Video Object Discovery and Co-Segmentation with Extremely Weak Supervision

Le Wang, *Member, IEEE*, Gang Hua, *Senior Member, IEEE*, Rahul Sukthankar, *Member, IEEE*, Jianru Xue, *Member, IEEE*, Zhenxing Niu, *Member, IEEE*, and Nanning Zheng, *Fellow, IEEE*

Abstract—We present a spatio-temporal energy minimization formulation for simultaneous video object discovery and co-segmentation across multiple videos containing irrelevant frames. Our approach overcomes a limitation that most existing video co-segmentation methods possess, i.e., they perform poorly when dealing with practical videos in which the target objects are not present in many frames. Our formulation incorporates a spatio-temporal auto-context model, which is combined with appearance modeling for superpixel labeling. The superpixel-level labels are propagated to the frame level through a multiple instance boosting algorithm with spatial reasoning, based on which frames containing the target object are identified. Our method only needs to be bootstrapped with the frame-level labels for a few video frames (e.g., usually 1 to 3) to indicate if they contain the target objects or not. Extensive experiments on four datasets validate the efficacy of our proposed method: 1) object segmentation from a single video on the SegTrack dataset, 2) object co-segmentation from multiple videos on a video co-segmentation dataset, and 3) joint object discovery and co-segmentation from multiple videos containing irrelevant frames on the MOViCS dataset and XJTU-Stevens, a new dataset that we introduce in this paper. The proposed method compares favorably with the state-of-the-art in all of these experiments.

Index Terms—Video object discovery, video object co-segmentation, spatio-temporal auto-context model, Spatial-MILBoost

1 Introduction

We address the problem of simultaneously segmenting a common category of objects from two or more videos, which is known as video object co-segmentation. The goal is to label each pixel in a set of videos according to whether it belongs to the unknown common object. Such capacity can be useful for a number of computer vision tasks, such as object centric video summarization, and content-based video retrieval. Compared with object segmentation from a single image, the benefit is that the appearance and/or structure information of the target objects across multiple videos are leveraged for object segmentation in each individual frame.

Several previous methods [1], [2], [3], [4], [5] have attempted to harness such information for video object co-segmentation. However, they all made the assumption that all frames from all videos contain the target object, i.e., all frames are relevant. Moreover, a closer look at the video datasets employed in previous papers reveals that the object instances in different videos are frequently the same object [1], or only exhibit small variations in color, shape, pose, size, and location [2], [3], [4], [5]. These limitations render such methods less applicable to real-world videos, such as those online videos gathered from a search engine in response to a specific query. The common objects in these videos are usually just of the same category, exhibiting dramatic variations in color, size, shape, pose, and viewpoint. Moreover, it is not uncommon for such videos to contain many irrelevant frames where the target objects are not present. This suggests that a practical video object co-segmentation method should also be capable of identifying the frames that contain the objects, i.e., discovering the objects. Fig. 1 illustrates the problem we intend to address.

We present a spatio-temporal energy minimization formulation to simultaneously discover and co-segment the target objects from multiple videos containing irrelevant frames. The flowchart of our method is presented in Fig. 2. Bootstrapped from just a few (often 1 to 3) labeled frames indicating whether they are relevant or not, our method performs a top-down modeling to propagate the frame-level label to the superpixels through a multiple instance boosting algorithm with spatial reasoning, namely Spatial-MILBoost. From bottom up, the labels of the superpixels are jointly determined by a spatio-temporal auto-context model induced from the Spatial-MILBoost algorithm and an appearance model using colors.

The learning of the spatio-temporal auto-context model, cast together with the color based appearance model as the data term, is embedded in a spatio-temporal energy minimization framework for joint object discovery and co-segmentation. Due to the embedded formulation, the learning of the spatio-temporal auto-context model (hence the object discovery), and the minimization of the energy function conducted by min-cut [6], [7] (hence the object co-segmentation), are performed iteratively until convergence. The final

- *L. Wang, J. Xue, and N. Zheng are with the Institute of Artificial Intelligence & Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China. E-mail: {lewang, jrxue, nnzheng}@mail.xjtu.edu.cn.*
- *G. Hua is with the Microsoft Research, Beijing 100080, China. E-mail: ganghua@gmail.com.*
- *R. Sukthankar is with Google Research, New York, NY 10011. E-mail: rahulsukthankar@gmail.com.*
- *Z. Niu is with the School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. E-mail: zxniu@xidian.edu.cn.*

Manuscript received 19 Aug. 2015; revised 24 May 2016; accepted 12 Sept. 2016. Date of publication 25 Oct. 2016; date of current version 12 Sept. 2017. Recommended for acceptance by J. J Corso.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2612187

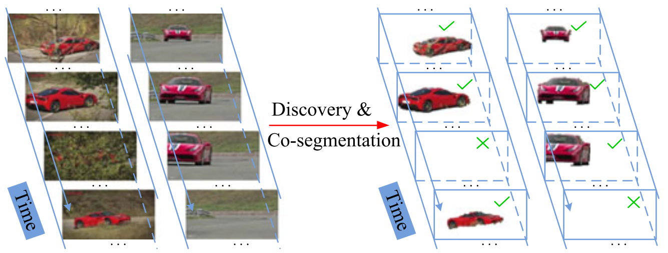


Fig. 1. Problem setting: **Input**—multiple videos capturing a common category of objects. Some of which may contain irrelevant frames. **Output**—a label for each frame indicating if it is relevant, and a detailed pixel labeling of the common object for each relevant frame identified.

output of our method includes a frame-level label for each frame indicating if it contains the target object, and a superpixel-level labeling of the target object for each identified relevant frame.

As a key component of our formulation, our proposed spatio-temporal auto-context model extends the original auto-context model [8] to also capture the temporal context. Our embedded formulation also facilitates learning the model with only weak supervision with frame-level labels using the Spatial-MILBoost algorithm. Spatial-MILBoost allows information to be propagated between the frame level and the superpixel level, and hence facilitates both the discovery and the co-segmentation of the target objects by effectively exploiting the spatio-temporal context across multiple videos.

To summarize, the key contributions of this paper are:

1) We propose a method to address the problem of simultaneous discovery and co-segmentation of a common category of objects from multiple videos containing irrelevant frames.
2) To facilitate both object discovery and co-segmentation, we model the spatio-temporal contextual information across multiple videos by a spatio-temporal auto-context model learned from a Spatial-MILBoost algorithm.
3) To exactly evaluate the proposed method, we collect and release a new 10-category video object co-segmentation and classification dataset with ground truth frame-level labels for all frames and pixel-wise foreground labels for all relevant frames.

We perform extensive studies to evaluate our method in three aspects, and compare with state-of-the-art in terms of both qualitative and quantitative results, including 1) object segmentation from a single video on the SegTrack dataset [9], [10], 2) object co-segmentation from multiple videos on the video co-segmentation dataset [2], [11], [12], and 3) joint object discovery and co-segmentation from multiple videos containing irrelevant frames on the MOViCS dataset [13] and a new 10-category video object co-segmentation and classification dataset collected by ourselves.

Furthermore, to better understand the contributions of different aspects of our proposed method, we implement four variants of our method to conduct extensive ablative studies; we also implement two groups of experiments to evaluate the impacts of different components we leveraged in our method. It is shown that our method compares favorably with the state-of-the-art, and has the ability to simultaneously discover and co-segment the target objects from multiple videos containing irrelevant frames.

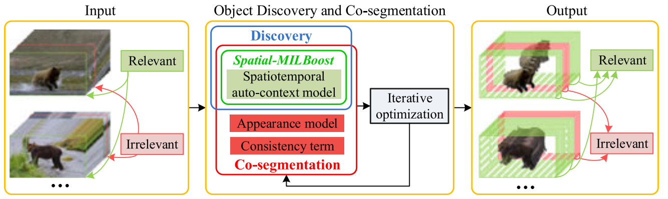


Fig. 2. The flowchart of our video object discovery and co-segmentation method.

This paper is an extension of our conference paper [14]. Compared with it, first of all, this paper provides a more comprehensive and systematic report of our work. Second, the most recent related work after the publication of our conference paper are also added in this paper. Moreover, we provide more details of the problem formulation and implementation. Last but not least, the experimental section is fully reorganized, and more extensive experiments are conducted to validate our method and its variants.

In Section 2, we give a review on related work. In Section 3, we present the problem formulation. In Section 4, we present the optimization procedure. In Section 5, we evaluate the framework on four datasets with detailed discussions. In Section 6, we conclude the paper.

2 Related Work

Since our work addresses the problem of object discovery and co-segmentation from multiple videos, we review related work in video object discovery, video object segmentation and co-segmentation, and image co-segmentation.

2.1 Video Object Discovery

Video object discovery has recently been extensively studied, in both unsupervised [15], [16], [17], [18] or weakly supervised [19], [20] settings. Liu and Chen [15] proposed a latent topic model for unsupervised object discovery in videos by combining Probabilistic Latent Semantic Analysis (PLSA) with Probabilistic Data Association (PDA) filter. Zhao et al. [16] proposed a topic model by incorporating a word co-occurrence prior into Latent Dirichlet Allocation (LDA) for efficient discovery of topical video objects from a set of key frames. Kwak et al. [17] proposed an algorithm to automatically localize the objects as spatio-temporal tubes in an unlabeled video set by combining object discovery and tracking. Yang et al. [18] proposed a method to detect primary objects by integrating the local saliency and global appearance consistency. Liu et al. [19] engaged human in the loop to provide a few labels at the frame level to roughly indicate the main object of interest. Prest et al. [20] proposed a fully automatic method to learn a class-specific object detector from weakly annotated real-world videos.

