Joint Video Object Discovery and Segmentation by Coupled Dynamic Markov Networks

Ziyi Liu, *Student Member, IEEE* Le Wang, *Member, IEEE* Gang Hua, *Senior Member, IEEE*
Qilin Zhang, *Member, IEEE* Zhenxing Niu, *Member, IEEE* Ying Wu, *Fellow, IEEE*
and Nanning Zheng, *Fellow, IEEE*

***Abstract*—It is a challenging task to extract segmentation mask of a target from a single noisy video, which involves object discovery coupled with segmentation. To solve this challenge, we present a method to jointly discover and segment an object from a noisy video, where the target disappears intermittently throughout the video. Previous methods either only fulfill video object discovery, or video object segmentation presuming the existence of the object in each frame. We argue that jointly conducting the two tasks in a unified way will be beneficial. In other words, video object discovery and video object segmentation tasks can facilitate each other. To validate this hypothesis, we propose a principled probabilistic model, where two dynamic Markov networks are coupled – one for discovery and the other for segmentation. When conducting the Bayesian inference on this model using belief propagation, the bi-directional message passing reveals a clear collaboration between these two inference tasks. We validated our proposed method in five datasets. The first three video datasets, *i.e.*, the SegTrack dataset, the YouTube-Objects dataset, and the Davis dataset, are not noisy, where all video frames contain the objects. The two noisy datasets, *i.e.*, the XJTU-Stevens dataset, and the Noisy-ViDiSeg dataset, newly introduced in this paper, both have many frames that do not contain the objects. When compared with state-of-the-art, it is shown that although our method produces inferior results on video datasets without noisy frames, we are able to obtain better results on video datasets with noisy frames.**

***Index Terms*—Object segmentation, Object discovery, Dynamic Markov Networks, Probabilistic graphical model.**

I. Introduction

THE problem of separating out a foreground object from the background across all frames of a video is known as video object segmentation. The goal is to label each pixel in all video frames according to whether it belongs to the unknown target object or not. The resulting segmentation is a spatio-temporal object tube delineating the boundaries of the object throughout a video. Such capacity can be useful for a variety of computer vision tasks, such as object centric video summarization, action analysis, video surveillance, and content-based video retrieval.

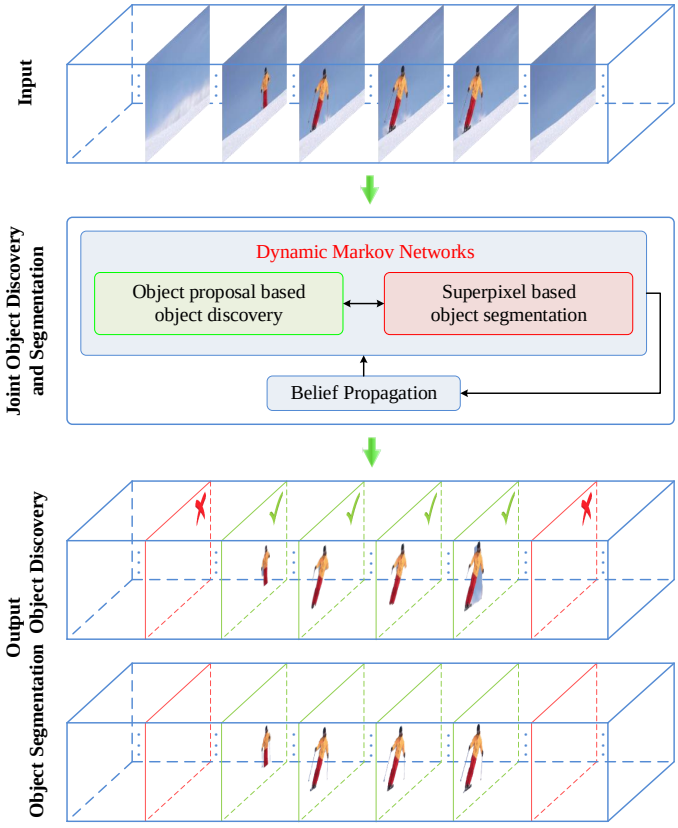


Fig. 1. Illustration of the proposed joint video object discovery and segmentation framework.

Video object segmentation has received great progress in recent years, mainly including fully automatic methods [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], semi-supervised methods [11], [12], [13], [14], [15], [16], and interactive methods [17], [18], [19], [20], [21]. Nevertheless, there are still three issues need to be further addressed.

Firstly, an unrealistically optimistic assumption is often

Manuscript received February 10, 2018; revised June 18, 2018; accepted July 16, 2018. Date of publication July 31, 2018; date of current version September 4, 2018. This work was supported partly by National Key R&D Program of China Grant 2017YFA0700800, National Natural Science Foundation of China Grants 61629301, 61773312, 91748208, and 61503296, China Postdoctoral Science Foundation Grants 2017T100752 and 2015M572563, National Science Foundation Grants IIS-1217302 and IIS-1619078, and the Army Research Office ARO W911NF-16-1-0138. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tolga Tasdizen. (*Corresponding author: Le Wang.*)

Z. Liu, L. Wang, and N. Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: liuziyi@stu.xjtu.edu.cn; lewang@xjtu.edu.cn; nnzheng@mail.xjtu.edu.cn).

G. Hua is with Microsoft Research, Redmond, WA 98052, USA (e-mail: ganghua@gmail.com).

Q. Zhang is with HERE Technologies, Chicago, IL 60606, USA (e-mail: samqzhang@gmail.com).

Z. Niu is with Alibaba Group, Hangzhou, Zhejiang 311121, China (e-mail: zhenxing.nzx@alibaba-inc.com).

Y. Wu is with Northwestern University, Evanston, IL 60208, USA (e-mail: yingwu@eecs.northwestern.edu).

made in these methods, that the target object is present in all (or most) video frames. Therefore, methods robust to a large number of "noisy" frames (*i.e.*, irrelevant frames devoid of the target object) are urgently needed.

Moreover, most of them emphasized on leveraging the low-level features (*i.e.*, color and motion) or contextual information shared among individual or consecutive frames to find the common regions, and simply employed the short-term motion (*e.g.*, optical flow) between consecutive frames to smooth the spatio-temporal segmentation. Therefore, they often encountered difficulties when the objects exhibit large variations in appearance, motion, size, pose, and viewpoint.

Furthermore, several methods [4], [22], [23], [24], [25], [26] employed the mid-level representation of objects (*i.e.*, object proposals [27]) as an additional cue to facilitate the segmentation of the object, with object discovery and object segmentation conveniently isolated as two independent tasks and performed in a two-step manner [28], [29]. Unfortunately, the disregard of their dependencies often leads to suboptimal performances, *e.g.*, object segmentation dramatically failing at focusing on the target, object discovery providing wildly inaccurate object proposals.

To address the above three issues, we present a method to jointly discover and segment an object from a single video with many noisy frames, benefiting from the collaboration of object discovery and object segmentation. Fig. 1 illustrates the proposed framework. We propose a principled probabilistic model, where one dynamic Markov Network for video object discovery and one dynamic Markov Network for video object segmentation are coupled. When conducting the Bayesian inference on this model using belief propagation, the bi-directional propagation of the beliefs of the object's posteriors on an object proposal graph and a superpixel graph reveals a clear collaboration between these two inference tasks. More specifically, object discovery is conducted through the object proposal graph representing the correlations of object proposals among multiple frames, which is built under the help of the spatio-temporal object segmentation tube obtained by object segmentation on the superpixel graph. Object segmentation is achieved on the superpixel graph representing the connections of superpixels, which is benefited from the spatio-temporal object proposal tube generated by object discovery through the object proposal graph.

We validated our proposed method in five video datasets, including 1) object segmentation from a single video without noisy frames on three video datasets where all video frames contain the objects, *i.e.*, the SegTrack dataset [30], [31], the YouTube-Objects dataset [32], and the Davis dataset [33], and 2) joint object discovery and segmentation from a single video with noisy frames on two video datasets where the videos in both datasets have many frames not containing the objects, *i.e.*, the XJTU-Stevens dataset [34], [35], and the Noisy-ViDiSeg dataset, newly introduced in this paper. When compared with state-of-the-art, it is shown that although our method produces inferior results on video datasets without noisy frames, we are able to obtain better results on video datasets with noisy frames. Indeed, the more noisy frames the videos contain, the better our method performs when compared with competing methods.

The key contributions of this paper are:

- We present an unsupervised method to jointly discover and segment an object from a single noisy video, where the target object disappears intermittently throughout the video.
- We propose a principled probabilistic model, where two dynamic Markov networks are coupled – one for discovery and the other for segmentation.
- To accurately evaluate our proposed method, we establish a noisy video object discovery and segmentation dataset, named Noisy-ViDiSeg dataset, in which the overall percentage of noisy frames is up to 33.1%.

The paper is organized as follows. Section II discusses the related work. Then, we present the principled probabilistic model for joint object discovery and segmentation in Section III, the inference algorithm in Section IV, and the implementation details in Section V. Experimental results are provided in Section VI. Finally, we conclude the paper in Section VII.

II. Related Work

We review related work in video object segmentation, mainly including unsupervised and supervised methods. Since our proposed method leverages the object proposals, we also review the object proposal based video object segmentation methods. Moreover, as some video object co-segmentation methods can separate a common object from multiple noisy videos, we briefly introduce them.

A. Unsupervised Video Object Segmentation

Unsupervised video object segmentation methods aim at automatically extracting an object from a single video. These methods exploited features such as clustering of point trajectories [1], [2], motion characteristics [3], appearance [4], [5], or saliency [3], [6], [7] to achieve object segmentation. Recently, Jang *et al.* [8] separated a primary object from its background in a video based on an alternating convex optimization scheme. Jain *et al.* [9] proposed an end-to-end learning framework to combine motion and appearance information to produce a pixel-wise binary segmentation for each frame. Luo *et al.* [10] proposed a complexity awareness framework which exploits local clips and their relationships.

B. Supervised Video Object Segmentation

Supervised video object segmentation methods require user annotations about a primary object, and can be roughly categorized into label propagation based methods and interactive segmentation methods.

In label propagation based segmentation, an object is manually delineated in one or more frames, and then propagated to the remaining ones [11], [13], [14], [15], [16]. Badrinarayanan *et al.* [11] proposed a probabilistic graphical model for label propogation. Xiang *et al.* [12] proposed an online web-data-driven framework for moving object segmentation with online prior learning and 3D Graph cuts. Jain and Grauman [13]

proposed a foreground propagation method using higher order supervoxel potentials. Tsai *et al.* [14] considered video object segmentation and optical flow estimation simultaneously, where the combination improved both. Marki *et al.* [15] utilized the segmentation mask of the first frame to construct appearance models for the objects, and then inferred the segmentation by optimizing an energy on a regularly sampled bilateral grid. Caelles *et al.* [16] adopted Fully Convolutional Networks (FCNs) to tackle video object segmentation, given the mask of the first frame.

In interactive segmentation, user annotations on a few frames are iteratively added during the object segmentation procedure [17], [18], [19], [20], [21]. Although they can guarantee a high quality segmentation, the needs of tedious human efforts render them unable to handle a large number of videos. Thus, they are only suitable for specific applications, such as video editing and post-processing.

