Net and Gaze2Mask-Net. We exploit the SALICON dataset [19] to train our EG-Net. The loss for EG-Net is formulated as the cross entropy loss between predicted eye gaze map and ground truth. For the Gaze2Mask-Net, we produce the eye gaze and object mask pair on the PASCAL VOC dataset [14]. Specifically, we utilize the EG-Net to generate the eye gaze map inside the ground truth instance mask. The Gaze2Mask-Net takes the RGB

image and eye gaze map as input to predict the object mask, which would be used to calculate the cross entropy loss with the corresponding ground truth. We use the Adam optimization with learning rate of 0.001 and batch size of 32 to train both EG-Net and Gaze2Mask-Net.

Training of WCS-Net. The overall loss of our WCS-Net is defined as:

$$L = \lambda_1 L_{Hotspot} + \lambda_2 L_{Mask} \tag{6}$$

where $L_{Hotspot}$ is the cross entropy loss between predicted hotspots with ground truth. L_{Mask} is the cross entropy loss between predicted masks and ground truth mask. For hotspot detection, we generate the ground truth based on SALICON data [19]. The WCS-Net is trained in two steps. We first pre-train the WCS-Net on the static image datasets of SALICON [19] and DUTS [43]. We use the human eye gaze annotations from SALICON to generate sufficient synthetic hotspot data for training our hotspot branch. Specifically, we use a large threshold to extract the peak distribution and take the focal regions as hotspot ground truth. We randomly combine the images from both datasets. The trade-off λ_1 is set to 1 if the data comes from SALICON, and 0 vise versa. The Adam with learning rate of 0.0001 and batch size of 15 is used in this stage to train the network. At the second stage, the WCS-Net is trained on the DAVIS-2016 training set [34] and SALICON datasets [19]. The network is trained using Adam with learning rate 0.0001 and batch size 10. Similarly, we integrate the data from two benchmarks and take the same setting on the trade-offs as the first stage.

4 Experiments

4.1 Dataset and Metrics

Dataset. To evaluate the performance of our proposed model, we conduct comparison experiments on three public VOS datasets, including DAVIS-2016 [34], SegTrackv2 [29] and Youtube-Objects [12]. The DAVIS-2016 dataset [34] contains 50 high-quality videos with 3455 densely annotated frames. All the videos in DAVIS-2016 are annotated with only one foreground object and they are splitted into 30 for training and 20 for testing. The SegTrackv2 dataset [29] is another challenging VOS dataset, which has 14 video sequences with densely annotated pixel-level ground truth. The videos in SegTrackv2 have large variations in resolution, object size and occlusion. The Youtube-Objects dataset is a large-size dataset, which contains 126 video sequences of 10 object categories. The ground truth in Youtube-Objects is sparsely annotated in every 10 frames.

Metrics. To compare the performance of our model with other state-of-the-arts in UVOS, we exploit two metrics, which are mean region similarity ($\mathcal{J}$ Mean) and mean contour accuracy ($\mathcal{F}$ Mean) [35]. To evaluate our hotspot tracking results, we exploit CC, SIM, KLD and NSS [1]. Besides, we also provide the run time of our method for efficiency evaluation. All the experiments are conducted on one NVIDIA 1080Ti GPU.

Table 1: Overall comparison with state-of-the-arts on the DAVIS-2016 validation dataset. The "✓" is used to indicate whether the method contains First Frame (FF), Online Finetuning (OF) or Post-processing (PP).

Method	FF	OF	PP	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
OSVOS [2]	✓	✓	✓	79.8	80.6
PLM [37]	✓	✓	✓	70.0	62.0
SegFlow [5]	✓	✓		74.8	74.5
RGMP [51]	✓			81.5	**82.0**
TrackPart [4]	✓	✓		77.9	76.0
OSMN [52]	✓			74.0	72.9
Siammask [44]	✓			71.7	67.8
FSEG [20]		✓		70.0	65.3
LMP [41]		✓	✓	70.0	65.9
LVO [42]		✓	✓	75.9	72.1
ARP [24]		✓		76.2	70.6
PDB [39]			✓	77.2	74.5
MOA [38]		✓	✓	77.2	78.2
AGS [48]			✓	79.7	77.4
COSNet [32]			✓	80.5	79.4
AGNN [45]			✓	80.7	79.1
Ours				**82.2**	**80.7**

Table 2: Overall comparison with state-of-the-arts on SegTrackv2 dataset.