Tuytelaars et al. [21] surveyed the unsupervised object discovery methods, but with the focus on still images. Wang et al. [22] summarized the abundant literature of visual pattern discovery, and discussed both bottom-up and top-down techniques as well as their diverse applications. In contrast, our video object discovery is achieved by propagating superpixel-level labels to frame level through a Spatial-MILBoost algorithm.

2.2 Video Object Segmentation and Co-Segmentation

Video object segmentation refers to the task of separating the objects from the background in a video, either interactively [9], [23], [24], [25], [26], [27] or automatically [10], [12], [28], [29], [30], [31], [32], [33]. A number of methods focus on finding the object-like proposals for this problem [26], [28], [29], [30], [31], [33]. Several methods track feature points or local regions over frames, and then cluster the resulting tracks based on pairwise [9], [25] or triplet similarity measures [10]. Tang et al. [24] proposed an algorithm for annotating spatio-temporal segments based on video-level labels. Grundmann et al. [12] clustered a video into spatio-temporal consistent supervoxels. Jain and Grauman [27] recently proposed a higher order supervoxel label consistency potential for semi-supervised foreground segmentation. Fragkiadaki et al. [32] segmented moving objects by ranking spatio-temporal segment proposals according to moving objectness. Perazzi et al. [33] employed a fully connected spatio-temporal graph built over object proposals for video segmentation.

Only a few video object co-segmentation methods [1], [2], [3], [4], [5] have been proposed recently to simultaneously segment a common category of objects from two or more videos. They all leveraged the low-level categorized features (i.e., color and texture) shared between multiple videos to achieve object co-segmentation, and thus often encountered difficulties when the objects of the same category in different videos exhibit large variations in color, size, shape, pose, and viewpoint. Moreover, they made the assumption that all frames from all videos should contain the target object, and thus cannot deal with noisy web videos which contain irrelevant frames.

There are also several methods focusing on multi-class video object co-segmentation from multiple videos [13], [34], [35], [36], where the number of object classes and the number of object instances are unknown in each frame and video. Chiu and Fritz [13] proposed a non-parametric algorithm to cluster pixels into different regions by using a global appearance model and a spatio-temporal segmentation prior. However, this method may not be robust to appearance variations caused by pose change of the target objects in different videos. Fu et al. [34] presented a co-selection graph to formulate correspondences between different videos, and extended this framework to handle multiple objects using a multi-state selection graph model. Lou and Gevers [35] employed the appearance, saliency and motion gradient consistency of object proposals to extract the primary objects, but they can only extract the objects of one common category each time. Zhang et al. [36] proposed an algorithm for object co-segmentation by selecting object proposal tracklets that are spatially salient and temporally consistent, and by iteratively extracting weighted groupings of objects with similar shape and appearance. Although it can handle multiple objects, temporary occlusions, and objects going in and out of view, it may encounter difficulties when handling objects with large intra-category variations, such as appearance and shape.

The differences between our work and the above works are that: 1) we address the problem of simultaneously discovering and segmenting the objects of interest from noisy videos, in which many frames do not contain the target objects, 2) we cast the tasks of object discovery and co-segmentation into a unified spatio-temporal energy minimization framework, and 3) we leverage the spatio-temporal contextual information to facilitate both object discovery and co-segmentation of the target objects of the common category from multiple videos.

TABLE 1
Principal Notations

$\mathcal{V}$	A collection of N videos
$\mathcal{L}$	The frame-level labels of $\mathcal{V}$
$\mathcal{B}$	A segmentation of $\mathcal{V}$
V^n	The nth video in $\mathcal{V}$ with N^n frames
L^n	The frame-level labels of V^n
B^n	A segmentation of V^n
f_i^n	The ith frame of V^n with N_i^n superpixels
l_i^n	The label of f_i^n, $l_i^n \in \{0,1\}$, where 1 means that f_i^n is relevant, i.e., f_i^n contains the target object
b_i^n	A segmentation of f_i^n
s_{ij}^n	The jth superpixel in f_i^n
b_{ij}^n	The label of s_{ij}^n, $b_{ij}^n \in \{0,1\}$, where 1 means that s_{ij}^n belongs to the target object

2.3 Image Co-Segmentation

Our work is also related to image co-segmentation [37], [38], [39], [40], [41], [42], [43], [44], [45], where the appearance or structure consistency of the foreground objects across the image collection is exploited to benefit object segmentation. The objective of image co-segmentation is to jointly segment a specific object from two or more images, and it is assumed that all images contain that object. There are also several image co-segmentation methods [46], [47] that further conduct the co-segmentation of multiple objects of multiple categories, in which they assumed that each image should contain at least one object among the multiple categories.

Recently, a few methods have been proposed to conduct the joint discovery and co-segmentation of the objects of a common category from noisy web image collections [48], [49], in which several images do not contain the target objects. In our work, we focus on video object discovery and co-segmentation with noisy video collections, where many frames may not contain the target objects.

3 Problem Formulation

For ease of presentation, we first summarize the main notations in Table 1. Then we present the proposed spatio-temporal energy minimization framework for simultaneous object discovery and co-segmentation across multiple videos, along with details of the spatio-temporal context model and the Spatial-MILBoost algorithm.

Given a set of videos $\mathcal{V}$, our objective is to obtain a frame-level label l_i^n for each frame f_i^n indicating if it is a relevant frame that contains the target objects, and a superpixel-level labeling b_i^n of the target object for each identified relevant frame f_i^n (i.e., $l_i^n = 1$). We cast this problem into a spatio-temporal energy minimization framework. Then, our energy function for simultaneous object discovery and co-segmentation from multiple videos $\mathcal{V}$ becomes

$$
\begin{aligned}
E(\mathcal{B}) = &\sum_{s_{ij}^n \in \mathcal{V}} D_j^{cont}(b_{ij}^n) + \sum_{s_{ij}^n \in V^n} D_j^{col}(b_{ij}^n) \\
&+ \sum_{s_{ij}^n, s_{ik}^n \in \mathcal{N}_j} S_{jk}^{intra}(b_{ij}^n, b_{ik}^n) + \sum_{s_{ij}^n, s_{uk}^n \in \bar{\mathcal{N}}_j} S_{jk}^{inter}(b_{ij}^n, b_{uk}^n), \qquad (1) \\
&n = 1, \ldots, N, i = 1, \ldots, N^n, j = 1, \ldots, N_i^n,
\end{aligned}
$$

where $D_j^{cont}(b_{ij}^n)$ and $D_j^{col}(b_{ij}^n)$ compose the data term, measuring the cost of labeling superpixel s_{ij}^n to be b_{ij}^n from a spatio-temporal auto-context model and a color based appearance model, respectively. The spatio-temporal auto-context model builds a multi-layer Boosting classifier on context features surrounding a superpixel to predict whether it is associated with the target concept or not, where subsequent layer is working on the probability maps from the previous layer, detailed below in Section 3.1. Hence, $D_j^{cont}(b_{ij}^n)$ relies on the discriminative probability maps estimated by a learned spatio-temporal auto-context model, which captures the spatio-temporal contextual information across multiple videos $\mathcal{V}$. Thus, it is video independent. While the appearance model is estimated by capturing the color distributions of the target objects and the backgrounds for each video V^n, and thus is video dependent.

$S_{jk}^{intra}(b_{ij}^n, b_{ik}^n)$ and $S_{jk}^{inter}(b_{ij}^n, b_{uk}^n)$ compose the consistency term, constraining the segmentation labels to be spatially consistent from a color based intra-frame consistency model, and temporally consistent from a spatio-temporal auto-context feature based inter-frame consistency model, respectively. $\mathcal{N}_j$ is the spatial neighborhood of s_{ij}^n in f_i^n. $\bar{\mathcal{N}}_j = \{s^{\leftarrow n}_{ij}, s^{\rightarrow n}_{ij}\}$ is the temporal neighborhood of s_{ij}^n, i.e., its corresponding next superpixel $s^{\rightarrow n}_{ij}$ in f_{i+1}^n and previous superpixel $s^{\leftarrow n}_{ij}$ in f_{i-1}^n. The superpixels are computed by using SLIC [50], due to its superiority in terms of adherence to boundaries, as well as computational and memory efficiency. However, the proposed method is not tied to any specific superpixel method, and one can choose others.

The particular spatio-temporal auto-context model embedded in the energy function is learned through a multiple instance learning algorithm with spatial reasoning (i.e., Spatial-MILBoost), and hence it can propagate information between the frame level and the superpixel level. From top down, the label of frame is propagated to the superpixel level to facilitate the energy minimization for co-segmentation; from bottom up, the labels of superpixels are propagated to the frame level to identify which frame is relevant. Bootstrapped from just a few frame-level labels, the learning of the spatio-temporal auto-context model (hence the object discovery), and the minimization of the energy function conducted by min-cut [6], [7] (hence the object co-segmentation) are performed iteratively until it converges. At each iteration, the spatio-temporal auto-context model, the appearance model, and the consistency term are updated based on the new segmentation $\mathcal{B}$ of $\mathcal{V}$.

We proceed to present the spatio-temporal auto-context model and the Spatial-MILBoost algorithm in Section 3.1, the appearance model in Section 3.2, the consistency term in Section 3.3, and the optimization procedure for object discovery and co-segmentation in Section 4.

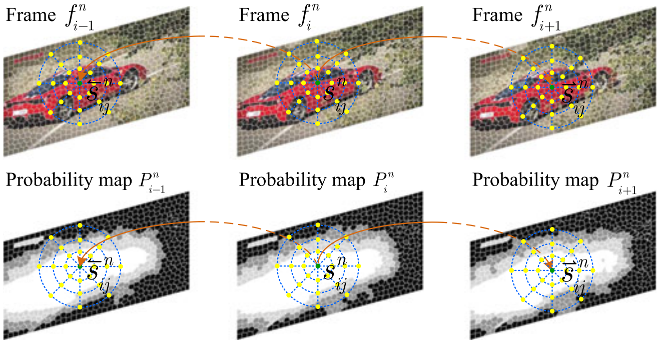

Frame f_{i-1}^n
Frame f_i^n
Frame f_{i+1}^n
$\overleftarrow{s}_{ij}^n$
s_{ij}^n
$\overrightarrow{s}_{ij}^n$
Probability map P_{i-1}^n
Probability map P_i^n
Probability map P_{i+1}^n

Fig. 3. The spatio-temporal auto-context feature.