C. Object Proposal Based Video Object Segmentation

A large number of methods [4], [22], [23], [24], [25], [26] leveraged the notion of "what is an object" (*i.e.*, object proposals [36], [27]) to facilitate video object segmentation. Lee *et al.* [4] automatically discovered key segments and grouped them to predict the foreground object in a video. Ma and Latecky [22] cast video object segmentation as finding a maximum weighted clique in a locally connected region graph with mutex constraints.

Zhang *et al.* [23] segmented the primary video object through a layered directed acyclic graph, which combined unary edges measuring the objectness of the object proposal and pairwise edges modeling the affinities between them. Fragkiadaki *et al.* [24] segmented the moving objects by ranking spatio-temporal segment proposals according to a moving objectness. Perazzi *et al.* [25] employed a fully connected spatio-temporal graph built over object proposals for video segmentation. Koh and Kim [26] identified the primary object region from the object proposals per frame by an augmentation and reduction process, and then achieved object segmentation.

D. Video Object Co-segmentation

There are several methods focusing on video object co-segmentation from multiple videos [37], [38], [39], [40], [34], [35], [41], where the numbers of both the object classes and object instances are unknown in each frame and each video. Chiu and Fritz [37] proposed a non-parametric algorithm to cluster pixels into different regions. Fu *et al.* [38] presented a selection graph to formulate correspondences between different videos. Lou and Gevers [39] employed the appearance, saliency and motion consistency of object proposals together to extract the primary objects.

Zhang *et al.* [40] proposed an object co-segmentation method by selecting spatially salient and temporally consistent object proposal tracklets. Wang *et al.* [34], [35] proposed a spatio-temporal energy minimization formulation for video object discovery and co-segmentation from multiple videos, but the method needed to be bootstrapped with a few frame-level labels. However, they almost always encountered difficulties when the videos have a large number of noisy frames.

The differences between our method and the above methods are two-fold. One is that we address the problem of simultaneously discovering and segmenting the object of interest from a single video with a large number of noisy frames. The other one is that we cast the two tasks of video object discovery and video object segmentation into a principled probabilistic model by coupling two dynamic Markov networks, in which object discovery and object segmentation can benefit each other. The proposed method is the first one that can jointly discover and segment the object from a single noisy video with a principled probabilistic model.

III. Model

Given a video $\mathbf{V} = \{f_t\}_{t=1}^T$ with a significant number of noisy frames, our goal is to jointly find an object discovery labeling $\mathbf{L}$ and an object segmentation labeling $\mathbf{B}$ from $\mathbf{V}$. $\mathbf{L} = \{\mathbf{L}_t\}_{t=1}^T$ is a spatio-temporal region (object) proposal tube of $\mathbf{V}$. $\mathbf{L}_t = \{l_{t,i}\}_{i=1}^K$ is the object discovery label of each frame f_t, where $l_{t,i} \in \{0,1\}$ and $\sum_{i=1}^K l_{t,i} \leq 1$, *i.e.*, no more than one region proposal among all the K proposals in f_t will be identified as the object. $\mathbf{B} = \{\mathbf{B}_t\}_{t=1}^T$ is a spatio-temporal object segmentation tube of $\mathbf{V}$. $\mathbf{B}_t = \{b_{t,j}\}_{j=1}^J$ is the object segmentation label of f_t, where $b_{t,j} \in \{0,1\}$ denotes that each of the J superpixels either belongs to the object ($b_{t,j} = 1$) or the background ($b_{t,j} = 0$).

The image observations associated with $\mathbf{L}$, $\mathbf{L}_t$, $\mathbf{B}$, and $\mathbf{B}_t$ are denoted by $\mathbf{O} = \{\mathbf{O}_t\}_{t=1}^T$, $\mathbf{O}_t = \{o_{t,i}\}_{i=1}^K$, $\mathbf{S} = \{\mathbf{S}_t\}_{t=1}^T$ and $\mathbf{S}_t = \{s_{t,j}\}_{j=1}^J$, respectively. $o_{t,i}$ and $s_{t,j}$ are the representations of a region proposal (*e.g.*, generated by [27]) and a superpixel (*e.g.*, computed by SLIC [42]), respectively.

Specifically, the beneficial information are encouraged to be propagated between the joint inference of $\mathbf{L}$ and $\mathbf{B}$, and hence video object discovery and video object segmentation can naturally benefit each other. As illustrated in Fig. 2 (a), we employ a Markov network [43], [44], [45] to characterize the joint object discovery and segmentation from $\mathbf{V}$. The undirected link represents the mutual influence of object discovery and object segmentation, and is associated with a potential compatibility function $\Psi(\mathbf{L}, \mathbf{B})$. The directed links represent the image observation processes, and are associated with two image likelihood functions $p(\mathbf{O}|\mathbf{L})$ and $p(\mathbf{S}|\mathbf{B})$. According to the Bayesian rule, it is easy to obtain

$$p(\mathbf{L}, \mathbf{B}|\mathbf{O}, \mathbf{S}) = \frac{1}{Z_Q}\Psi(\mathbf{L}, \mathbf{B})p(\mathbf{O}|\mathbf{L})p(\mathbf{S}|\mathbf{B}), \tag{1}$$

where Z_Q is a normalization constant. The above Markov network is a generative model at one time instant.

When putting the above Markov network into temporal context by accommodating dynamic models, we construct two coupled dynamic Markov networks as shown in Fig. 2 (b). The subscript t represents the time index. In addition, we denote the collective image observations associated with the object discovery labels from the beginning to t by $\underrightarrow{\mathbf{O}}_t = \{\mathbf{O}_1, \ldots, \mathbf{O}_t\}$, and reversely from the end to t by $\underleftarrow{\mathbf{O}}_t = \{\mathbf{O}_T, \ldots, \mathbf{O}_t\}$. The collective image observations associated with the object segmentation labels are built in the same way, *i.e.*, $\underrightarrow{\mathbf{S}}_t =$

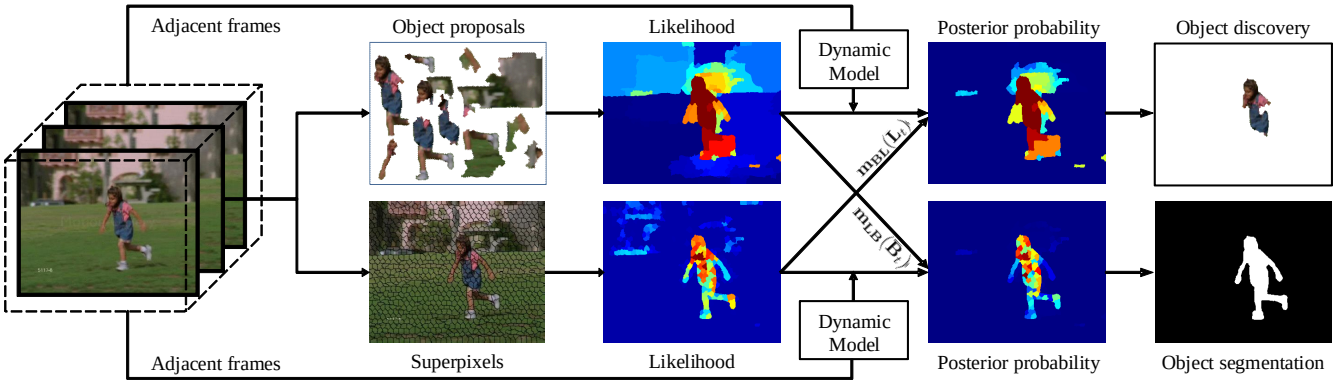


Fig. 3. The inference process of the two coupled dynamic Markov networks to obtain the joint video object discovery and segmentation.

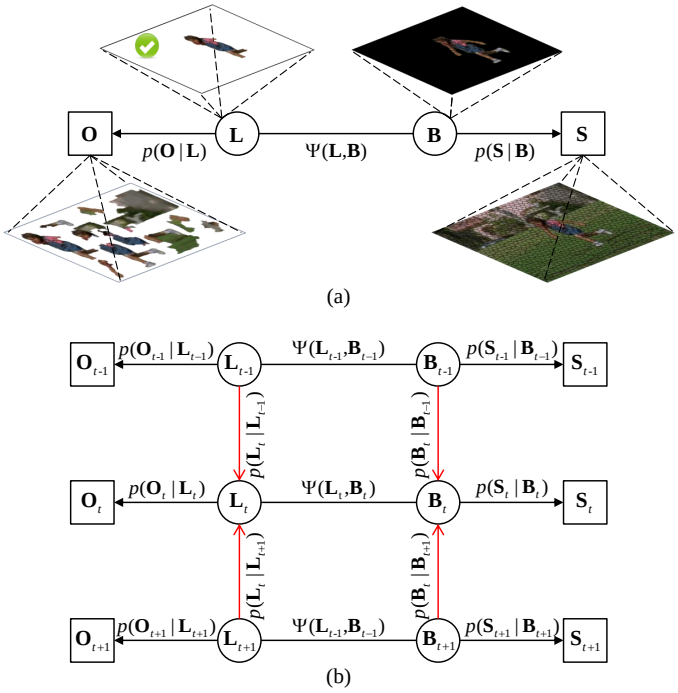


Fig. 2. The (a) Markov network and (b) the two coupled dynamic Markov networks for joint video object discovery and segmentation.

$\{\mathbf{S}_1, \ldots, \mathbf{S}_t\}$ and $\underleftarrow{\mathbf{S}}_t = \{\mathbf{S}_T, \ldots, \mathbf{S}_t\}$. In this formulation, the problem of joint video object discovery and segmentation from a single noisy video is to perform Bayesian inference of the dynamic Markov networks to obtain the marginal posterior probabilities $p(\mathbf{L}_t|\mathbf{O}, \mathbf{S})$ and $p(\mathbf{B}_t|\mathbf{O}, \mathbf{S})$.

IV. Inference

We first perform Bayesian inference of the Markov network in Fig. 2 (a) to obtain the marginal posterior probabilities $p(\mathbf{L}|\mathbf{O}, \mathbf{S})$ and $p(\mathbf{B}|\mathbf{O}, \mathbf{S})$. With loop-less graph models in Bayesian inference, belief propagation guarantees the exact inference through a local message passing process [46], [47]. As is the case in Fig. 2 (a), Bayesian inference is performed using belief propagation. For ease of reading, the detailed derivation of the formula for the inference is summarized in Appendix I. They are calculated by iterating the message passing until convergence as

$$p(\mathbf{L}|\mathbf{O}, \mathbf{S}) \propto p(\mathbf{O}|\mathbf{L})\mathbf{m}_{\mathbf{BL}}(\mathbf{L}), \tag{2}$$

$$p(\mathbf{B}|\mathbf{O}, \mathbf{S}) \propto p(\mathbf{S}|\mathbf{B})\mathbf{m}_{\mathbf{LB}}(\mathbf{B}), \tag{3}$$

where $\mathbf{m}_{\mathbf{BL}}(\mathbf{L})$ and $\mathbf{m}_{\mathbf{LB}}(\mathbf{B})$ are the local messages passing from $\mathbf{B}$ to $\mathbf{L}$ and from $\mathbf{L}$ to $\mathbf{B}$, respectively.