Method	KEY [26]	FST [33]	NLC [11]	FSEG [20]	MSTP [18]	Ours
$\mathcal{J}$ Mean	57.3	52.7	67.2	61.4	70.1	**72.2**

4.2 Evaluation on Unsupervised Video Object Segmentation

DAVIS 2016. We compare our proposed model with state-of-the-art approaches on DAVIS-2016 dataset [34]. In Tab. 1, we list comparison results with methods from both SVOS and UVOS. Besides, we provide the indicator of some operations in the existing methods, including first frame annotation (FF), online finetuning (OF) and post-processing (PP). Compared with the existing UVOS methods, our model outperforms the second-best AGNN [45] by 1.8%, 2.0% on $\mathcal{J}$ Mean and $\mathcal{F}$-Mean, respectively. Note that we do not implement any post-processing in our model as other methods. We also propose the comparison results between our model and SVOS methods. We can observe that even without providing first frame's ground truth during testing, our model can perform favorably against most SVOS approaches (without online training).

SegTrackv2. We also illustrate the comparison results on SegTrackv2 dataset in Tab. 2. We report the $\mathcal{J}$ Mean performance as suggested by [29, 18]. As can

Table 3: Overall comparison with state-of-the-arts on Youtube-Objects dataset. We report the per-category $\mathcal{J}$ Mean and the average result. Since COSNet and AGNN use dense-CRF post-processing, we also report our method with the same post-processing, denoted as Ours*.

Method	PP	Airplane (6)	Bird (6)	Boat (15)	Car (7)	Cat (16)	Cow (20)	Dog (27)	Horse (14)	Motor (10)	Train (5)	Avg.
FST [33]		70.9	70.6	42.5	65.2	52.1	44.5	65.3	53.5	44.2	29.6	53.8
ARP [24]		73.6	56.1	57.8	33.9	30.5	41.8	36.8	44.3	48.9	39.2	46.2
PDB [39]	✓	78.0	80.0	58.9	76.5	63.0	64.1	70.1	67.6	58.3	35.2	65.4
FSEG [20]		81.7	63.8	**72.3**	74.9	68.4	68.0	69.4	60.4	62.7	62.2	68.4
AGS [48]	✓	**87.7**	76.7	72.2	78.6	**69.2**	64.6	73.3	64.4	62.1	48.2	69.7
COSNet [32]	✓	81.1	75.7	71.3	77.6	66.5	**69.8**	**76.8**	67.4	67.7	46.8	70.5
AGNN [45]	✓	81.1	75.9	70.7	78.1	67.9	69.7	77.4	67.3	68.3	47.8	70.8
Ours		81.8	81.1	67.7	79.2	64.7	65.8	73.4	68.6	**69.7**	49.2	70.5
Ours*	✓	81.8	**81.2**	67.6	**79.5**	65.8	66.2	73.4	**69.5**	69.3	**49.7**	**70.9**

Table 4: Run time comparison on the DAVIS-2016 dataset. "Ours" is the model implemented on EfficientNet [40] and "Ours†" is the model built on Resnet101 [15].

Method	Siammask [44]	OSMN [52]	RGMP [51]	AGS [48]	Ours	Ours†
Time (s)	0.02	0.14	0.13	0.60	0.03	0.04

be seen, our model significantly outperforms state-of-the-art methods and has an improvement of 3% on $\mathcal{J}$ Mean against MSTP [18].