3.1 Spatio-Temporal Auto-Context Model

We extend the auto-context model originally proposed by Tu [8] and later tailored by Wang et al. [49], [51], [52], [53] for video object discovery and co-segmentation. The original auto-context model builds a multi-layer Boosting classifier on image and context features surrounding a pixel to predict if it is associated with the target concept, where subsequent layer is working on the probability maps from the previous layer. In previous works, it just modeled the spatial contextual information, either from a single image [51], [53], or a set of labeled [8] or unlabeled [49], [52] images. Here, we extend it to capture both the spatial and temporal contextual information across multiple videos, and the extended model operates on superpixels instead of pixels.

Spatio-Temporal Auto-Context Feature. Let $\mathbf{c}_{ij}^n$ denote the context feature of superpixel s_{ij}^n, $P^n \in \mathcal{P}$ the probability map set for video V^n, P_i^n the probability map for frame f_i^n, p_{ij}^n the probability value of superpixel s_{ij}^n. The sampling structure of the spatio-temporal auto-context model on the discriminative probability maps are illustrated in Fig. 3. $\mathbf{c}_{ij}^n$ consists of a previous-frame part, a current-frame part and a next-frame part as

$$
\mathbf{c}_{ij}^n = \{\{\overleftarrow{p}_{ij}^n(k)\}, \{p_{ij}^n(k)\}, \{\overrightarrow{p}_{ij}^n(k)\}\}_{k=1}^{N_c}, \qquad (2)
$$

where $p_{ij}^n(k)$, $p^{\leftarrow n}_{ij}(k)$ and $p^{\rightarrow n}_{ij}(k)$ are the probability values of the kth point on the sampling structure centered at superpixel s_{ij}^n in P_i^n, its corresponding previous superpixel $s^{\leftarrow n}_{ij}$ in P_{i-1}^n, and its corresponding next superpixel $s^{\rightarrow n}_{ij}$ in P_{i+1}^n, respectively. N_c is the number of sampled points on the sampling structure for the current superpixel in each frame, and it is set to be 41 in our experiments. Here, we find the corresponding previous and next superpixels of current superpixel between neighboring frames using optical flow [54], due to its high accuracy and low time consumption. If the number of pixels in the intersection between a superpixel in the current frame and its corresponding superpixel in neighboring frames, identified from the optical flow vector displacements of the current superpixel, is greater than half of the number of pixels in the current superpixel, it is selected as the temporal neighbor.

Update the Spatio-Temporal Auto-Context Classifier. In the first round of the iterative learning of the spatio-temporal auto-context model, the training set is built on multiple videos $\mathcal{V}$ with a few manually annotated frame-level labels as

$$\begin{aligned}\mathbf{S}_1 = \{\{\mathbf{C}_{i'}^n(\alpha), l_{i'}^n(\alpha)\} | n = 1, \ldots, N; i' = 1', \ldots, N^{n\prime}; \\ \alpha = 0, 1\},\end{aligned} \tag{3}$$

where i' is the index of frame $f_{i'}^n$ that was manually labeled by the user as relevant (i.e., $l_{i'}^n = 1$) or irrelevant (i.e., $l_{i'}^n = 0$). $N^{n\prime}$ is the number of labeled frames in video V^n, and it is set to be 1 to 3 in our experiments. $\mathbf{C}_{i'}^n = \{\mathbf{c}_{i'j}^n\}_{j=1}^{N_{i'}^n}$ are the context features of superpixels in $f_{i'}^n$, and $\mathbf{C}_{i'}^n(\alpha)$ are the context features in the object (i.e., $\alpha = 1$) or background (i.e., $\alpha = 0$) of $f_{i'}^n$. We treat $\mathbf{C}_{i'}^n(\alpha)$ as a *bag*, and $\mathbf{c}_{i'j}^n$ as an *instance*. $l_{i'}^n(\alpha)$ is the label of bag $\mathbf{C}_{i'}^n(\alpha)$, and it equals to 1 when both $l_{i'}^n$ and α equal to 1, and 0 otherwise. In other words, we treat the objects of the relevant frames as positive bags, the backgrounds of the relevant frames and both the objects and backgrounds of the irrelevant frames as negative bags. The initial segmentations $\mathcal{B}$ for $\mathcal{V}$ are obtained by using an objectness measure [55] and a saliency measure [56], and the probability maps $\mathcal{P}$ for $\mathcal{V}$ are initialized by averaging the scores returned by objectness and saliency.

Then, the first spatio-temporal auto-context classifier $H(\cdot)$ is learned on $\mathbf{S}_1$ using a Spatial-MILBoost algorithm, detailed immediately below. We proceed to use the learned classifier to classify all the context features of the objects and backgrounds of all frames in $\mathcal{V}$, and obtain the new probability map set $\mathcal{P}$ for $\mathcal{V}$. This way, the spatio-temporal contextual information extracted from a few frames of $\mathcal{V}$ are leveraged to help estimating the probability of each of the superpixels in $\mathcal{V}$ belonging to the target object. The new probability of superpixel s_{ij}^n being positive is updated by the learned classifier as

$$p_{ij}^n = \frac{1}{1 + \exp(-H(\mathbf{c}_{ij}^n))}. \tag{4}$$

The data term based on the spatio-temporal auto-context model in Eq. (1) is defined as

$$D_j^{cont}(b_{ij}^n) = -\log p_{ij}^n. \tag{5}$$

The probability of the object or background (*bag*) of frame f_i^n being positive is a Noisy-OR defined as

$$p_i^n(\alpha) = 1 - \prod_{j=1}^{N_i^n(\alpha)} (1 - p_{ij}^n), \tag{6}$$

where $N_i^n(\alpha)$ denotes the number of superpixels (*instances*) in the object or background (*bag*) of frame f_i^n. In this way, the trained spatio-temporal auto-context classifier can propagate superpixel-level labels (indicating if the superpixels belong to the target objects) to the object-level label (indicating if it contains the target object).

From the second round of the iterative learning process, we update the training set built on all frames of $\mathcal{V}$ as

$$\begin{aligned}\mathbf{S}_2 = \{\{\mathbf{C}_i^n(\alpha), l_i^n(\alpha)\} | n = 1, \ldots, N; i = 1, \ldots, N^n; \\ \alpha = 0, 1\},\end{aligned} \tag{7}$$

and learn a new spatio-temporal auto-context classifier on the updated context features, which are based on the discriminative probability map set $\mathcal{P}$ obtained from the previous iteration. Then, the new $\mathcal{P}$ for $\mathcal{V}$ are computed by the new spatio-temporal auto-context classifier. This process will iterate until convergence, where $\mathcal{P}$ no longer changes. Indeed, the spatio-temporal auto-context model is alternatively updated with the iterative co-segmentation of $\mathcal{V}$, i.e., the iterative minimization of the energy in Eq. (1).

Since the training set built in each iteration consists of the context features of superpixels in the form of bags (i.e., objects and backgrounds of the relevant and irrelevant frames from multiple videos $\mathcal{V}$), the spatio-temporal auto-context classifier learned on it naturally captures both the spatial and temporal contextual information from $\mathcal{V}$ to predict the probability of a superpixel belonging to the target object. Thus, the spatio-temporal auto-context model benefits both object discovery and object co-segmentation.

Algorithm 1. Spatial-MILBoost-Training

```
Input: Training set $\{\mathbf{x}_i, l_i\}_{i=1}^N$ of $N$ bags, where each bag $\mathbf{x}_i = \{x_{ij}\}_{j=1}^{N_i}$ containing $N_i$ instances, the bag label $l_i \in \{0, 1\}$.
  1) Initialize the instance weights $w_{ij} = 2(l_i - 0.5)$ and the instance classifier $H = 0$
  2) Initialize estimated margins $\{\hat{y}_{ij}\}_{i,j=1}^{N,N_i}$ to 0
  3) For $t = 1, \ldots, T$
       a. Set $\bar{x}_{ij} = \{\hat{y}_{ik} | x_{ik} \in \mathrm{Nbr}(x_{ij})\}$
       b. Train weak data classifier $h_t^d$ on the data $\{x_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as
            $$h_t^d(x_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(x_{ij}) w_{ij}$$
       c. Train weak spatial classifier $h_t^s$ on the data $\{\bar{x}_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as
            $$h_t^s(\bar{x}_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(\bar{x}_{ij}) w_{ij}$$
       d. Set $\begin{cases} \epsilon^d = \sum_{i,j} \omega_{ij} |h_t^d(x_{ij}) - l_i| \\ \epsilon^s = \sum_{i,j} \omega_{ij} |h_t^s(\bar{x}_{ij}) - l_i| \end{cases}$
       e. Set $h_t(x_{ij}) = \begin{cases} h_t^d(x_{ij}) & \text{if} \quad \epsilon^d < \epsilon^s \\ h_t^s(\bar{x}_{ij}) & \text{otherwise} \end{cases}$
       f. Find $\lambda_t$ via line search to minimize likelihood $L(H) = \prod_i (q_i)^{l_i} (1 - q_i)^{(1-l_i)}$ as $\lambda_t = \arg\max_\lambda L(H + \lambda h_t)$
       g. Update margins $\hat{y}_{ij}$ to be $\hat{y}_{ij} = H(x_{ij}) = \hat{y}_{ij} + \lambda_t h_t(x_{ij})$
       h. Compute the instance probability $q_{ij} = \frac{1}{1+\exp(-\hat{y}_{ij})}$
       i. Compute the bag probability $q_i = 1 - \prod_{j=1}^{N_i} (1 - q_{ij})$
       j. Update the instance weights $w_{ij} = \frac{\partial \log L(H)}{\partial y_{ij}} = \frac{l_i - q_i}{q_i} q_{ij}$
Output: Instance classifier $H(x_{ij}) = \sum_{t=1}^T \lambda_t h_t(x_{ij})$.
```

Spatial-MILBoost Algorithm. The training and testing details of Spatial-MILBoost are presented in Algorithms 1 and 2, respectively. Compared to the original MILBoost algorithm [57], we incorporate the spatial information between the neighboring superpixels [58] into the multiple instance boosting algorithm [19], [57] to infer whether the superpixel is positive or not, and name this algorithm Spatial-MILBoost.