Then, we generalize to infer the marginal posterior probabilities $p(\mathbf{L}_t|\mathbf{O}, \mathbf{S})$ and $p(\mathbf{B}_t|\mathbf{O}, \mathbf{S})$ on the two coupled dynamic Markov networks in Fig. 2 (b), as detailed in Appendix II. They are computed by combing the incoming messages from both its forward and backward neighborhood as

$$\begin{aligned} p(\mathbf{L}_t|\mathbf{O}, \mathbf{S}) &= p(\mathbf{O}_t|\mathbf{L}_t)\mathbf{m}_{\mathbf{BL}}(\mathbf{L}_t) \\ &\times \int_{\mathbf{L}_{t-1}} p(\mathbf{L}_t|\mathbf{L}_{t-1}) p(\mathbf{L}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1}, \underrightarrow{\mathbf{S}}_{t-1}) d\mathbf{L}_{t-1} \\ &\times \int_{\mathbf{L}_{t+1}} p(\mathbf{L}_t|\mathbf{L}_{t+1}) p(\mathbf{L}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1}, \underleftarrow{\mathbf{S}}_{t+1}) d\mathbf{L}_{t+1}, \end{aligned} \tag{4}$$

$$\begin{aligned} p(\mathbf{B}_t|\mathbf{O}, \mathbf{S}) &= p(\mathbf{S}_t|\mathbf{B}_t)\mathbf{m}_{\mathbf{LB}}(\mathbf{B}_t) \\ &\times \int_{\mathbf{B}_{t-1}} p(\mathbf{B}_t|\mathbf{B}_{t-1}) p(\mathbf{B}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1}, \underrightarrow{\mathbf{S}}_{t-1}) d\mathbf{B}_{t-1} \\ &\times \int_{\mathbf{B}_{t+1}} p(\mathbf{B}_t|\mathbf{B}_{t+1}) p(\mathbf{B}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1}, \underleftarrow{\mathbf{S}}_{t+1}) d\mathbf{B}_{t+1}, \end{aligned} \tag{5}$$

where $\mathbf{m}_{\mathbf{BL}}(\mathbf{L}_t)$ and $\mathbf{m}_{\mathbf{LB}}(\mathbf{B}_t)$ are messages updating at time t from $\mathbf{B}_t$ to $\mathbf{L}_t$ and from $\mathbf{L}_t$ to $\mathbf{B}_t$ in both directions. $p(\mathbf{L}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1}, \underrightarrow{\mathbf{S}}_{t-1})$ and $p(\mathbf{B}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1}, \underrightarrow{\mathbf{S}}_{t-1})$ are the inference results at the previous time step $t-1$, and $p(\mathbf{L}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1}, \underleftarrow{\mathbf{S}}_{t+1})$ and $p(\mathbf{B}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1}, \underleftarrow{\mathbf{S}}_{t+1})$ are the inference results at the next time step $t+1$. Fig. 3 illustrates the inference process of the two coupled dynamic Markov networks to obtain the joint video object discovery and segmentation.

V. Implementation Details

In this section, we further present the detailed definitions of the likelihood functions, the compatibility functions, and the dynamic models of object discovery and object segmentation.

A. Likelihood Functions

Likelihood function of object discovery. As illustrated in Fig. 4, the object proposals generated for each frame (*e.g.*,

(a) (b) (c)

Fig. 4. Illustration of three types of object proposals: (a) object region, (b) possible object region, and (c) non-object region.

by [27]) have three forms: (1) *object region*, which is part of (or exactly) the object; (2) *possible object region*, which simultaneously contains parts of the object and the background; and (3) *non-object region*, which is part of (or exactly) the background.

It is ideal to select the "*object region*" that almost exactly contains the object instead of the "*possible object region*" and "*non-object region*". Then the question becomes: *how to measure the confidence of a region being an object?* We identified three useful measures: (1) **saliency**, which indicates that a region being most salient is more likely to be an object; (2) **objectness**, which requires the appearance of a region to be typical to a whole object; and (3) **motility**, which requires a region to have distinct motion patterns relative to its surrounding.

Thus, we define an object score by combining the above three measures to estimate how likely an object proposal $o_{t,i}$ is to be a whole object as

$$r(o_{t,i}) = r_s(o_{t,i}) \cdot r_a(o_{t,i}) \cdot r_m(o_{t,i}), \tag{6}$$

where $r_s(o_{t,i})$ is a saliency score, which is the mean value of the saliency values (*e.g.*, computed by [48]) within $o_{t,i}$; $r_a(o_{t,i})$ is an objectness score denoting the confidence that $o_{t,i}$ contains an object, which is computed by scoring the edge map described in [49]; and $r_m(o_{t,i})$ is a motion score, measuring the confidence that $o_{t,i}$ is a coherently moving object. It is computed similarly to $r_a(o_{t,i})$, but replacing the edge map with the motion boundary map [50].

Then, the likelihood function $p(\mathbf{O}_t|\mathbf{L}_t)$ of object discovery is calculated as

$$p(\mathbf{O}_t = o_{t,i}|\mathbf{L}_t) = \widehat{r}(o_{t,i}); i \in \{1, \cdots, K\}, \tag{7}$$

where $\widehat{r}(o_{t,i})$ is the object score normalized across $\mathbf{V}$, and K is the number of proposals that $\mathbf{O}_t$ contains.

Likelihood function of object segmentation. The object proposals in the spatio-temporal object proposal tube of $\mathbf{V}$ are treated as foreground objects, and the remaining parts are naturally treated as background. We learn two color Gaussian Mixture Models (GMMs) for the object and the background across $\mathbf{V}$, and denote them as $\mathbf{h}_1$ and $\mathbf{h}_0$, respectively. The likelihood function of object segmentation is then defined as

$$p(\mathbf{S}_t = s_{t,j}|\mathbf{B}_t) = \mathbf{h}_{b_{t,j}}(s_{t,j}); j \in \{1, \cdots, J\}, \tag{8}$$

where J is the number of superpixels that $\mathbf{S}_t$ contains.

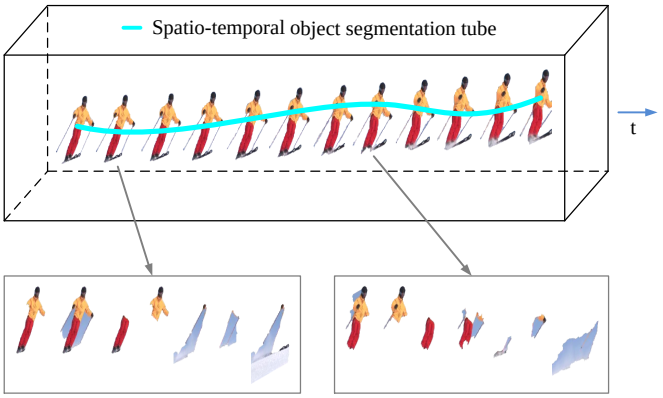


Fig. 5. The object proposals ranked by the compatibility function based on the spatio-temporal object segmentation tube obtained by object segmentation.

B. Compatibility Functions

The object proposal selected by object discovery should have a large overlap with the foreground object obtained by object segmentation. Thus, the compatibility function $\Psi_{\mathbf{LB}}(\mathbf{L}_t, \mathbf{B}_t)$ (from $\mathbf{L}_t$ to $\mathbf{B}_t$) is defined as

$$\Psi_{\mathbf{LB}}(\mathbf{L}_t, \mathbf{B}_t) = \mathrm{IoU}(o_{t,i}, \mathbf{B}_t(1)); i \in \{1, \cdots, K\}, \tag{9}$$

which means the intersection-over-union score (IoU) of $o_{t,i}$ and the segmented foreground $\mathbf{B}_t(1)$ of frame f_t, calculated by Eq. (16). The object proposals ranked by the compatibility function are illustrated in Fig. 5.

The compatibility function $\Psi_{\mathbf{BL}}(\mathbf{B}_t, \mathbf{L}_t)$ (from $\mathbf{B}_t$ to $\mathbf{L}_t$) is defined as

$$\Psi_{\mathbf{BL}}(\mathbf{B}_t, \mathbf{L}_t) = \frac{|s_{t,j} \cap \mathbf{O}_t(1)|}{|s_{t,j}|}; j \in \{1, \cdots, J\}, \tag{10}$$

which is the rate that superpixel $s_{t,j}$ covered by the selected object proposal $\mathbf{O}_t(1)$.

C. Dynamic Models

Dynamic model of object discovery. The object discovery labeling $\mathbf{L}$ should be temporally consistent throughout $\mathbf{V}$. Thus, the dynamic model of object discovery is defined as

$$p(\mathbf{L}_t = l_{t,m}|\mathbf{L}_{t-1}) = p^o_m; m \in \{1, \cdots, K\}, \tag{11}$$

where

$$p^o_m = \delta_m \cdot (\exp(-\alpha_m) + \exp(-\beta_m)), \tag{12}$$

is the transition probability between $o_{t,m}$ and its temporally adjacent object proposal $o_{t-1,i}$, where i is found by

$$i = \underset{i' \in \{1, \cdots, K\}}{\arg\max}\ \mathrm{IoU}(o_{t,m}, \mathrm{Warp}(o_{t-1,i'})), \tag{13}$$

where $\mathrm{Warp}(o_{t-1,i'})$ is the warped region from $o_{t-1,i'}$ in frame f_{t-1} to its neighboring frame f_t by optical flow [51]. $\delta_m = \delta(l_{t-1,i}, l_{t,m})$ is an indicator variable. It is 1 when $l_{t-1,i} \neq l_{t,m}$, *i.e.*, the object discovery labels of $o_{t-1,i}$ and $o_{t,m}$ are inconsistent, and 0 otherwise. $\alpha_m = \mathrm{EMD}(\mathbf{h}_c(o_{t-1,i}), \mathbf{h}_c(o_{t,m}))$ is the earth mover's distance (EMD) [52] between the color histograms of $o_{t-1,i}$ and

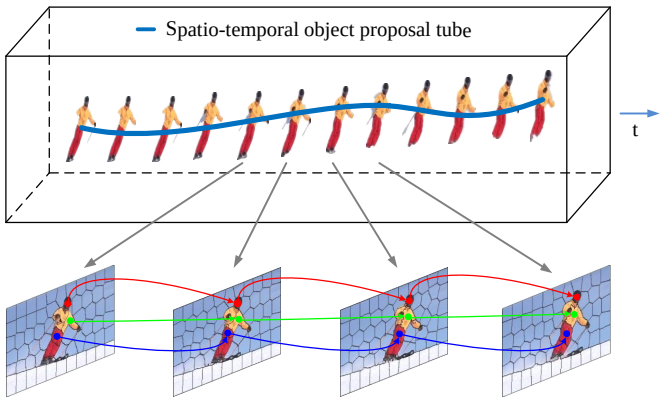


Fig. 6. The temporally adjacent superpixels found under the guidance of the spatio-temporal object proposal tube by object discovery.

$o_{t,m}$. $\beta_m = \chi^2_{shape}(o_{t-1,i}, o_{t,m})$ is the χ^2-distance between the HOG descriptors [53] of $o_{t-1,i}$ and $o_{t,m}$.