Youtube-Objects. Tab. 3 lists the $\mathcal{J}$ Mean results of the state-of-the-arts on Youtube-Objects. Our model achieves the best results on the categories of bird, car, horse, motorbike and train. It is also comparable with recent UVOS methods AGS [48], COSNet [32] and AGNN [45] across all categories. For a fair comparison with methods using Dense-CRF [25], we evaluate our method with the same post-processing (named as "Ours*"), and it achieves the best performance on this dataset.

Qualitative results. We illustrate the visual results of our model in Fig. 4. For the video object segmentation, our model can keep tracking the target object with shape deformation and occlusion and produce accurate segmentation masks with well-defined details.

Run Time. We also report the comparison on run time to verify the efficiency of our model. The run time of our model and other state-of-the-arts on DAVIS-2016 are shown in Tab. 4. The results illustrate that our model can not only produce accurate segmentation and eye gaze results, but achieves real-time inference speed. The Siammask [44] performs much faster than our model. However, our method does not need any ground truth indicator as [44] and we

Fig. 4: Qualitative results of our proposed model on DAVIS-2016. For the two sequences, the top row is the hotspot tracking results and the bottom row lists the object segmentation. Zoom in to see the details.

significantly outperform Siammask in accuracy with 15% and 19% improvement on $\mathcal{J}$ Mean and $\mathcal{F}$ Mean, respectively.

4.3 Evaluation on Hotspot Tracking

Hotspot ground truth for testing. As illustrated in Sec.1, the existing video eye gaze data is noisy and unstable due to the saccade in eye movement. It can not meet the problem formulation of our hotspot tracking. To evaluate the effectiveness of our hotspot tracking, we provide the hotspot data on DAVIS-2016 validation dataset. To annotate the hotspots for the target object in DAVIS-2016, we exploit the mouse click to stimulate the human attention as [19]. The reason why we don't use the eye tracker is two folds. First, it is hard to constrain the eye gaze inside the object using eye tracker. Second, the delay between human eye gaze and fast object movement makes it difficult for users to keep tracking the salient region. Such cases would produce flicker annotations inside the object (See the third row in Fig. 1. Compared with the eye tracker, using the mouse click is more suitable for our task to produce a clean and consistent hotspot ground truth [19]. Specifically, we sample the video clips from DAVIS-2016 test set in every 10 frame for annotating. We randomly disrupt the frame order in each video sequence and ask five users to annotate. The users are required to first determine the salient part inside the object and provide consistent mouse click on that region along the video sequence.

Comparison results. In our experiment, we exploit the existing metrics in eye gaze to evaluate the stability of our hotspot tracking [48]. The comparison results with other video eye tracking models are shown in Tab. 5. We can observe that our model outperforms state-of-the-arts on all metrics. The qualitative results in Fig 4 illustrate that our model can produce clean and consistent hotspot tracking for the video objects.

Table 5: Quantitative results of hotspot tracking on the DAVIS-2016 dataset.

Method	KLD ↓	NSS ↑	CC↑	SIM↑
DeepVS [22]	3.148	2.010	0.189	0.089
AGS [48]	2.704	2.944	0.275	0.138
Ours	**2.592**	**3.174**	**0.429**	**0.333**
Ours w/o. mask branch	2.701	3.025	0.399	0.291

Table 6: Ablation study on the DAVIS-2016 validation dataset.

Model Setting		$\mathcal{J}$ Mean	$\Delta\mathcal{J}$ Mean
Full Model		82.2	-
Multi-level feature	45	74.5	-7.7
	345	78.0	-4.2
	2345	79.9	-2.3
w/o. weighted correlation		75.3	-6.9
w/o. weighted pooling		78.3	-3.9
w/o. hotspot branch		81.5	-0.7
Resnet101		81.7	-0.5

4.4 Ablation Study

In this section, we analyze the contribution of each component in our weighted correlation siamese network. The results in terms of $\mathcal{J}$ Mean on DAVIS-2016 dataset are shown in Tab. 6.