To present the algorithm in a more general sense, we use $\mathbf{x}_i$, l_i and $x_{ij} \in \mathbf{x}_i$ instead of $\mathbf{C}_i^n(\alpha)$, $l_i^n(\alpha)$ and $\mathbf{c}_{ij}^n \in \mathbf{C}_i^n(\alpha)$ to denote the *bag*, its *label* and its *instance*, respectively.

Algorithm 2. Spatial-MILBoost-Testing

```
Input: Testing set $\{x_{ij}\}_{i,j=1}^{N,N_i}$, and the instance classifier $H(\cdot)$.
    1)  Initialize estimated margins $\{\hat{y}_{ij}\}_{i,j=1}^{N,N_i}$ to 0
    2)  For $t = 1, \ldots, T$
        a.  Set $\bar{x}_{ij} = \{\hat{y}_{ik} | x_{ik} \in \mathrm{Nbr}(x_{ij})\}$
        b.  Update margins $\hat{y}_{ij}$ to be $\hat{y}_{ij} = \hat{y}_{ij} + \lambda_t h_t(x_{ij})$
Output: Labels $\{\hat{y}_{ij}\}_{i,j=1}^{N,N_i}$.
```

The score of the instance x_{ij} is $y_{ij} = H(x_{ij})$, where $H(x_{ij}) = \sum_{t=1}^{T} \lambda_t h_t(x_{ij})$ is a weighted sum of weak classifiers. The probability of the instance x_{ij} being positive is defined as a standard logistic function,

$$q_{ij} = \frac{1}{1 + \exp(-y_{ij})}. \tag{8}$$

The probability of the bag $\mathbf{x}_i$ being positive is a Noisy-OR,

$$q_i = 1 - \prod_{j=1}^{N_i} (1 - q_{ij}). \tag{9}$$

The goal now is to estimate λ_t and h_t, so q_{ij} approaches its true value. The likelihood assigned to a set of training bags is $L(H) = \prod_i (q_i)^{l_i} (1 - q_i)^{(1 - l_i)}$, and is maximum when $q_i = l_i$, where $l_i \in \{0, 1\}$ is the label of bag $\mathbf{x}_i$. To find an instance classifier that maximizes the likelihood, we compute the derivative of the log-likelihood with respect to y_{ij} as $\frac{\partial \log L(H)}{\partial y_{ij}} = w_{ij} = \frac{l_i - q_i}{q_i} q_{ij}$.

In each round t of gradient descent, one solves the optimal weak *instance* classifier $h_t(x_{ij})$. Here, we train a weak *data* classifier on the data $\{x_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as $h_t^d(x_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(x_{ij}) w_{ij}$. Meanwhile, we train a weak *spatial* classifier on the data $\{\bar{x}_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as $h_t^s(\bar{x}_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(\bar{x}_{ij}) w_{ij}$, where $\bar{x}_{ij} = \{\hat{y}_{ik} | x_{ik} \in \mathrm{Nbr}(x_{ij})\}$ are the predicted labels of the neighbors $\mathrm{Nbr}(x_{ij})$ of the current instance x_{ij}. The classifier with lower training error is selected as the weak *instance* classifier $h_t(x_{ij})$,

$$h_t(x_{ij}) = \begin{cases} h_t^d(x_{ij}) & \text{if} \quad \epsilon^d < \epsilon^s \\ h_t^s(\bar{x}_{ij}) & \text{otherwise} \end{cases}, \tag{10}$$

where $\epsilon^d = \sum_{i,j} \omega_{ij} |h_t^d(x_{ij}) - l_i|$ and $\epsilon^s = \sum_{i,j} \omega_{ij} |h_t^s(\bar{x}_{ij}) - l_i|$ are the training errors of $h_t^d(x_{ij})$ and $h_t^s(\bar{x}_{ij})$, respectively. This is the major difference between the proposed Spatial-MILBoost algorithm and the traditional MILBoost algorithm [19], [57].

The parameter λ_t is determined using a line search as $\lambda_t = \arg\max_\lambda L(H + \lambda h_t)$. Then, the instance classifier $H(\cdot)$ is updated by $H(\cdot) \leftarrow H(\cdot) + \lambda_t h_t(\cdot)$.

3.2 Appearance Model

Since the appearance of the object instances (also the backgrounds) are similar in color within each video V^n, while exhibiting large variations across multiple videos $\mathcal{V}$, we independently learn the color distributions of the target objects and the backgrounds for each video V^n.

In detail, with a segmentation $\mathcal{B}$ for $\mathcal{V}$, we estimate two color Gaussian Mixture Models (GMMs) for the target objects and the backgrounds of each video V^n, denoted as $\mathbf{h}_1^n$ and $\mathbf{h}_0^n$, respectively. The corresponding data term based on the appearance model in Eq. (1) is defined as

$$D_j^{col}(b_{ij}^n) = -\log \mathbf{h}_{b_{ij}^n}^n (s_{ij}^n), \tag{11}$$

where $D_j^{col}(b_{ij}^n)$ measures the contribution of labeling s_{ij}^n to be b_{ij}^n, based on the appearance model learned from V^n.

3.3 Consistency Term

The consistency term is composed of an intra-frame consistency model and an inter-frame consistency model, and is leveraged to constrain the segmentation labels to be both spatially and temporally consistent.

Intra-Frame Consistency Model. The intra-frame consistency model encourages the spatially adjacent superpixels in the same frame to have the same label. As the spatially adjacent superpixels in the same frame either have similar color or distinct color contrast, we adopt the well-known standard contrast-dependent function [29], [36] to constrain the labels of spatially adjacent superpixels with similar color to be consistent. In Eq. (1), the consistency term computed between spatially adjacent superpixels s_{ij}^n and s_{ik}^n in frame f_i^n of video V^n is defined as

$$S_{jk}^{intra}(b_{ij}^n, b_{ik}^n) = \delta(b_{ij}^n, b_{ik}^n) \exp(-||\mathbf{I}_{ij}^n - \mathbf{I}_{ik}^n||_2^2), \tag{12}$$

where $\mathbf{I}$ is the color vector of the superpixel. b_{ij}^n and b_{ik}^n are the segmentation labels of s_{ij}^n and s_{ik}^n, respectively. $\delta(\cdot)$ is an indicator variable, which is 1 when $b_{ij}^n \neq b_{ik}^n$, and 0 otherwise.

Inter-Frame Consistency Model. The inter-frame consistency model encourages the temporally adjacent superpixels in consecutive frames to have the same label. Since there often exist large variations of motion, shape and lighting between temporally adjacent superpixels from consecutive frames, resulting in large appearance differences between them, the spatial-temporal auto-context feature capturing both spatial and temporal contextual information across multiple videos has better invariance against these variations. Thus, we use a L_1 distance based function to assign the same label to temporally adjacent superpixels that have similar spatial-temporal auto-context features. In Eq. (1), the consistency term computed between temporally adjacent superpixels s_{ij}^n and s_{uk}^n in consecutive frames of video V^n is defined as

$$S_{jk}^{inter}(b_{ij}^n, b_{uk}^n) = \delta(b_{ij}^n, b_{uk}^n) \exp(-||\mathbf{c}_{ij}^n - \mathbf{c}_{uk}^n||_1), \tag{13}$$

where $\mathbf{c}$ is the context vector of the superpixel. b_{ij}^n and b_{uk}^n are the segmentation labels of s_{ij}^n and s_{uk}^n, respectively. s_{uk}^n is the temporal neighbor of s_{ij}^n, i.e., its corresponding next superpixel $s^{\rightarrow n}_{ij}$ in f_{i+1}^n or previous superpixel $s^{\leftarrow n}_{ij}$ in f_{i-1}^n.

4 Optimization

The proposed approach is bootstrapped from a few manually annotated relevant and irrelevant frames (e.g., usually 1 to 3), and an objectness measure [55] and a saliency measure [56] to initialize the segmentation $\mathcal{B}$ and the discriminative probability map set $\mathcal{P}$ of $\mathcal{V}$. We proceed to start the first round learning of the spatio-temporal auto-context model, and propagate the superpixel labels estimated from the learned auto-context classier $H(\cdot)$ to frame-level labels $\mathcal{L}$ of $\mathcal{V}$ through the Spatial-MILBoost algorithm. We then update the spatio-temporal auto-context model together with the appearance model and consistency term, and perform energy minimization on Eq. (1) by using min-cut [6], [7] to obtain an updated segmentation $\mathcal{B}$ of $\mathcal{V}$.

The learning of the spatio-temporal auto-context model (the object discovery), and the minimization of the energy function in Eq. (1) (the object co-segmentation) are iteratively performed until convergence, which returns not only a frame-level label $\mathcal{L}$ of $\mathcal{V}$ and a segmentation $\mathcal{B}$ of $\mathcal{V}$, but also a spatio-temporal auto-context model.