Dynamic model of object segmentation. The object segmentation labeling $\mathbf{B}$ should also be temporally consistent throughout $\mathbf{V}$. Thus, the dynamic model of object segmentation is defined as

$$p(\mathbf{B}_t = b_{t,n}|\mathbf{B}_{t-1}) = p^b_n; n \in \{1, \cdots, J\}, \tag{14}$$

where

$$p^b_n = \delta_n \exp(-\omega_n) + \sigma_n \exp(\mu_n), \tag{15}$$

is the transition probability between $s_{t,n}$ and its temporally adjacent superpixel $s_{t-1,j}$. $\delta_n = \delta(b_{t-1,j}, b_{t,n})$ is an indicator variable, defined identical to δ_m in Eq. (12). $\omega_n = ||\mathbf{h}_m(s_{t-1,j}) - \mathbf{h}_m(s_{t,n})||_2$ is the Euclidean distance between the histograms of oriented optical flow (HOOF) [54] of $s_{t-1,j}$ and $s_{t,n}$. σ_n is also an indicator variable, which is 1 when $s_{t,n}$ and $s_{t-1,j}$ both belong to the spatio-temporal object proposal tube obtained by object discovery, and 0 otherwise. $\mu_n = \text{IoU}(s_{t,n}, \text{Warp}(s_{t-1,j}))$ is the IoU score of $s_{t,n}$ and the warped region from $s_{t-1,j}$ to its neighboring frame f_t.

In this way, p^b_n will encourage the temporally adjacent superpixels that both belong to the spatio-temporal object proposal tube obtained by object discovery to have the same segmentation labels, as illustrated in Fig. 6. Besides, p^b_n will encourage the segmentation labels of temporally adjacent superpixels that have similar motion to be consistent. This ensures that we can handle the object with large motion.

D. Unsupervised Initialization

Given $\mathbf{V}$, each frame f_t is represented by $\mathcal{F}(f_t) \in \mathbb{R}^n$, which is obtained by using a pre-trained ResNet-152 [55] on ImageNet [56] followed by PCA [57] to generate a compact representation. We then leverage a classifier to obtain a confidence score for each frame to be a noisy frame. To train the classifier, we build an initial training set, in which the negative examples are gathered from the Google-30 dataset [58], [59], and the positive examples are uniformly sampled from $\mathbf{V}$ in some (*e.g.*, 5) frames. We proceed to retrain the classifier by treating the top ranked frames as positive, and the low ranked frames as negative. This process will iterate upon convergence. Specifically, benefitted from the iterative training, the impact of noisy frames in the positive examples on training accuracy is very limited.

TABLE I
THE STATISTICAL DETAILS OF FIVE BENCHMARK DATASETS OR THEIR SUBSETS USED FOR EVALUATION OF OUR JOINT VIDEO OBJECT DISCOVERY AND SEGMENTATION METHOD.

Dataset	Group	Video	Frame			Noise
			Total	Pos.	Neg.	(%)
SegTrack	8	8	785	785	0	0
YouTube-Objects	8	83	12941	12890	51	0.4
DAVIS	50	50	3455	3455	0	0
XJTU-Stevens	10	101	13398	12907	491	3.7
Noisy-ViDiSeg	11	11	1961	1312	649	33.1

VI. Experiments and Discussions

A. Experimental Setting

Evaluation datasets. We conduct extensive experiments on five video datasets to evaluate our joint video object discovery and segmentation method. We first evaluate the object segmentation performance from a single video without noisy frames on the SegTrack dataset [30], [31], the YouTube-Objects dataset [32], and the DAVIS dataset [33], where all video frames contain the objects. We proceed to evaluate the joint object discovery and segmentation performance from a single video with noisy frames on the XJTU-Stevens dataset [34], [35] and a newly introduced Noisy-ViDiSeg dataset in this paper, both have many frames that do not contain the objects. Some of the statistics of the above datasets (or their subsets) used for evaluation are summarized in Table I. They are

- **SegTrack dataset** [30], [31] is one of the most widely used video object segmentation dataset. It contains 14 videos of $1,066$ frames with pixel-wise annotations. As our method focuses on single object segmentation, we use the 8 videos that contain only one object.
- **YouTube-Objects dataset** [32], [13], [60] is mainly used for video object detection evaluation, while its subset indicated in [60] and the ground truth provided by [13] are often used for video object segmentation evaluation. This subset has 126 challenging videos of 10 categories with $20,101$ frames, where $2,127$ frames are labeled. As there are videos containing multiple objects, we only use the 83 videos of 8 categories containing only one object, with $12,941$ frames in total and $1,379$ labeled frames.
- **DAVIS dataset** [33] is the latest and most challenging video object segmentation dataset. It includes 50 high-quality videos of $3,455$ frames, and has pixel-wise labels for the prominent moving objects. The videos are unconstrained in nature and exhibit occlusions, motion blur, and large variation in appearance.
- **XJTU-Stevens dataset** [34], [35] is a video object co-segmentation and classification dataset. It contains 10 categories of 101 publicly available web videos for a total of $13,398$ frames, and 3.7% of them are noisy frames not containing the objects. The objects in each

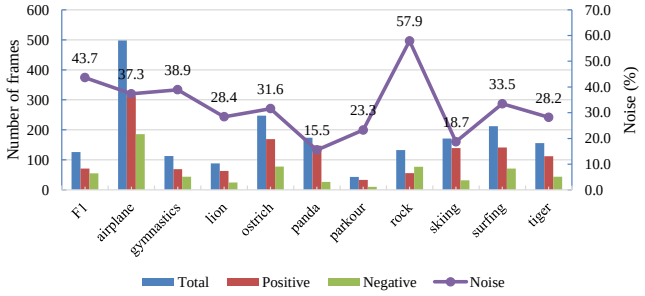


Fig. 7. The numbers of total, positive and negative frames, and the percentage of noisy frames of each category of the new Noisy-ViCoSeg dataset.

video category exhibit large differences in appearance, size, shape, viewpoint, and pose.

- **Noisy-ViDiSeg dataset** is a video object discovery and segmentation dataset newly introduced in this paper, in order to accurately evaluate our proposed method and to build a benchmark for future research. It includes 11 videos of 11 categories with $1,961$ frames in total, and each video contains a large number of noisy frames. The percentage of noisy frames is 33.1%. Fig. 7 details the statistics. As shown in Fig. 8, we manually assign the noisy frames with frame-level labels indicating if they contain the object, and the positive frames with both frame-level labels and pixel-wise segmentation labels.

Evaluation metric. The intersection-over-union score is used for object segmentation evaluation, and is defined as

$$\text{IoU} = \frac{|Seg \cap GT|}{|Seg \cup GT|}, \tag{16}$$

where Seg is the segmentation result, and GT is the ground truth segmentation.

The labeling accuracy is employed for object discovery evaluation, and is defined as

$$\text{Acc} = \frac{TP + TN}{Total}, \tag{17}$$

where TP, TN and $Total$ are the numbers of true positive, true negative and total frames, respectively.

Baselines. To fully evaluate our proposed method, we compare our method with six state-of-the-art methods, including four single video object segmentation methods (VOS [4], FOS [3], BVS [15], and OSS [16]) and two multi-video object co-segmentation methods (VOC [40] and VDC [35]). They are

- VOS [4]: an unsupervised single video object segmentation method which automatically discovers key segments and groups them to predict the foreground object.
- FOS [3]: an unsupervised single video object segmentation method which separates the target object via a rapid estimate of which pixels are inside the object.
- BVS [15]: a semi-supervised single video object segmentation method which separates the target objects based on operations in the bilateral space. It exploits the object segmentation mask of the first frame.
- OSS [16]: a semi-supervised single video object segmentation method which separates the object from the background based on a fully-convolutional neural network, given the mask of the first frame.
- VOC [40]: an unsupervised multi-video object co-segmentation method which can segment multiple objects by sampling, tracking and matching object proposals via a regulated maximum weight clique extraction scheme.
- VDC [35]: a supervised multi-video object discovery and co-segmentation method which can discover and segment the common objects from multiple videos with a few noisy frames, given the frame-level discovery labels of three video frames.

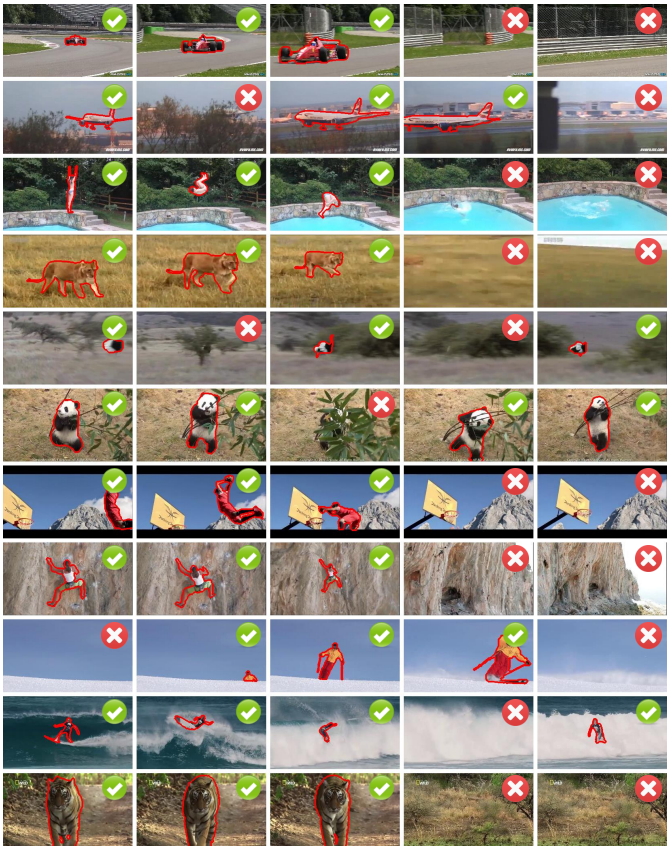


Fig. 8. Some example frames and their annotations of the Noisy-ViDiSeg dataset. The red cross indicates the noisy frame; the green tick indicates the positive frame containing the object, which is depicted by the red edge.

B. Object Segmentation from a Single Video without Noisy Frames

We first evaluate the object segmentation performance from a single video without noisy frames of our method on the SegTrack dataset [30], [31], YouTube-Objects dataset [32], and DAVIS dataset [33]. All video frames of these three video datasets contain the objects.

Evaluation on the SegTrack dataset. As our method focuses on single object segmentation, we test our method on the eight videos that contain only one object, and compare with four single video object segmentation methods (VOS [4], FOS [3], BVS [15], and OSS [16]). The average IoU scores and some example results of them are presented in Table II and

TABLE II
THE AVERAGE IOU SCORES OF OUR METHOD AND FOUR SINGLE VIDEO OBJECT SEGMENTATION METHODS ON EIGHT VIDEOS THAT CONTAIN ONLY ONE OBJECT OF THE SEGTRACK DATASET. HIGHER VALUES ARE BETTER.