Effectiveness of multi-level construction. In our model, we implement the weighted correlation block on the multiple side-outputs to calculate the correspondence between template and search frames. The generated multi-level correlation features are further fed into two sub-branches to produce both hotspot and object mask. To demonstrate the effectiveness of such multi-level architecture, we gradually remove the skip connection from each feature level. The results can be referred in the rows named "Multi-level feature" in Tab 6. For example, the item named "45" indicates that only features from the 4-th and 5-th encoder blocks are used to produce the correlation feature and generate the final results. The results in Tab. 6 verify the efficacy of our multi-level architecture.

Effectiveness of joint training. Our model builds a unified architecture for joint video object segmentation and hotspot tracking. Specifically, with a shared siamese encoder, we implement two parallel decoder branches for both tasks. The decoder branches are jointly trained with both object mask and hotspot annotations using Eq. 6. To investigate the efficacy of joint training strategy on video object segmentation, we remove the hotspot tracking branch and train the network for only video object segmentation. Its comparison result between full

model ("w/o. hotspot tracking branch" in Tab. 6) demonstrate the effectiveness for such joint training method. We also implement the ablation experiment on the hotspot tracking branch. The result "Ours w/o. mask branch" in Tab. 5 verifies the joint training strategy on hotspot tracking task.

Effectiveness of weighted correlation. To verify the effect of weighted correlation block, we build a baseline network. Instead of conducting the correlation operation, we first concatenate the template and search feature in channel wise. Then the concatenated feature would be fed into the decoder block to produce the object mask. Note that the multi-level construction is also implemented in this baseline network. The result of this baseline network is illustrated in "w/o. weighted correlation" of Tab. 6. The comparison result with baseline verifies the efficacy of our weighted correlation block.

Effectiveness of weighted pooling. In our weighted correlation block, we transfer the template feature map into 1×1 feature vector via weighted pooling with the initial mask and hotspot. To demonstrate the effect of weighted pooling, we implement a model with the original correlation operation as Siammask [44]. Specifically, the size of template frame is half of the search frame, and they are fed into the siamese encoder to produce multi-level features. Instead of conduct the weighted pooling operation, we directly use the template and search features to calculate correlation features for mask generation. From the results in Tab. 6, we can observe that the weighted pooling is more effective compared with the original correlation operation in tracking methods [27, 28, 44].

Implementation using Resnet101. We implement our WCS-Net on the Resnet101 [15]. The results of $\mathcal{J}$ Mean and the run time are listed in Tab. 6 and Tab. 4, respectively. They demonstrate that our WCS-Net also works on the Resnet on both accuracy and efficiency.

5 Conclusion

In this paper, we propose a Weighted Correlation Siamese Network (WCS-Net) for joint unsupervised video object segmentation and hotspot tracking. We introduce a novel weighted correlation block (WCB) to calculate the cross-correlation between template frame and the search frame. The correlation feature from WCB is used in both sub-branches for generating the track-lets of mask and hotspots. The experimental results on three benchmarks demonstrate our proposed model outperforms existing competitors on both unsupervised video object segmentation and hotspot tracking with a significantly faster speed of 33 FPS.

Acknowledgements. The paper is supported in part by the National Key R&D Program of China under Grant No. 2018AAA0102001 and National Natural Science Foundation of China under grant No. 61725202, U1903215, 61829102, 91538201, 61771088, 61751212 and the Fundamental Research Funds for the Central Universities under Grant No. DUT19GJ201 and Dalian Innovation leaders support Plan under Grant No. 2018RD07.