4.1 Object Discovery

Object discovery is to identify the relevant frames containing the target objects from multiple videos $\mathcal{V}$. As we obtained a current frame-level label $\mathcal{L}$, segmentation $\mathcal{B}$, and discriminative probability map set $\mathcal{P}$ estimated by the spatio-temporal auto-context model from the previous iteration, the probability of frame f_i^n containing the target object is updated as

$$p_i^n = 1 - (1 - p_i^n(1))(1 - p_i^n(0)), \tag{14}$$

which is a Noisy-OR on $p_i^n(1)$ and $p_i^n(0)$ calculated with Eqs. (6) and (4), indicating the probabilities of the current segmented object and background of f_i^n being positive, respectively. It is consistent with the practical situations that 1) if both the segmented object and background of f_i^n do not contain any part of the target object, f_i^n will certainly not contain the target object, and 2) if at least one of the segmented object and background of f_i^n contains some parts of the target object, f_i^n will contain the target object. This way, no matter the current segmentation of f_i^n is accurate or not, as long as the segmented object or background contains parts of the target object, f_i^n will be predicted to contain the target object (hence the object discovery).

Then, the label l_i^n indicating whether f_i^n is relevant can be predicted by binarizing p_i^n,

$$l_i^n = \begin{cases} 1 & \text{if} \quad p_i^n \geq \lambda \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

where the threshold λ is fixed to be 0.45 empirically. l_i^n equals to 1 when f_i^n is relevant, and equals to 0 otherwise. This way, the label l_i^n can be inferred from the probabilities of the segmented object and background inside f_i^n indicating if they contain the target object or not; while the probability of the segmented object (or background) can be inferred from the probabilities of the superpixels inside it denoting if they belong to the target object.

4.2 Object Co-Segmentation

The video object co-segmentation is to simultaneously find a superpixel-level labeling $\mathcal{B}$ for the relevant frames identified from $\mathcal{V}$. As we obtain a current frame-level label $\mathcal{L}$, segmentation $\mathcal{B}$ and discriminative probability map set $\mathcal{P}$ estimated by the spatio-temporal auto-context model from the previous iteration, we update the video independent spatio-temporal auto-context model. Naturally, the spatio-temporal contextual information across $\mathcal{V}$ are leveraged for the segmentation of each frame. The new segmentation B^n of each video V^n also serves to update the corresponding video dependent appearance model and consistency term. We then minimize the energy function in Eq. (1) using min-cut [6], [7] to obtain the new segmentation $\mathcal{B}$ of $\mathcal{V}$.

5 Experiments and Discussions

5.1 Experimental Setup

Evaluation Datasets. We conduct extensive experiments to evaluate our method in 3 cases, i.e., 1) *object segmentation* from a single video on the SegTrack dataset [9], [10], 2) *object co-segmentation* from multiple videos only containing relevant frames on the video co-segmentation dataset [2], [11], [12], and 3) *joint object discovery and co-segmentation* from multiple videos containing irrelevant frames on the MOViCS dataset [13] and a new 10-category video object co-segmentation and classification dataset collected by ourselves. As there are indeed specific assumptions for each of the above three tasks, we use different initialization to bootstrap our method according to the specific assumptions, and accurately evaluate our method on the corresponding benchmark dataset or its subset.

Evaluation Metric. We employ the intersection-over-union (IoU) score [13] for the evaluation of segmentation performance, which is one of the most widely adopted metric to evaluate the performance of image/video segmentation methods. It 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.

Baselines. To fully and exactly evaluate our proposed method, we compare our method with 12 state-of-the-art methods, including five single video segmentation methods (VS [12], VOS [28], VST [10], SVOS [29], and FOS [30]), three video object co-segmentation methods (VC [2], VOC [3], and VCA [4]), one multi-class image co-segmentation method (MIC [47]), and 3 multi-class video co-segmentation methods (MVC [13], MFVC [34], and MVOC [36]). They are

- VS [12], a video segmentation method which achieves hierarchical segmentations of a video by using a hierarchical graph-based algorithm.
- VOS [28], a video object segmentation method which automatically discovers key segments and groups them to predict the foreground object in a video.
- VST [10], a video segmentation method which is achieved by simultaneously tracking multiple holistic figure-ground segments on each frame.
- SVOS [29], a video object segmentation method which segments the primary object from a single video in a layered directed acyclic graph framework.
- FOS [30], a video object segmentation method which separates the target objects from a video based on a rapid estimate of which pixels are inside the object.

TABLE 2
The Average IoU Scores of Our Methods and Five Competing Single Video Segmentation Methods on Eight Videos That Contain Only One Object on the SegTrack Dataset

Video	VS [12]	VOS [28]	SVOS [29]	FOS [30]	VST [10]	woC	wSC	wMIL	VODC (wSV)
birdfall	57	49	**71**	59	63	52	65	68	**70**
girl	32	88	82	73	89	63	88	90	**91**
parachute	69	**96**	94	91	93	76	90	91	92
frog	67	75	74	77	72	65	77	81	**83**
worm	35	**84**	60	74	83	57	74	78	80
soldier	67	67	60	69	84	55	82	84	**85**
monkey	62	79	62	65	85	71	86	90	**90**
bird of paradise	87	92	-	66	94	69	89	92	**95**
Avg.	60	79	72	72	83	64	81	84	**86**

Higher values are better.

- VC [2], a video co-segmentation method which gathers information from multiple videos to jointly separate the foreground object from the background.
- VOC [3], a video object co-segmentation method which is realized by subspace clustering and a subsequent quadratic pseudo-boolean optimization.
- VCA [4], a video co-segmentation method for common action extraction by using dense trajectories.
- MIC [47], a multi-class image co-segmentation method which jointly segments a large number of images into regions of multiple classes.
- MVC [13], a multi-class video co-segmentation method which produces a segmentation of multiple classes from multiple videos by formulating a non-parametric bayesian model across multiple videos.
- MFVC [34], an object-based multiple foreground video co-segmentation method which can handle multiple foreground co-segmentation with a multi-state selection graph model.
- MVOC [36], a 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.

Ablative Studies. To better understand the contributions of different components of our method, we also implement four variants of our full video object discovery and co-segmentation method (VODC) to perform extensive ablative studies. They are

- woC, a variant of VODC without the spatio-temporal auto-context model, which becomes an object segmentation method with an appearance model learned on a video and a consistency model computed on a frame. Thus, we can only evaluate its object segmentation performance from a single video.
- wSV, a variant of VODC by learning the spatio-temporal auto-context model from a single video instead of multiple videos, which becomes a video segmentation method.
- wSC, a variant of VODC by replacing the spatio-temporal auto-context model with spatial auto-context model, which is also a video object discovery and co-segmentation method.

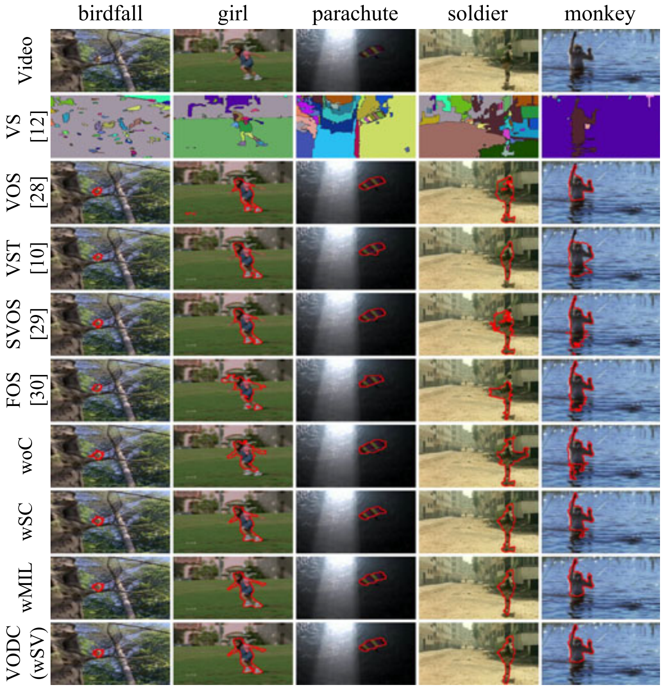


Fig. 4. Some visual example results of our methods and five state-of-the-art single video segmentation methods on eight videos that contain only one object on the SegTrack dataset.

- wMIL, a variant of VODC by replacing the proposed Spatial-MILBoost algorithm with MILBoost [57], which is also a video object discovery and co-segmentation method.

5.2 Object Segmentation from a Single Video

We first evaluate the performance of object segmentation from a single video on the SegTrack dataset [9], [10]. The SegTrack v1 dataset [9] consists of six videos, in which three videos contain only one object, and the others contain multiple adjacent/interacting objects. The SegTrack v2 dataset [10] extends the SegTrack v1 dataset to contain eight additional videos, in which five videos contain only one object, and the rest of them contain multiple ones. The videos have full pixel-level annotations on the objects at each frame.

As our method focuses on single object segmentation, we evaluate it on the eight videos that contain only one object, and compare with five single video segmentation methods (VS [12], VOS [28], VST [10], SVOS [29], and FOS [30]). By initializing all frames as relevant, we segment each video using our full methods (VODC) and four variants of it (woC, wSV, wSC, and wMIL), respectively. Here, VODC equals to wSV, as it learns the spatio-temporal auto-context model on a single video. The average IoU scores are presented in Table 2, and some example results are given in Fig. 4.

The results show that, 1) VODC has the ability to segment the objects with certain variations in appearance (bird of paradise), shape (girl and frog), size (soldier), and backgrounds (parachute). It is superior among all other methods on the five videos, but has encountered some difficulties when the objects are too small (birdfall), or the background are too complex (birdfall and parachute), or the boundaries between the objects and background are too weak (worm).