Video	VOS [4]	FOS [3]	BVS [15]	OSS [16]	Ours Dis.	Ours Seg.
birdfall2	49.4	17.5	**63.5**	38.1	34.1	63.0
bird of paradise	92.4	81.8	91.7	67.4	84.5	**94.5**
frog	75.7	54.1	76.4	71.0	53.6	**82.1**
girl	64.2	54.9	79.1	**87.9**	56.0	85.5
monkey	82.6	65.0	85.9	**88.2**	69.7	69.0
parachute	**94.6**	76.3	93.8	79.8	86.3	91.2
soldier	60.8	39.8	56.4	**85.8**	56.1	80.5
worm	62.2	72.8	65.5	63.1	63.5	**78.8**
Avg.	72.7	57.8	76.5	72.2	59.2	**80.6**

TABLE III
THE AVERAGE IOU SCORES OF OUR METHOD AND THREE SINGLE VIDEO OBJECT SEGMENTATION METHODS ON THE VIDEOS CONTAINING ONLY ONE OBJECT OF THE YOUTUBE-OBJECTS DATASET. HIGHER VALUES ARE BETTER.

Video	FOS [3]	BVS [15]	OSS [16]	Ours Dis.	Ours Seg.
aeroplane	83.9	**90.8**	84.4	73.9	88.1
bird	80.9	**89.5**	85.6	76.1	88.1
boat	35.1	72.7	**75.1**	58.0	71.8
car	69.1	64.5	**69.3**	53.0	68.8
cat	57.8	62.7	**73.8**	41.2	65.9
dog	54.8	78.2	**87.7**	46.8	72.4
motorbike	21.8	55.8	**68.0**	33.9	55.3
train	21.8	53.5	54.4	54.9	**71.9**
Avg.	53.1	71.0	**74.8**	54.7	72.8

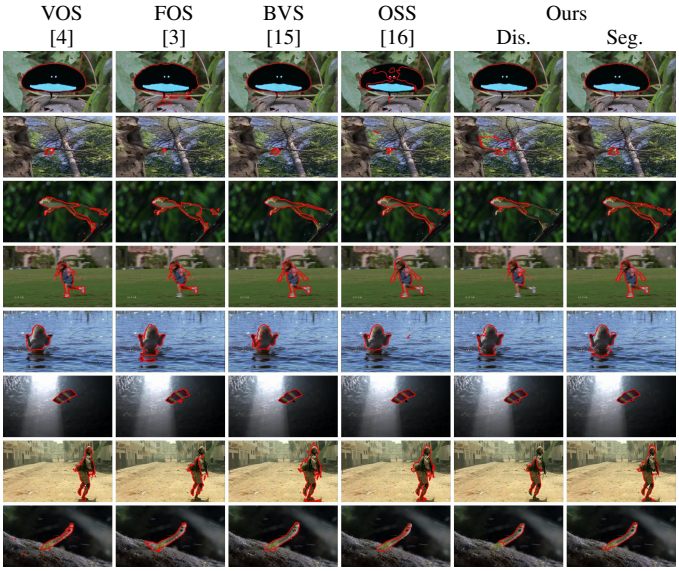


Fig. 9. Some example results of our method and four single video object segmentation methods on eight videos that contain only one object of the SegTrack dataset.

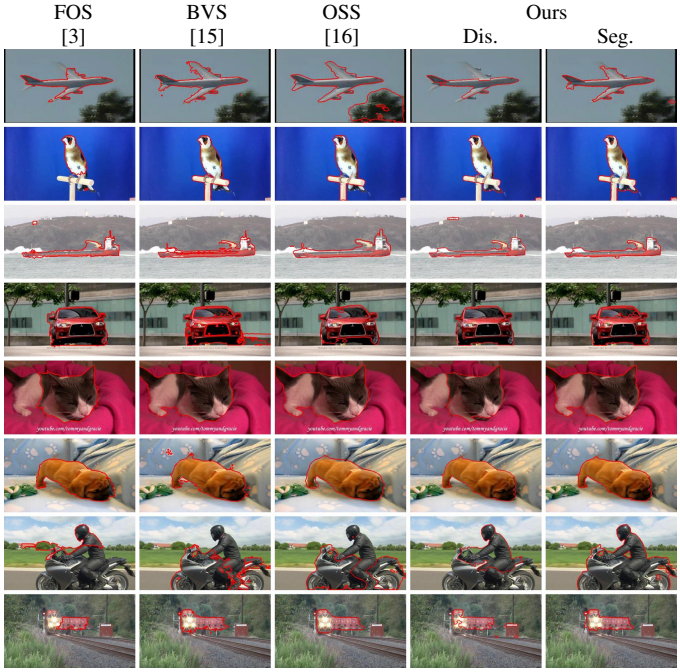


Fig. 10. Some example results of our methods and three single video object segmentation methods on the videos containing only one object of the YouTube-Objects dataset.

Fig. 9, respectively. Besides the qualitative and quantitative results obtained by object segmentation of our method, we also present the average IoU scores and some example object regions obtained by object discovery of our method.

The results show that our method outperforms all other state-of-the-art methods. But on the videos of monkey and soldier, our method erroneously segment the shadow of the monkey in water and the shadow of the soldier as foreground objects, and thus does not performs well. The above results clearly demonstrate that our method can handle certain variations in shape (frog and worm), appearance (bird of paradise), and illumination (parachute), but has encountered difficulties when there are large shadows that have similar motion or color with the objects (monkey and soldier).

Evaluation on the YouTube-Objects dataset. Similarly, we evaluate our method and compare with three single video object segmentation methods (FOS [3], BVS [15], and OSS [16]) on the 83 videos that contain only one object. We present the average IoU scores of them in Table III, and some example results of them in Fig. 10. For fair comparison, we computed the IoU scores of BVS [15] and OSS [16] using the final segmentation masks provided by them, respectively.

The results show that our method outperforms FOS [3] and BVS [15], but performs poorer than OSS [16]. This is because the semi-supervise method OSS [16] can leverage the segmentation mask of the first frame to separate the object from its ambiguous surrounding, while our method segments the object and its connective surrounding with similar motion as a whole. As illustrated by the videos of motorbike and boat in Fig. 11, the persons on the motorbike and boat are all labeled as background in the ground truth, although they move together with the motorbike and boat.

Evaluation on the DAVIS dataset. We test our method and compare with four single video object segmentation methods (VOS [4], FOS [3], BVS [15], and OSS [16]) on all 50 videos of the DAVIS dataset. The average IoU scores and some qualitative results of them are presented in Table IV

Fig. 11. Some examples of the ground truth segmentations provided by [13] of the YouTube-Objects dataset.

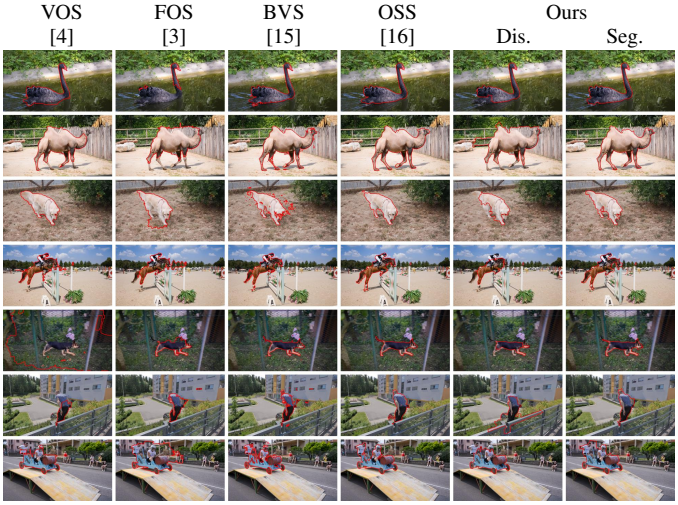


Fig. 12. Some visual example results of our methods and four single video object segmentation methods on the DAVIS dataset.

and Fig. 12, respectively.

The results reveal that our method largely outperforms VOS [4], FOS [3], and BVS [15] by a margin from 7.9% to 17.5%, although BVS [15] exploits the segmentation mask of the first frame to facilitate the segmentation procedure. There is a margin of 5.4% between our method and OSS [16]. This is mainly because the semi-supervised method OSS [16] uses not only the segmentation mask of the first frame of each video, but also a large video set (30 of 50 videos) of the DAVIS dataset for training to obtain their final results on the remaining 20 videos, while our method is unsupervised.

C. Joint Object Discovery and Segmentation from a Single Video with Noisy Frames

We further evaluate the joint object discovery and segmentation performance from a single video with noisy frames of our method on the XJTU-Stevens dataset [34], [35] and Noisy-ViDiSeg dataset, both of them have many noisy frames that do not contain the objects.

Evaluation on the XJTU-Stevens dataset. The XJTU-Stevens dataset is a video object co-segmentation and classification dataset, in which 3.7% of the frames are noisy frames. Besides the four single video segmentation methods (VOS [4], FOS [3], BVS [15], and OSS [16]), we also compare our method with two multi-video object co-segmentation methods (VOC [40] and VDC [35]). We implement two versions of VDC [35], one is its original version operating on multiple videos, and the other one is operating on one single video instead of multiple videos, which becomes a single video object discovery and segmentation method VDS [35].

We present the average IoU scores of object segmentation in Table V, the labeling accuracies of object discovery in Table VI, and some qualitative results in Fig. 13. As they show,

TABLE IV
THE AVERAGE IOU SCORES OF OUR METHOD AND FOUR SINGLE VIDEO OBJECT SEGMENTATION METHODS ON THE DAVIS DATASET. THE 30 VIDEOS USED BY OSS FOR TRAINING ARE ANNOTATED BY "-". HIGHER VALUES ARE BETTER.

Video	VOS [4]	FOS [3]	BVS [15]	OSS [16]	Ours Dis.	Ours Seg.
bear	89.1	89.8	**95.5**	-	85.0	91.3
Blackswan	84.2	73.2	**94.3**	94.2	83.4	91.5
Bmx-Bumps	30.9	24.1	43.4	-	10.6	**45.2**
Bmx-Trees	19.3	18.0	38.2	**55.5**	33.5	41.1
Boat	6.5	36.1	**64.4**	-	56.5	63.1
Breakdance	54.9	46.7	50.0	**70.8**	42.5	52.9
Breakdance-Flare	55.9	61.6	**72.7**	-	45.8	60.2
Bus	78.5	82.5	86.3	-	74.1	**87.0**
Camel	57.9	56.2	66.9	**85.1**	64.7	82.7
Car-Roundabout	64.0	80.8	85.1	**95.3**	68.3	75.2
Car-Shadow	58.9	69.8	57.8	**93.7**	59.3	75.9
Car-Turn	80.6	85.1	84.4	-	75.0	**85.9**
Cows	33.7	79.1	89.5	**94.6**	65.6	88.7
Dance-Jump	**74.8**	59.8	74.5	-	38.9	64.2
Dance-Twirl	38.0	45.3	49.2	**67.0**	44.9	60.6
Dog	69.2	70.8	72.3	**90.7**	66.3	86.4
Dog-Agility	13.2	28.0	34.5	-	45.2	**68.4**
Drift-Chicane	18.8	66.7	3.3	**83.5**	6.2	71.5
Drift-Straight	19.4	**68.3**	40.2	67.6	56.4	66.6
Drift-Turn	25.5	53.3	29.9	-	50.2	**58.1**
Elephant	67.5	82.4	85.0	-	55.9	**89.2**
Flamingo	69.2	81.7	**88.1**	-	55.9	81.3
Goat	70.5	55.4	66.1	**88.0**	62.2	79.4
Hike	89.5	88.9	75.5	-	75.5	**89.8**
Hockey	51.5	46.7	**82.9**	-	50.5	64.8
Horsejump-High	37.0	57.8	80.1	78.0	63.8	**80.9**
Horsejump-Low	63.0	52.6	60.1	-	49.9	**75.8**
Kite-Surf	58.5	27.2	42.5	68.6	48.9	**68.7**
Kite-Walk	19.7	64.9	**87.0**	-	67.0	71.6
Libby	61.1	50.7	77.6	**80.8**	34.2	79.9
Lucia	84.7	64.4	**90.1**	-	68.9	85.4
Mallard-Fly	58.5	60.1	**60.6**	-	38.8	42.9
Mallard-Water	78.5	8.7	**90.7**	-	38.3	74.6
Motocross-Bumps	68.9	61.7	40.1	-	58.6	**83.3**
Motocross-Jump	28.8	60.2	34.1	**81.6**	44.6	68.6
Motorbike	57.2	55.9	56.3	-	53.6	**66.1**
Paragliding	86.1	72.5	87.5	-	74.8	**90.2**
Paragliding-Launch	55.9	50.6	**64.0**	62.5	56.7	60.0
Parkour	41.0	45.8	75.6	**85.5**	64.4	77.9
Rhino	67.5	77.6	78.2	-	73.6	**83.8**
Rollerblade	51.0	31.8	58.8	-	39.9	**77.0**
Scooter-Black	50.2	52.2	33.7	**71.1**	47.1	44.5
Scooter-Gray	36.3	32.5	50.8	-	47.0	**66.1**
Soapbox	75.7	41.0	78.9	**81.2**	58.6	79.4
Soccerball	87.9	84.3	84.4	-	76.8	**88.5**
Stroller	75.9	58.0	76.7	-	52.3	**87.8**
Surf	89.3	47.5	49.2	-	70.4	**92.5**
Swing	71.0	43.1	78.4	-	59.9	**83.8**
Tennis	76.2	38.8	73.7	-	36.8	**78.6**
Train	45.0	83.1	87.2	-	46.0	**91.4**
Avg.	56.9	57.5	66.5	**79.8**	54.9	74.4