References

1. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)
2. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017)
3. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: CVPR (2018)
4. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: CVPR (2018)
5. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: CVPR (2017)
6. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R^3Net: Recurrent residual refinement network for saliency detection. In: IJCAI (2018)
7. Ding, H., Cohen, S., Price, B., Jiang, X.: Phraseclick: Toward achieving flexible interactive segmentation by phrase and click. In: ECCV (2020)
8. Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: ICCV (2019)
9. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: CVPR (2018)
10. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Semantic correlation promoted shape-variant context for segmentation. In: CVPR (2019)
11. Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC (2014)
12. Ferrari, V., Schmid, C., Civera, J., Leistner, C., Prest, A.: Learning object class detectors from weakly annotated video. In: CVPR (2012)
13. Gegenfurtner, K.R.: The interaction between vision and eye movements. Perception **45**(12), 1333–1357 (2016)
14. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: CVPR (2011)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR (2020)
17. Hu, Y.T., Huang, J.B., Schwing, A.: Maskrnn: Instance level video object segmentation. In: Advances in Neural Information Processing Systems (2017)
18. Hu, Y.T., Huang, J.B., Schwing, A.G.: Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: ECCV (2018)
19. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: ICCV (2015)
20. Jain, S.D., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: CVPR (2017)
21. Jang, W.D., Lee, C., Kim, C.S.: Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In: CVPR (2016)
22. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: ECCV (2018)

23. Keuper, M., Andres, B., Brox, T.: Motion trajectory segmentation via minimum cost multicuts. In: CVPR (2015)
24. Koh, Y.J., Kim, C.S.: Primary object segmentation in videos based on region augmentation and reduction. In: CVPR (2017)
25. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems. pp. 109–117 (2011)
26. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011)
27. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR (2019)
28. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR (2018)
29. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013)
30. Li, S., Seybold, B., Vorobyov, A., Fathi, A., Huang, Q., Jay Kuo, C.C.: Instance embedding transfer to unsupervised video object segmentation. In: CVPR (2018)
31. Li, S., Seybold, B., Vorobyov, A., Lei, X., Jay Kuo, C.C.: Unsupervised video object segmentation with motion-based bilateral networks. In: ECCV (2018)
32. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019)
33. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
34. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
35. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
36. Rommelse, N.N., Van der Stigchel, S., Sergeant, J.A.: A review on eye movement studies in childhood and adolescent psychiatry. Brain and cognition **68**(3), 391–414 (2008)
37. Shin Yoon, J., Rameau, F., Kim, J., Lee, S., Shin, S., So Kweon, I.: Pixel-level matching for video object segmentation using convolutional neural networks. In: ICCV (2017)
38. Siam, M., Jiang, C., Lu, S., Petrich, L., Gamal, M., Elhoseiny, M., Jagersand, M.: Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In: 2019 International Conference on Robotics and Automation (2019)
39. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: ECCV (2018)
40. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
41. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: CVPR (2017)
42. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: CVPR (2017)
43. Wang, L., Lu, H., Wang, Y., Feng, M., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)

44. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR (2019)
45. Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L.: Zero-shot video object segmentation via attentive graph neural networks. In: CVPR (2019)
46. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: CVPR (2018)
47. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR (2015)
48. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C.H., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (2019)
49. Wei, Z., Wang, B., Hoai, M., Zhang, J., Shen, X., Lin, Z., Mech, R., Samaras, D.: Sequence-to-segments networks for detecting segments in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
50. Wei, Z., Wang, B., Nguyen, M.H., Zhang, J., Lin, Z., Shen, X., Mech, R., Samaras, D.: Sequence-to-segment networks for segment detection. In: Advances in Neural Information Processing Systems. pp. 3507–3516 (2018)
51. Wug Oh, S., Lee, J.Y., Sunkavalli, K., Joo Kim, S.: Fast video object segmentation by reference-guided mask propagation. In: CVPR (2018)
52. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: CVPR (2018)
53. Yang, Z., Wang, Q., Bertinetto, L., Hu, W., Bai, S., Torr, P.H.S.: Anchor diffusion for unsupervised video object segmentation. In: ICCV (2019)
54. Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., Hoai, M.: Predicting goal-directed human attention using inverse reinforcement learning. In: CVPR (2020)
55. Zhang, L., Dai, J., Lu, H., He, Y.: A bi-directional message passing model for salient object detection. In: CVPR (2018)
56. Zhang, L., Lin, Z., Zhang, J., Lu, H., He, Y.: Fast video object segmentation via dynamic targeting network. In: ICCV (2019)
57. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: CVPR (2019)