TABLE 3
The Average IoU Scores of Our Methods and 12 State-of-the-Art Methods on the VCoSeg Dataset

Algorithm	chachacha	ice skater	kite surfer	Avg.
VS [12]	55.1	51.2	38.3	48.2
VOS [28]	54.2	83.3	68.7	68.7
VST [10]	74.8	75.0	44.9	64.9
SVOS [29]	32.1	57.4	36.8	42.1
FOS [30]	65.1	83.2	43.0	63.8
VC [2]	57.6	47.2	14.1	39.6
VOC [3]	70.4	81.5	62.2	71.4
VCA [4]+[59]	80.1	72.4	69.5	74.0
MIC [47]	36.3	44.8	14.5	31.9
MVC [13]	59.5	67.8	36.6	54.6
MFVC [34]	73.1	84.1	72.8	76.7
MVOC [36]	55.5	67.2	50.6	57.8
wSV	78.2	81.7	76.3	78.7
wSC	81.8	85.6	80.7	82.7
wMIL	82.1	87.5	82.9	84.2
VODC	**83.3**	**89.9**	**84.1**	**85.8**

Higher values are better.

2) woC is only better than VS [12], as it simply leverages the color information and lacks the consistency constraint between adjacent frames. 3) wSC performs worse than three methods, since it only uses a spatial auto-context model without employing the temporal contextual information. 4) wMIL using the original MILBoosting outperforms all other methods except VODC, because it does not consider the spatial reasoning while predicting the segmentation label.

5.3 Object Co-Segmentation from Multiple Videos

We then evaluate the performance of object co-segmentation from multiple videos of our method on the video co-segmentation (VCoSeg) dataset [2], [11], [12], which consists of three categories of videos, i.e., four videos of the chachacha category from [11], three videos of the kite surfer category and three videos of the ice skater category both from [2] and [12].

As all frames of all videos in each category contain the target objects, we simultaneously segment the videos of each category using our methods (VODC, wSV, wSC, and wMIL) by treating all frames of each category as relevant, and compare with 12 state-of-the-art methods (i.e., five single video segmentation methods [10], [12], [28], [29], [30], three video object co-segmentation methods [2], [3], [4], one multi-class image co-segmentation method [47], and three multi-class video co-segmentation methods [13], [34], [36]). Since VCA [4] produces the results in terms of dense trajectories, they use the method in [59] to turn their trajectory labels into pixel labels for comparison. For the five single video segmentation methods [10], [12], [28], [29], [30], each video is individually segmented by them. For MIC [47], all frames of each category are treated as a set of individual images.

The average IoU scores and some example results of our methods and the above methods are presented in Table 3 and Fig. 5, respectively. They show that, 1) our full method (VODC) is better than the five single video segmentation methods [10], [12], [28], [29], [30]. This is because that VODC can leverage the information of the target object across multiple videos for co-segmentation, but the single video segmentation methods can only use the appearance and motion cues from one video. 2) VODC, which leverages a spatio-temporal auto-context model, outperforms the compared video object co-segmentation methods [2], [3], [4] that only utilize the low-level appearance and motion cues across multiple videos. Moreover, VODC is not limited to the initial segmentation generated by combining the objectness and saliency measures that VC [2] is sensitive to. 3) VODC performs better than the image co-segmentation method [47], as MIC [47] lacks the consideration of temporal consistency between frames. 4) VODC achieves better performances than the three multi-class video co-segmentation methods [13], [34], [36]. This is due to that MVC [13] often separates some parts instead of the integral foreground object, and MFVC [34] and MVOC [36] often encounter difficulties when selecting the accurate object proposal for each frame. In addition, the reasons why wSC and wMIL are poorer than VODC have already been discussed in Section 5.2.

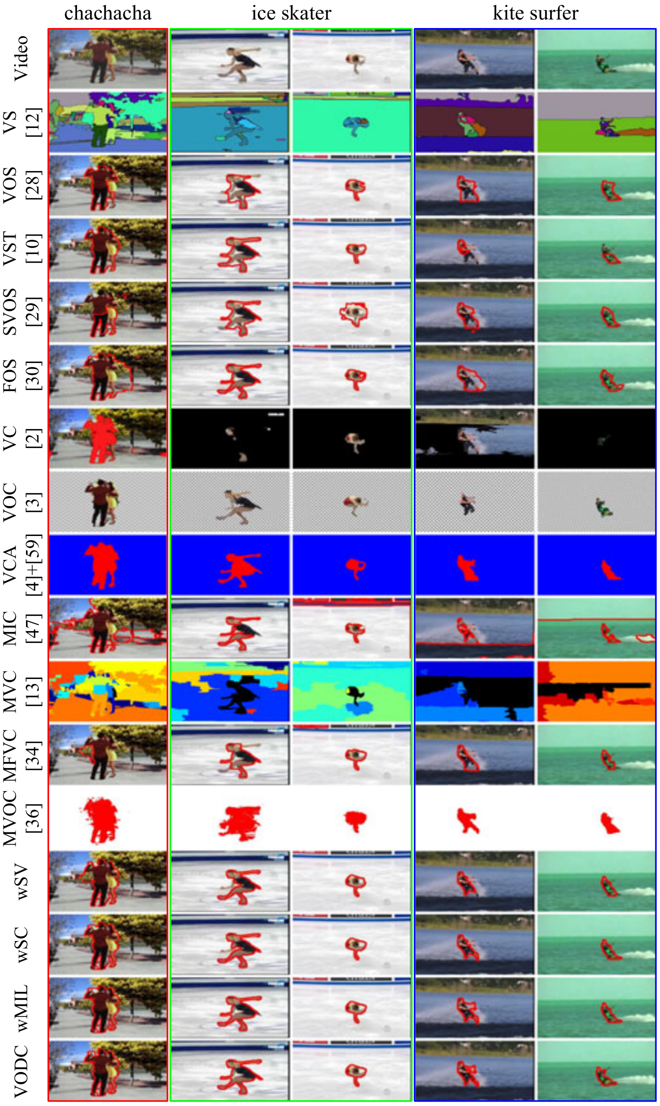


Fig. 5. Some visual example results of our methods and 12 state-of-the-art methods on the VCoSeg dataset.

5.4 Joint Object Discovery and Co-Segmentation from Multiple Videos

We further evaluate the performance of joint object discovery and co-segmentation of our method from multiple

TABLE 4
The Average IoU Scores of Our Methods and Four State-of-the-Art Methods on the MOViCS Dataset

Algorithm	chicken	giraffe	lion	Tiger	Avg.
MIC [47]	46.7	41.9	59.6	42.4	47.7
MVC [13]	70.5	56.4	67.2	53.0	61.8
MFVC [34]	87.2	66.8	82.8	71.4	77.1
MVOC [36]	83.9	58.1	80.7	53.1	69.0
wSC	86.5	67.3	81.3	77.4	78.1
wMIL	89.2	69.7	83.5	78.7	80.3
VODC	**91.5**	**71.1**	**86.2**	**80.6**	**82.4**

Higher values are better.

videos on the MOViCS dataset [13] and a new 10-category video object co-segmentation and classification dataset.

1) Evaluation on MOViCS Dataset

The MOViCS dataset [13] includes four groups of videos which has 11 videos in total. Each video group contains one or two objects, and five frames of each video have pixelwise ground truth.

In the experiments, we provide each video with one relevant or irrelevant frame to bootstrap our method. For the videos of chicken, giraffe and tiger, as all frames of all videos contain one primary object (i.e., the chicken, giraffe and tiger), we randomly select one frame from each video as the relevant one. As the initial segmentations of all frames cover the primary object, our method can co-segment the primary object from the video set without the interference of the other object (i.e., the turtle or elephant). For the four videos of lion, all frames of two videos contains the lion, some frames of one video do not contain the lion because of occlusion, and all frames of one video do not contain the lion at all. We randomly select one frame from each video as relevant or irrelevant according to its ground truth label. As the initial segmentations of all the relevant frames cover the lion, our method can identify all the relevant frames from the irrelevant ones, and meanwhile co-segment the lion out.

Since the videos of each group (except the tiger group) in the MOViCS dataset [13] contain two objects, we compare our methods (VODC, wSC, and wMIL) with one multi-class image co-segmentation method (MIC [47]) and three multi-class video co-segmentation methods (MVC [13], MFVC [34], and MVOC [36]). The average IoU scores and some example results of them are presented in Table 4 and Fig. 6, respectively.

MIC [47] does not perform well because it do not employ motion cue to enhance the temporal smoothness of the target object. Since MVC [13] relies on pixel-level features, the segmentation results tend to be over-segmented. Moreover, it may link different objects from different videos, as shown in the third tiger video in Fig. 6. MFVC [34] formulates video co-segmentation as a co-selection graph to connect object proposals in multiple videos, and thus its segmentation results are highly dependent on the method of generating object proposals. MVOC [36] makes a strong assumption that the object proposal tracklets for the same class of objects should have similar appearance both within a video and across videos. Thus, it may assign incorrect class labeling for the object of the same class, as illustrated in the second video of giraffe and the third video of tiger.

As the above results shown, our methods outperform all the compared methods, and the average improvements of wSC, wMIL, and VODC are more than 1, 3, and 5 percent, respectively. This strongly demonstrates the efficacy of the spatio-temporal auto-context model which captures the categorized contextual information across multiple videos, and the Spatial-MILBoost algorithm which considers the spatial reasoning while predicting the segmentation label. Moreover, our methods identifies all the relevant frames from the irrelevant ones, while most of the above methods do not assign correct class labeling for most of the frames (MIC [47], MVC [13], and MVOC [36]).