TABLE V
THE AVERAGE IOU SCORES OF OUR METHOD, FOUR SINGLE VIDEO OBJECT SEGMENTATION METHODS, AND TWO MULTI-VIDEO OBJECT CO-SEGMENTATION METHODS ON THE XJTU-STEVENS DATASET. HIGHER VALUES ARE BETTER.

Video	VOS [4]	FOS [3]	BVS [15]	OSS [16]	VOC [40]	VDC [35]	VDS [35]	Ours Dis.	Ours Seg.
airplane	19.7	69.9	35.1	80.3	61.2	**86.4**	75.2	54.6	78.4
balloon	77.4	60.3	90.8	86.5	87.4	**94.6**	86.5	81.7	94.0
bear	88.3	80.7	82.0	**92.8**	85.9	90.5	86.1	79.5	89.7
cat	30.3	66.8	42.2	74.0	80.7	**92.1**	79.7	66.4	75.7
eagle	37.3	69.2	37.7	65.6	79.5	**89.5**	80.9	48.2	74.2
ferrari	36.0	70.7	50.5	84.0	62.1	**87.7**	75.4	71.7	86.7
figure skating	62.4	25.5	48.7	58.4	65.8	**88.5**	74.6	45.3	72.5
horse	75.7	72.3	76.8	91.9	86.2	**92.0**	85.8	68.6	86.3
parachute	52.3	48.3	72.9	73.1	84.7	**94.0**	83.9	58.5	88.0
single diving	59.7	49.2	30.7	70.3	72.0	**87.7**	76.8	54.2	73.1
Avg.	53.9	61.3	54.3	77.7	76.6	**90.3**	80.5	62.9	81.9

TABLE VI
THE LABELING ACCURACIES OF OBJECT DISCOVERY OF OUR METHOD, FOUR SINGLE VIDEO OBJECT SEGMENTATION METHODS, AND TWO MULTI-VIDEO OBJECT CO-SEGMENTATION METHODS ON THE XJTU-STEVENS DATASET. HIGHER VALUES ARE BETTER.

Video	APR	VOS [4]	FOS [3]	BVS [15]	OSS [16]	VOC [40]	VDC [35]	VDS [35]	Ours
airplane	96.5	95.1	98.1	96.5	98.4	96.5	**100.0**	96.5	99.4
balloon	95.5	95.5	96.0	94.9	96.1	95.5	**99.8**	95.5	98.2
bear	95.8	96.3	97.7	96.6	97.6	95.8	99.8	95.8	**99.9**
cat	97.6	97.8	87.8	97.6	97.6	97.6	**99.2**	97.6	97.3
eagle	97.8	97.8	95.4	96.7	98.4	97.8	**99.5**	97.8	97.2
ferrari	97.8	97.8	99.0	97.8	98.9	97.8	**99.5**	97.8	99.4
figure skating	95.1	96.2	93.3	95.1	95.1	95.1	**100.0**	95.1	**100.0**
horse	95.4	95.8	97.3	95.4	97.2	95.4	99.9	95.4	**100.0**
parachute	97.3	97.5	95.8	97.3	96.7	97.3	**99.9**	97.3	96.6
single diving	94.8	94.1	92.5	88.9	97.1	94.8	**99.6**	94.8	97.9
Avg.	96.3	96.3	95.8	95.6	97.4	96.3	**99.7**	96.3	98.6

our method outperforms all other methods in terms of both IoU scores for object segmentation and labeling accuracies for object discovery, except VDC [35].

In terms of object segmentation, our method is greatly superior in IoU score to not only four single video object segmentation methods (VOS [4], FOS [3], BVS [15], and OSS [16]) by a margin from 4.2% to 28%, but also the multi-video object co-segmentation method VOC [40] by a margin of 5.3%.

Although our method is inferior to the multi-video object discovery and co-segmentation method VDC [35], our method is better than its variant VDS [35], *i.e.*, a single video object discovery and segmentation method. The reasons are two-fold, one is that VDC [35] can leverage the contextual information of the common objects from multiple videos to facilitate both the object discovery and object segmentation of each single video, and the other one is that VDC [35] is bootstrapped with the frame-level object discovery labels for three frames of each video.

In terms of object discovery, our method achieves a higher labeling accuracy than VOS [4], FOS [3], BVS [15], OSS [16], VOC [40], and VDS [35], but is slightly lower than VDC [35]. The reasons are three-fold, the first one is that VOS [4], FOS [3], BVS [15], VOC [40], and VDS [35] almost all cannot distinguish the positive frames that contain the object from the noisy frames, thus their labeling accuracies of object discovery are equal to or lower than the actual positive rate (APR) of the frames of each video category.

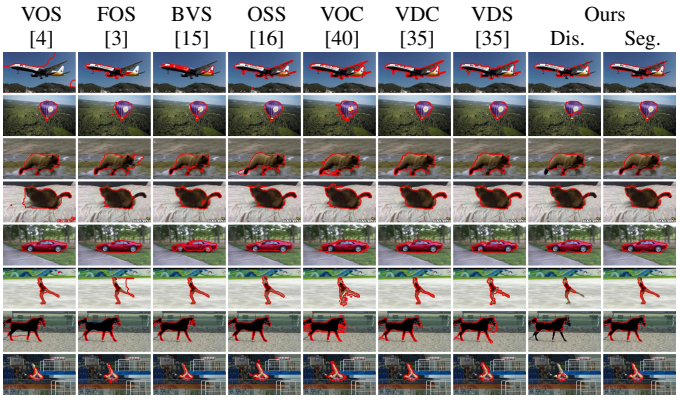


Fig. 13. Some qualitative results of our method, four single video object segmentation methods, and two multi-video object co-segmentation methods on the XJTU-Stevens dataset.

The second one is that only 3.7% of the frames are noisy frames in the video dataset, and most of the noisy frames come from the different video shot with the positive frames, thus it is easy to identify the noisy frames. The last but the most important one is that our method and VDC [35] indeed are able to identify the object from the noisy video, where VDC [35] needs to be bootstrapped by three frame-level discovery labels, while our method does not need any supervision.

Evaluation on the Noisy-ViDiSeg dataset. The Noisy-ViDiSeg dataset is a newly introduced object discovery and segmentation dataset in this paper, in which 33.1% of the frames are noisy frames. We test our method and compare with four single video object segmentation methods (VOS [4], FOS [3], BVS [15], and OSS [16]) and two multi-video object co-segmentation methods (VOC [40] and VDC [35]). Because there is only one video in each video category, VOC [40] becomes a single video object segmentation method, and VDC [35] becomes a single video object discovery and segmentation method, *i.e.*, VDS [35].

The average IoU scores of object segmentation, the labeling accuracies of object discovery, and some qualitative results are presented in Table VII, Table VIII and Fig. 14, respectively. They show that, our method outperforms all other methods in terms of both object segmentation and object discovery. This strongly validates the efficacy of our joint object discovery and segmentation method.

For object segmentation, our method improves the state-of-the-art methods by a margin from 4.2% to 53.4%. This is mainly because all the other methods encounter difficulties when the object in each video may disappear at any time and exhibits complex temporary occlusions and dramatic changes in appearance, size, and shape, while our method can better handle these cases.

For object discovery, our method outperforms the state-of-the-art methods by a significant margin from 8.4% to 32.3%. The reason is that our method is able to distinguish the video frames that contain the object from the noisy frames in a single

TABLE VII
THE AVERAGE IOU SCORES OF OUR METHOD, FOUR SINGLE VIDEO OBJECT SEGMENTATION METHODS, AND TWO MULTI-VIDEO OBJECT CO-SEGMENTATION METHODS ON THE NOISY-VIDISEG DATASET. HIGHER VALUES ARE BETTER.

Video	VOS [4]	FOS [3]	BVS [15]	OSS [16]	VOC [40]	VDC [35]	Ours Dis.	Ours Seg.
F1	77.2	8.6	26.9	77.3	8.9	78.2	68.2	**81.9**
airplane	30.0	48.8	34.6	34.6	8.3	57.6	43.7	**65.8**
gymnastics	14.2	55.6	16.6	61.9	10.6	70.8	68.7	**76.9**
lion	27.3	62.6	51.1	**79.0**	27.4	71.4	61.0	76.1
ostrich	57.4	57.4	2.5	**70.7**	1.1	61.3	60.2	63.0
panda	29.1	33.6	75.3	82.1	62.7	79.5	57.2	**85.7**
parkour	66.9	69.4	54.7	82.7	56.8	80.6	59.5	**85.1**
rock	6.9	44.4	18.0	**79.0**	3.2	70.8	55.4	73.0
skiing	66.3	62.6	2.8	65.2	46.6	67.9	79.0	**85.0**
surfing	56.3	55.2	35.2	57.7	1.6	57.2	45.3	**61.0**
tiger	63.2	54.1	58.2	**94.8**	16.7	76.7	51.2	77.7
Avg.	45.0	50.2	34.2	71.4	22.2	70.2	59.0	**75.6**

TABLE VIII
THE LABELING ACCURACIES OF OBJECT DISCOVERY OF OUR METHOD, FOUR SINGLE VIDEO OBJECT SEGMENTATION METHODS, AND TWO MULTI-VIDEO OBJECT CO-SEGMENTATION METHODS ON THE NOISY-VIDISEG DATASET. HIGHER VALUES ARE BETTER.

Video	APR	VOS [4]	FOS [3]	BVS [15]	OSS [40]	VOC [16]	VDC [35]	Ours
F1	56.3	95.2	56.3	69.0	99.2	56.3	94.2	**100.0**
airplane	62.6	62.9	61.4	62.7	71.7	62.7	81.9	**95.6**
gymnastics	61.1	69.9	98.2	61.1	61.1	61.1	87.4	**99.1**
lion	71.6	86.4	97.7	71.6	**98.9**	71.6	92.5	96.6
ostrich	68.4	81.8	**93.1**	65.6	91.1	68.4	89.1	91.5
panda	84.5	84.5	84.5	84.5	90.8	84.5	89.8	**99.4**
parkour	76.7	76.7	76.7	76.7	76.7	76.7	77.5	**93.0**
rock	42.1	42.1	**100.0**	42.1	96.2	42.1	97.1	99.2
skiing	81.3	90.6	95.9	36.8	**97.1**	81.3	83.6	93.0
surfing	66.5	82.5	81.6	66.5	**99.1**	66.5	87.1	91.0
tiger	71.8	71.8	71.8	71.8	**100.0**	71.8	74.3	**100.0**
Avg.	66.9	74.9	79.9	63.5	87.4	66.9	86.8	**95.8**

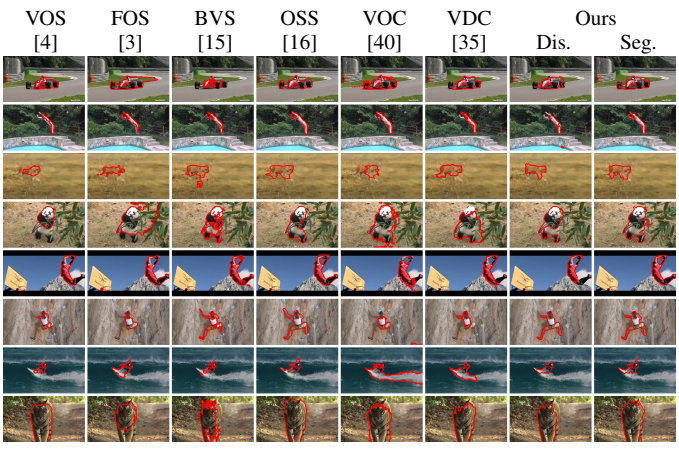


Fig. 14. Some visual example results of our method, four single video object segmentation methods, and two multi-video object co-segmentation methods on the Noisy-ViDiSeg dataset.

TABLE IX
THE AVERAGE IOU SCORES OF DIFFERENT CHOICES ON SUPERPIXEL AND OBJECT PROPOSAL ALGORITHMS ON THE NOISY-VIDISEG DATASET. HIGHER VALUES ARE BETTER.

Video	COP [36] GS[61]	COP [36] SLIC[42]	COP [36] ES[62]	GOP [27] GS[61]	GOP [27] SLIC[42]	GOP [27] ES[62]
F1	80.9	**82.3**	82.2	80.7	81.9	82.2
airplane	62.4	62.7	62.5	65.4	**65.8**	65.3
gymnastics	62.1	64.5	65.4	75.5	76.9	**80.1**
lion	78.6	78.2	**79.0**	76.1	76.1	76.2
ostrich	65.1	64.4	**65.4**	65.2	63.0	**65.4**
panda	83.6	84.5	84.1	85.5	85.7	**86.0**
parkour	83.2	84.4	**85.9**	82.9	85.1	85.7
rock	74.1	76.0	**77.3**	71.1	73.0	75.4
skiing	82.8	83.6	85.2	83.7	85.0	**85.4**
surfing	65.0	64.5	**74.0**	64.6	61.0	62.6
tiger	74.2	74.5	71.2	77.2	**77.7**	75.6
Avg.	73.8	74.5	75.7	75.3	75.6	**76.4**

video, while all the other methods do not have the ability or the ability is too weak, when there are a large number of noisy frames in a single video.

Please note that, we also present the average IoU scores and some examples of the object regions selected by object discovery of our method on the above five datasets. They show that the object regions selected by object discovery almost always focus on the object, and the majority of them belong to the type of "object region" as defined in Section V-C, this is due to the collaboration of object discovery and object segmentation of our method. Moreover, although the average IoU scores of the object regions selected by object discovery of our method are not high, compared to the average IoU scores obtained by object segmentation of our method and other state-of-the-art methods, they indeed facilitate the object segmentation procedure of our method.

Impact of superpixel and object proposal algorithms. To quantify the impact of the different superpixel algorithms, we compare the performance of our method with SLIC [42], GS [61] and ES [62]. To quantify the impact of different object proposal algorithms, we compare the performance of our method with GOP [27] and COP [36]. With these different variants of our methods, the average IoU scores on the Noisy-ViDiSeg dataset are summarized in Table IX and some qualitative examples are illustrated in Fig 15. As shown in Table IX, the performance differences are within 2.6%, demonstrating that our method is robust to these variations and not tied to specific superpixel or object proposal algorithms.

To summarize, the results on the above five datasets clearly reveal that, although our method produces inferior results on video datasets without noisy frames, we are able to obtain better results on video datasets with noisy frames, when compared with state-of-the-art. Moreover, as there are more noisy frames in the video dataset, the performance of our method becomes better, while other methods perform poorer. This strongly demonstrates that our method is capable of jointly discovering and segmenting the object from a single noisy video, where object discovery and object segmentation work in a collaborative way.

Fig. 15. Some visual examples of different choices on superpixel and object proposal algorithms on the Noisy-ViDiSeg dataset.

VII. Conclusion

We presented a method to jointly discover and segment an object from a single video, in which there are a large number of irrelevant frames devoid of the target object. Our method overcomes a limitation that previous methods either only fulfill video object discovery or video object segmentation requiring all video frames contain the object. We proposed a principle probabilistic model, in which video object discovery and video object segmentation are cast into two coupled dynamic Markov networks. The bi-directional message passing revealed the collaboration between the two tasks. Experiments on five video datasets validated the efficacy of our proposed method.

Appendix I

The exact inference of the marginal posterior probabilities $p(\mathbf{L}|\mathbf{O},\mathbf{S})$ and $p(\mathbf{B}|\mathbf{O},\mathbf{S})$ can be calculated by belief propagation algorithm through a local message passing process. The local messages passing from $\mathbf{B}$ to $\mathbf{L}$ and from $\mathbf{L}$ to $\mathbf{B}$ are

$$\mathbf{m_{BL}}(\mathbf{L}) \leftarrow \int_{\mathbf{B}} p(\mathbf{S}|\mathbf{B})\Psi(\mathbf{L},\mathbf{B})d\mathbf{B}, \tag{18}$$

$$\mathbf{m_{LB}}(\mathbf{B}) \leftarrow \int_{\mathbf{L}} p(\mathbf{O}|\mathbf{L})\Psi(\mathbf{L},\mathbf{B})d\mathbf{L}. \tag{19}$$

By iterating the message passing until convergence, the marginal posterior probabilities of $\mathbf{L}$ and $\mathbf{B}$ are obtained as

$$p(\mathbf{L}|\mathbf{O},\mathbf{S}) \propto p(\mathbf{O}|\mathbf{L})\mathbf{m_{BL}}(\mathbf{L}), \tag{20}$$

$$p(\mathbf{B}|\mathbf{O},\mathbf{S}) \propto p(\mathbf{S}|\mathbf{B})\mathbf{m_{LB}}(\mathbf{B}). \tag{21}$$

Appendix II

The belief propagation algorithm is extended to infer the marginal posterior probabilities $p(\mathbf{L}_t|\underline{\mathbf{O}}_t,\underline{\mathbf{S}}_t)$ and $p(\mathbf{B}_t|\underline{\mathbf{O}}_t,\underline{\mathbf{S}}_t)$ on the two coupled dynamic Markov networks. The dynamic models in object discovery and object segmentation are assumed to be independent

$$p(\mathbf{L}_t,\mathbf{B}_t|\mathbf{L}_{t-1},\mathbf{B}_{t-1}) = p(\mathbf{L}_t|\mathbf{L}_{t-1})p(\mathbf{B}_t|\mathbf{B}_{t-1}). \tag{22}$$

Given the inference results both at previous time $t-1$ ($p(\mathbf{L}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1},\underrightarrow{\mathbf{S}}_{t-1})$ and $p(\mathbf{B}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1},\underrightarrow{\mathbf{S}}_{t-1})$) and next time $t+1$ ($p(\mathbf{L}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1},\underleftarrow{\mathbf{S}}_{t+1})$, and $p(\mathbf{B}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1},\underleftarrow{\mathbf{S}}_{t+1})$), the messages updating at time t from $\mathbf{B}$ to $\mathbf{L}$ and from $\mathbf{L}$ to $\mathbf{B}$ are executed in a bi-directional way as

$$\begin{aligned}\mathbf{m_{BL}}(\mathbf{L}_t) \leftarrow \int_{\mathbf{B}_t} \Big[& p(\mathbf{S}_t|\mathbf{B}_t)\Psi_{\mathbf{BL}}(\mathbf{B}_t,\mathbf{L}_t) \\ & \times \int_{\mathbf{B}_{t-1}} p(\mathbf{B}_t|\mathbf{B}_{t-1})p(\mathbf{B}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1},\underrightarrow{\mathbf{S}}_{t-1})d\mathbf{B}_{t-1} \\ & \times \int_{\mathbf{B}_{t+1}} p(\mathbf{B}_t|\mathbf{B}_{t+1})p(\mathbf{B}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1},\underleftarrow{\mathbf{S}}_{t+1})d\mathbf{B}_{t+1}\Big] d\mathbf{B}_t,\end{aligned} \tag{23}$$

$$\begin{aligned}\mathbf{m_{LB}}(\mathbf{B}_t) \leftarrow \int_{\mathbf{L}_t} \Big[& p(\mathbf{O}_t|\mathbf{L}_t)\Psi_{\mathbf{LB}}(\mathbf{L}_t,\mathbf{B}_t) \\ & \times \int_{\mathbf{L}_{t-1}} p(\mathbf{L}_t|\mathbf{L}_{t-1})p(\mathbf{L}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1},\underrightarrow{\mathbf{S}}_{t-1})d\mathbf{L}_{t-1} \\ & \times \int_{\mathbf{L}_{t+1}} p(\mathbf{L}_t|\mathbf{L}_{t+1})p(\mathbf{L}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1},\underleftarrow{\mathbf{S}}_{t+1})d\mathbf{L}_{t+1}\Big] d\mathbf{L}_t.\end{aligned} \tag{24}$$

The marginal posterior probabilities of $\mathbf{L}$ and $\mathbf{B}$ at time t are computed by combing the incoming messages from both its forward and backward neighborhood as

$$\begin{aligned}p(\mathbf{L}_t|\mathbf{O},\mathbf{S}) = & p(\mathbf{O}_t|\mathbf{L}_t)\mathbf{m_{BL}}(\mathbf{L}_t) \\ & \times \int_{\mathbf{L}_{t-1}} p(\mathbf{L}_t|\mathbf{L}_{t-1})p(\mathbf{L}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1},\underrightarrow{\mathbf{S}}_{t-1})d\mathbf{L}_{t-1} \\ & \times \int_{\mathbf{L}_{t+1}} p(\mathbf{L}_t|\mathbf{L}_{t+1})p(\mathbf{L}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1},\underleftarrow{\mathbf{S}}_{t+1})d\mathbf{L}_{t+1},\end{aligned} \tag{25}$$

$$\begin{aligned}p(\mathbf{B}_t|\mathbf{O},\mathbf{S}) = & p(\mathbf{S}_t|\mathbf{B}_t)\mathbf{m_{LB}}(\mathbf{B}_t) \\ & \times \int_{\mathbf{B}_{t-1}} p(\mathbf{B}_t|\mathbf{B}_{t-1})p(\mathbf{B}_{t-1}|\underrightarrow{\mathbf{O}}_{t-1},\underrightarrow{\mathbf{S}}_{t-1})d\mathbf{B}_{t-1} \\ & \times \int_{\mathbf{B}_{t+1}} p(\mathbf{B}_t|\mathbf{B}_{t+1})p(\mathbf{B}_{t+1}|\underleftarrow{\mathbf{O}}_{t+1},\underleftarrow{\mathbf{S}}_{t+1})d\mathbf{B}_{t+1}.\end{aligned} \tag{26}$$

References

[1] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 282–295.
[2] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1846–1853.
[3] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1777–1784.
[4] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1995–2002.
[5] A. Khoreva, F. Galasso, M. Hein, and B. Schiele, "Classifier based graph construction for video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 951–960.
[6] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting." in *Proc. British Mach. Vis. Conf.*, vol. 2, no. 7, 2014, p. 8.
[7] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3395–3402.
[8] W.-D. Jang, C. Lee, and C.-S. Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 696–704.
[9] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3664–3673.