2) Evaluation on XJTU-Stevens Video Co-Segmentation and Classification Dataset

The existing video object co-segmentation datasets [1], [2], [3], [13], [34], [35], [36] have the following limitations: 1) they only contain the relevant frames or a small number of irrelevant frames, 2) the categories, the videos of each

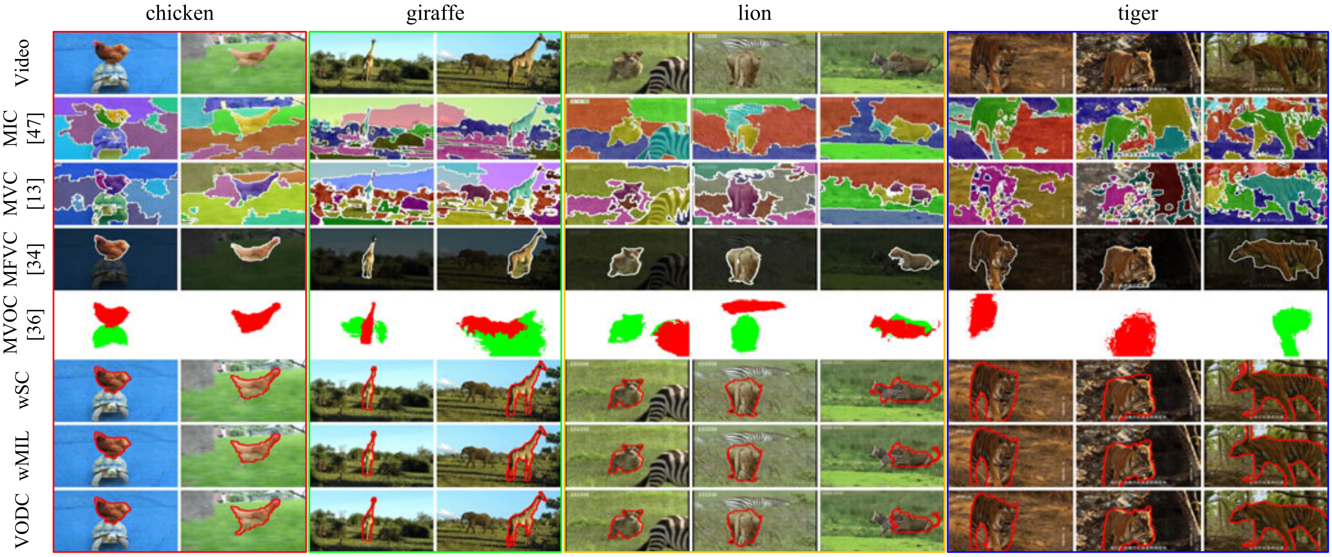


Fig. 6. Some visual example results of our methods and four state-of-the-art methods on the MOViCS dataset. The different colors in the 2^{nd}, 3^{rd}, and 5^{th} rows denote the class labels.

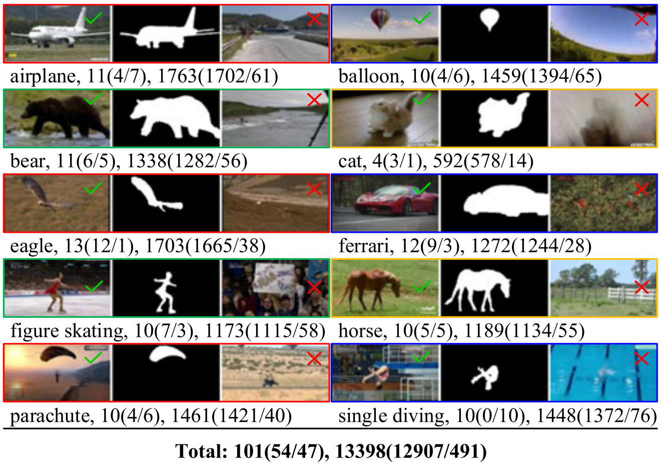

airplane, 11(4/7), 1763(1702/61)
balloon, 10(4/6), 1459(1394/65)
bear, 11(6/5), 1338(1282/56)
cat, 4(3/1), 592(578/14)
eagle, 13(12/1), 1703(1665/38)
ferrari, 12(9/3), 1272(1244/28)
figure skating, 10(7/3), 1173(1115/58)
horse, 10(5/5), 1189(1134/55)
parachute, 10(4/6), 1461(1421/40)
single diving, 10(0/10), 1448(1372/76)
Total: 101(54/47), 13398(12907/491)

Fig. 7. The example relevant ($\surd$) and irrelevant ($\times$) frames, the pixel-wise ground truth foreground labels (binary mask) for relevant frames, and the statistical details of the XJTU-Stevens video co-segmentation and classification dataset. For the airplane category, 11(4/7) denotes that the numbers of all videos, videos only containing relevant frames, and videos containing irrelevant frames are 11, 4, and 7, respectively; "1763(1702/61)" denotes that the numbers of all frames, relevant frames, and irrelevant frames are 1,763, 1,702, and 61, respectively.

category and the frames of each video are relatively limited, 3) the object instances in different videos are frequently the same object, or only exhibit small variations in color, shape, pose, size, and location, and 4) the annotations for the videos and frames are usually small. Thus, to exactly evaluate the efficacy of our method and to establish a benchmark for future research, we have collected a new dataset consisting of 10 categories of 101 publicly available internet videos (13,398 frames in total), and call it *XJTU-Stevens video co-segmentation and classification dataset*, in which some videos include irrelevant frames. The objects in videos of each category are of the common category, but exhibit large differences in appearance, size, shape, viewpoint, and pose. In support of the final evaluation of our end-to-end system, we manually assign each frame a label (1 for relevant and 0 for irrelevant) denoting whether the frame contains the target object, and also manually assign pixel-wise ground truth foreground labels for each relevant frame. We present some example relevant and irrelevant frames, the pixel-wise ground truth foreground labels for relevant frames, and the statistical details of the new dataset in Fig. 7.

Performance Evaluation. As the videos of each category always contain irrelevant frames which can be regarded as objects of different categories, we compare our video object discovery and co-segmentation methods (VODC, wSC, and wMIL) with three multi-class video co-segmentation methods (MVC [13], MFVC [34], and MVOC [36]) for fair comparison.

We first evaluate the discovery performance of our methods by varying the number of manually annotated relevant and irrelevant frames. The number of manually annotated relevant and irrelevant frames of each video is set from 1 to 3, and they are randomly selected from each video given the ground truth frame-level labels. We present the number of misclassified frames of each category in Table 5. As the results shown, all the irrelevant frames are misclassified as relevant ones by MVC [13], MFVC [34], and MVOC [36]. This is due to that all of them leverage the low-level color, shape and/or location cues. All of our methods work better than MVC [13], MFVC [34], and MVOC [36] even when we just provide each video with 1 relevant or irrelevant frame. Moreover, our full method (VODC) can identify almost all the relevant frames from multiple videos when we provide three relevant and irrelevant frames. This validates the efficacy of VODC and the spatio-temporal auto-context model learned through the Spatial-MILBoost algorithm.

TABLE 5
The Number of Misclassified Frames of Our Methods by Varying the Number of Manually Annotated Frames (I, II or III in the 2nd Row), and Three Multi-Class Video Co-Segmentation Methods on the XJTU-Stevens Video Co-Segmentation and Classification Dataset

<table>
<tr><th>Category</th><th>MVC</th><th>MFVC</th><th>MVOC</th><th colspan="3">wSC|wMIL|VODC</th></tr>
<tr><th></th><th>[13]</th><th>[34]</th><th>[36]</th><th>I</th><th>II</th><th>III</th></tr>
<tr><td>airplane</td><td>61</td><td>61</td><td>61</td><td>30|27|20</td><td>17|14|10</td><td>2|2|<b>0</b></td></tr>
<tr><td>balloon</td><td>65</td><td>65</td><td>65</td><td>17|17|13</td><td>8|7|4</td><td>4|4|<b>3</b></td></tr>
<tr><td>bear</td><td>56</td><td>56</td><td>56</td><td>10|7|3</td><td>5|4|3</td><td>4|3|<b>2</b></td></tr>
<tr><td>cat</td><td>14</td><td>14</td><td>14</td><td>11|9|<b>4</b></td><td>5|5|5</td><td>5|5|5</td></tr>
<tr><td>eagle</td><td>38</td><td>38</td><td>38</td><td>31|29|23</td><td>19|16|12</td><td>11|9|<b>8</b></td></tr>
<tr><td>ferrari</td><td>28</td><td>28</td><td>28</td><td>20|18|11</td><td>14|14|7</td><td>12|10|<b>6</b></td></tr>
<tr><td>figure skating</td><td>58</td><td>58</td><td>58</td><td>2|<b>0</b>|<b>0</b></td><td><b>0</b>|<b>0</b>|<b>0</b></td><td><b>0</b>|<b>0</b>|<b>0</b></td></tr>
<tr><td>horse</td><td>55</td><td>55</td><td>55</td><td>17|11|5</td><td>9|4|<b>1</b></td><td>4|2|<b>1</b></td></tr>
<tr><td>parachute</td><td>40</td><td>40</td><td>40</td><td>22|20|14</td><td>16|15|10</td><td>4|4|<b>2</b></td></tr>
<tr><td>single diving</td><td>76</td><td>76</td><td>76</td><td>35|26|18</td><td>26|21|13</td><td>17|11|<b>5</b></td></tr>
<tr><td><b>Avg.</b></td><td>49</td><td>49</td><td>49</td><td>20|16|11</td><td>12|10|7</td><td>6|5|<b>3</b></td></tr>
</table>

Lower values are better.