[10] B. Luo, H. Li, F. Meng, Q. Wu, and K. Ngan, "An unsupervised method to extract video object via complexity awareness and object local parts," *IEEE Trans. Circuits Syst. Video Technol.*, 2017.

[11] V. Badrinarayanan, F. Galasso, and R. Cipolla, "Label propagation in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3265–3272.

[12] X. Xiang, H. Chang, and J. Luo, "Online web-data-driven segmentation of selected moving objects in videos," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 134–146.

[13] S. D. Jain and K. Grauman, "Supervoxel-consistent foreground propagation in video," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 656–671.

[14] Y. H. Tsai, M. H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3899–3908.

[15] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 743–751.

[16] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 221–230.

[17] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: robust video object cutout using localized classifiers," in *ACM Trans. Graph.*, vol. 28, no. 3, 2009, p. 70.

[18] B. L. Price, B. S. Morse, and S. Cohen, "Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 779–786.

[19] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen, "Jumpcut: non-successive mask transfer and interpolation for video cutout," *ACM Trans. Graph.*, vol. 34, no. 6, p. 195, 2015.

[20] N. Shankar Nagaraja, F. R. Schmidt, and T. Brox, "Video segmentation with just a few strokes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3235–3243.

[21] Y. Lu, X. Bai, L. Shapiro, and J. Wang, "Coherent parametric contours for interactive video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 642–650.

[22] T. Ma and L. J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 670–677.

[23] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 628–635.

[24] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4083–4090.

[25] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung, "Fully connected object proposals for video segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3227–3234.

[26] Y. J. Koh and C.-S. Kim, "Primary object segmentation in videos based on region augmentation and reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3442–3450.

[27] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 725–739.

[28] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 105–112.

[29] J. Xue, L. Wang, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting," *Pattern Recognition*, vol. 46, no. 11, pp. 2874–2889, 2013.

[30] D. Tsai, M. Flagg, and J. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. British Mach. Vis. Conf.*, 2010, pp. 56–67.

[31] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2192–2199.

[32] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3282–3289.

[33] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 724–732.

[34] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 640–655.

[35] L. Wang, G. Hua, R. Sukthankar, Z. Niu, J. Xue, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2074–2088, 2017.

[36] I. Endres and D. Hoiem, "Category independent object proposals," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 575–588.

[37] W.-C. Chiu and M. Fritz, "Multi-class video co-segmentation with a generative multi-video model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 321–328.

[38] H. Fu, D. Xu, B. Zhang, and S. Lin, "Object-based multiple foreground video co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3166–3173.

[39] Z. Lou and T. Gevers, "Extracting primary objects by video co-segmentation," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2110–2117, 2014.

[40] D. Zhang, O. Javed, and M. Shah, "Video object co-segmentation by regulated maximum weight cliques," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 551–566.

[41] X. Lv, L. Wang, Q. Zhang, Z. Niu, N. Zheng, and G. Hua, "Video object co-segmentation from noisy videos by a multi-level hypergraph model," in *Proc. IEEE Int. Conf. Image Process.*, 2018.

[42] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[43] Y. Wu, G. Hua, and T. Yu, "Tracking articulated body by dynamic markov network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1094–1101.

[44] G. Hua and Y. Wu, "Variational maximum a posteriori by annealed mean field analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1747–1761, 2005.

[45] ——, "Multi-scale visual tracking by sequential belief propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 826–833.

[46] M. I. Jordan and Y. Weiss, "Graphical models: Probabilistic inference," *The handbook of brain theory and neural networks*, pp. 490–496, 2002.

[47] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 25–47, 2000.

[48] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.

[49] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1841–1848.

[50] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Learning to detect motion boundaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2578–2586.

[51] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1744–1757, 2012.

[52] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 840–853, 2007.

[53] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[54] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1932–1939.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[57] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

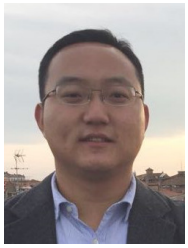
[58] L. Wang, G. Hua, J. Xue, Z. Gao, and N. Zheng, "Joint segmentation and recognition of categorized objects from noisy web image collection," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4070–4086, 2014.

[59] W. Liu, G. Hua, and J. R. Smith, "Unsupervised one-class learning for automatic outlier removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3826–3833.

[60] K. Tang, R. Sukthankar, J. Yagnik, and F. F. Li, "Discriminative segment annotation in weakly labeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2483–2490.

[61] G. Mori, "Guiding model search using segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 1417–1423.
[62] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, 2015.

Ziyi Liu received the B.S. degree in Control Science and Engineering from Xi'an Jiaotong University in 2015. He is currently a Ph.D. student with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University. His research interests include computer vision and machine learning. He is a student member of the IEEE.

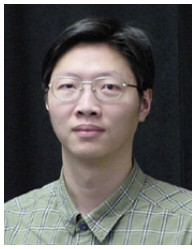


Le Wang (M'14) received the B.S. and Ph.D. degrees in Control Science and Engineering from Xi'an Jiaotong University in 2008 and 2014, respectively. From 2013 to 2014, he was a visiting Ph.D. student with Stevens Institute of Technology. From 2016 to 2017, he is a visiting scholar with Northwestern University. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University. His research interests include computer vision, machine learning, and their application for web images and videos. He is the author of more than 10 peer reviewed publications in prestigious international journals and conferences. He is a member of the IEEE.

Gang Hua (M'03-SM'11) was enrolled in the Special Class for the Gifted Young of Xi'an Jiaotong University (XJTU) in 1994 and received the B.S. degree in Automatic Control Engineering from XJTU in 1999. He received the M.S. degree in Control Science and Engineering in 2002 from XJTU, and the Ph.D. degree in Electrical Engineering and Computer Science at Northwestern University in 2006. He is currently a Principle Researcher/Research Manager at Microsoft Research. Before that, he was an Associate Professor of Computer Science at Stevens Institute of Technology. He also held an Academic Advisor position at IBM T. J. Watson Research Center between 2011 and 2014. He was a Research Staff Member at IBM Research T. J. Watson Center from 2010 to 2011, a Senior Researcher at Nokia Research Center, Hollywood from 2009 to 2010, and a Scientist at Microsoft Live Labs Research from 2006 to 2009. He is currently an Associate Editor in Chief for CVIU, and Associate Editors for IJCV, IEEE T-IP, IEEE T-CSVT, IEEE Multimedia, and MVA. He also served as the Lead Guest Editor on two special issues in TPAMI and IJCV, respectively. He is a program chair of CVPR'2019&2022. He is an area chair of CVPR'2015&2017, ICCV'2011&2017, ICIP'2012&2013&2016, ICASSP'2012&2013, and ACM MM 2011&2012&2015&2017. He is the author of more than 150 peer reviewed publications in prestigious international journals and conferences. He holds 19 issued US patents and has 20 more US patents pending. He is the recipient of the 2015 IAPR Young Biometrics Investigator Award for his contribution on Unconstrained Face Recognition from Images and Videos, and a recipient of the 2013 Google Research Faculty Award. He is an IAPR Fellow, an ACM Distinguished Scientist, and a senior member of the IEEE.

Qilin Zhang received the B.E. degree in Electrical Information Engineering from the University of Science and Technology of China, Hefei, China, in 2009, the M.S. degree in Electrical and Computer Engineering from University of Florida, Gainesville, Florida, USA in 2011, and the Ph.D. degree in Computer Science from Stevens Institute of Technology, Hoboken, New Jersey, USA, in 2016. He is currently a Senior Research Engineer with Here Technologies, Chicago, Illinois, USA. His research interests include computer vision, machine learning and autonomous driving. He is the author of more than 10 peer reviewed publications in international journals and conferences. He is a member of the IEEE.

Zhenxing Niu received the Ph.D. degree in Control Science and Engineering from Xidian University, Xi'an, China, in 2012. From 2013 to 2014, he was a visiting scholar with University of Texas at San Antonio, Texas, USA. He is a Researcher at Alibaba Group, Hangzhou, China. Before joining Alibaba Group, he is an Associate Professor of School of Electronic Engineering at Xidian University, Xi'an, China. His research interests include computer vision, machine learning, and their application in object discovery and localization. He served as PC member of CVPR, ICCV, and ACM Multimedia. He is a member of the IEEE.



Ying Wu (SM'06-F'16) received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, USA, in 1994, 1997, and 2001, respectively. From 1997 to 2001, he was a Research Assistant with the Beckman Institute for Advanced Science and Technology, UIUC. From 1999 to 2000, he was a Research Intern with Microsoft Research, Redmond, WA, USA. In 2001, he joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA, as an Assistant Professor. He was promoted as an Associate Professor in 2007 and a Full Professor in 2012. He is currently a Full Professor of Electrical Engineering and Computer Science with Northwestern University. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction. He received the Robert T. Chien Award by UIUC in 2001 and the NSF CAREER Award in 2003. He serves as an Associate Editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the SPIE Journal of Electronic Imaging, and the IAPR Journal of Machine Vision and Applications. He is a fellow of the IEEE.

Nanning Zheng (SM'94-F'06) graduated in 1975 from the Department of Electrical Engineering, Xi'an Jiaotong University (XJTU), received the ME degree in Information and Control Engineering from Xi'an Jiaotong University in 1981, and a Ph. D. degree in Electrical Engineering from Keio University in 1985. He is currently a Professor and the director with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University. His research interests include computer vision, pattern recognition, computational intelligence, and hardware implementation of intelligent systems. Since 2000, he has been the Chinese representative on the Governing Board of the International Association for Pattern Recognition. He became a member of the Chinese Academy Engineering in 1999. He is a fellow of the IEEE.