We further present the average IoU scores of our methods when providing with 3 relevant and irrelevant frames, and compare with MVC [13], MFVC [34], and MVOC [36] in Table 6. Some example results of them are presented in Fig. 8. Here, the average IoU score is computed on the relevant frames containing the target object. The results show that all of our methods can co-segment the intact objects with dramatic variations in appearance, size, pose, viewpoint, and shape out on most of the categories, and outperform MVC [13], MFVC [34], and

TABLE 6
The Average IoU Scores of Our Methods and Three Multi-Class Video Co-Segmentation Methods on the XJTU-Stevens Video Co-Segmentation and Classification Dataset

Category	MVC [13]	MFVC [34]	MVOC [36]	wSC	wMIL	VODC -NC	VODC -QS	VODC
airplane	58.4	46.4	61.2	83.1	84.7	85.4	**86.6**	86.4
balloon	86.9	91.5	87.4	93.2	93.9	94.1	94.5	**94.6**
bear	81.4	85.4	85.9	88.5	89.3	88.5	89.4	**90.5**
cat	75.6	70.4	80.7	85.2	89.4	90.7	91.5	**92.1**
eagle	72.8	81.4	79.5	82.3	86.2	87.4	88.2	**89.5**
ferrari	75.8	77.9	62.1	81.5	86.3	85.8	87.1	**87.7**
figure skating	62.1	55.4	65.8	83.4	86.9	86.2	86.7	**88.5**
horse	80.2	84.0	86.2	89.6	90.7	91.3	92.5	**92.0**
parachute	80.8	74.3	84.7	87.9	91.7	91.9	92.4	**94.0**
single diving	59.3	49.2	72.0	81.6	85.2	86.0	87.3	**87.7**
Avg.	73.3	71.6	76.6	85.6	88.4	88.7	89.6	**90.3**

Higher values are better.

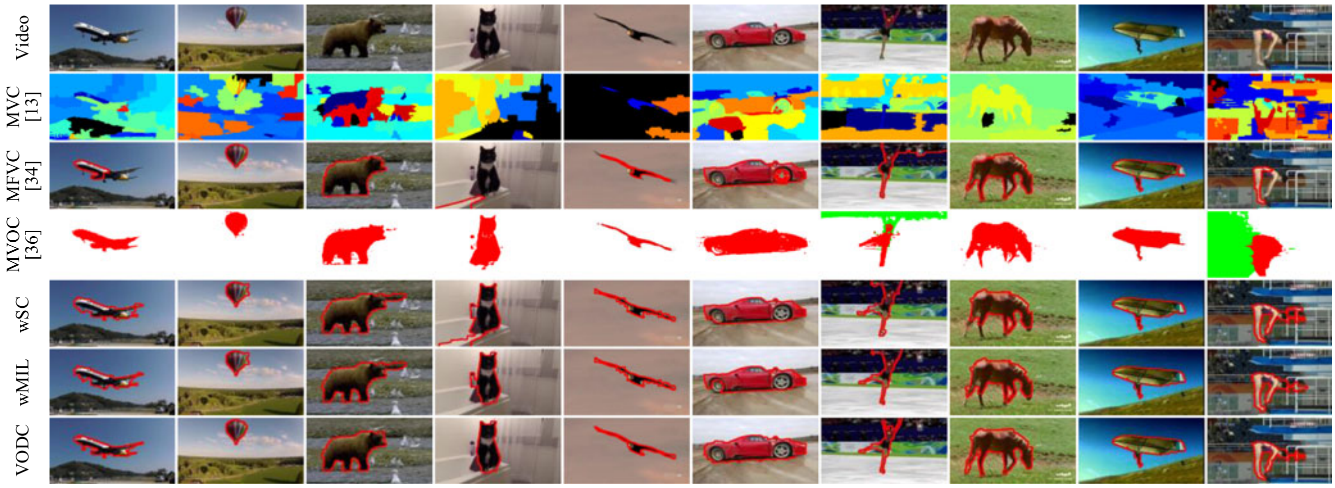


Fig. 8. Some visual example results of our methods and three multi-class video co-segmentation methods on the XJTU-Stevens video co-segmentation and classification dataset. The different colors in the 2nd and 4th rows denote the class labels.

MVOC [36] with an average improvement from 9 to 13.7 percent. The results of MVC [13] are over-segmented, and cannot capture the target object in its entirety (especially for the bear, ferrari and single diving categories). MFVC [34] highly relies on the object proposals, and thus sometimes cannot focus on the target object, as illustrated in the cat category. In summary, the above results demonstrate the advantages of our full method (VODC), the leveraged spatio-temporal auto-context model, and also the Spatial-MILBoost algorithm.

Parameter Analysis. N_c in Eq. (2) is the number of sampled points on the sampling structure of the spatio-temporal auto-context model. To evaluate its impact, we test our method by setting N_c to be 25, 33, 41, 49, and 57, respectively. We present the average IoU scores and the numbers of misclassified frames in Fig. 9. They show that both the object discovery and co-segmentation performances of our method almost always reach the best when N_c equals 41. Thus, N_c is set to be 41 in all our experiments.

Impact of Superpixel Methods. To study the impact of the superpixel methods to our method, we replace the superpixel method (SLIC [50]) used in our full method (VODC) with two other superpixel methods (NC [60] and QS [61]), and denote them as VODC-NC and VODC-QS, respectively. The results on the new dataset showed that they both have the same object discovery performance with VODC, and the differences of the average IoU scores are within 1.6 percent, as given in Table 6. This clearly demonstrates that our method is not tied to any specific superpixel method.

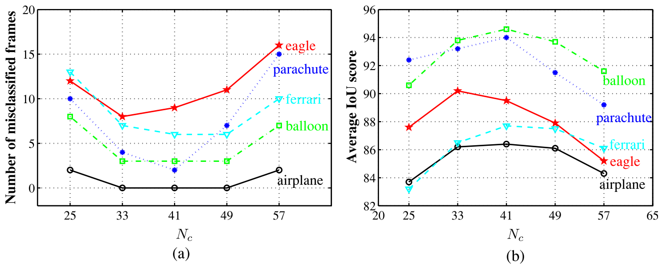


Fig. 9. (a) The numbers of misclassified frames and (b) the average IoU scores of our method by varying the value of N_c.

5.5 Strengths and Limitations

To summarize, as the above results on four datasets shown, our method 1) achieves superior performance of segmenting the target object from a single video compared to the state-of-the-art single video object segmentation methods, 2) compares favorably with the state-of-the-art video object co-segmentation methods in object co-segmentation from multiple videos only containing relevant frames, 3) outperforms the state-of-the-art multi-class video co-segmentation methods in joint object discovery and co-segmentation from multiple videos containing irrelevant frames. Moreover, the ablative studies demonstrate the advantages of the spatio-temporal auto-context model which captures the categorized spatio-temporal contextual information across multiple videos, and the Spatial-MILBoost algorithm which considers the spatial relationship of neighboring superpixels while predicting the segmentation label of superpixel.

Our method is capable of discovering and co-segmenting a common category of objects from multiple videos containing irrelevant frames, but it needs to be bootstrapped with 1 to 3 manually annotated frame-level labels. Since the reconstruction errors of an autoencoder are discriminative enough to well separate the inliers and outliers, and it has been proven that the autoencoder is a simple yet effective tool for separating inliers and outliers in an unsupervised fashion, thus in our future work, we plan to utilize the reconstruction errors of an autoencoder [62] to automatically select dozens of relevant frames (inliers) and irrelevant frames (outliers) to initialize our method.

Moreover, our method cannot handle the case where the common objects of multiple categories are present in the video. In our future work, we will extend our method to discover and co-segment the common objects of multiple categories from multiple videos. One possible solution is to generate initial segmentations for the common objects of each category (e.g., using a co-saliency measure [63]), and proceed to iteratively learn an appearance model, a spatio-temporal auto-context model, and a consistency term for the common objects of each category and perform optimization on Eq. (1). Through these steps, we may be able to extend our current method to not only discover the relevant frames, but also co-segment the common objects of multiple categories from the relevant frames, respectively.

6 Conclusion

We presented a spatio-temporal energy minimization formulation to simultaneously discover and co-segment a common category of objects from multiple videos containing irrelevant frames, which only requires extremely weak supervision (i.e., 1 to 3 frame-level labels). Our formulation incorporates a spatio-temporal auto-context model to capture the spatio-temporal contextual information across multiple videos. It facilitates both object discovery and co-segmentation through a multiple instance learning algorithm with spatial reasoning. Our method overcomes an important limitation of previous video object co-segmentation methods, which assume all frames from all videos contain the target objects. Experiments on four datasets demonstrated the superior performance of our proposed method.

Acknowledgments

This work was supported partly by NSFC Grants 61231018, 61629301, and 61273252, and National Key Research and Development Plan 2016YFB1001004. Le Wang was supported by NSFC Grant 61503296, China Postdoctoral Science Foundation Grant 2015M572563, the Fundamental Research Funds for the Central Universities Grant XJJ2015066, and China Scholarship Council. Zhenxing Niu was supported by NSFC Grant 61672402.

References

[1] D.-J. Chen, H.-T. Chen, and L.-W. Chang, "Video object cosegmentation," in *Proc. ACM Multimedia*, 2012, pp. 805–808.

[2] J. C. Rubio, J. Serrat, and A. López, "Video co-segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 13–24.

[3] C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang, "Video object co-segmentation via subspace clustering and quadratic pseudo-boolean optimization in an MRF framework," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 903–916, Jun. 2014.

[4] J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou, "Video co-segmentation for meaningful action extraction," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 2232–2239.

[5] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3137–3148, Oct. 2015.

[6] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[7] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient ND image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[8] Z. Tu, "Auto-context and its application to high-level vision tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[9] 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.

[10] 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.

[11] F. Tiburzi, M. Escudero, J. Bescós, and J. M. Martínez, "A ground truth for motion-based video-object segmentation," in *Proc. Int. Conf. Image Process.*, 2008, pp. 17–20.

[12] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2141–2148.

[13] 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.

[14] 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.

[15] D. Liu and T. Chen, "A topic-motion model for unsupervised video object discovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[16] G. Zhao, J. Yuan, and G. Hua, "Topical video object discovery from key frames by modeling word co-occurrence prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1602–1609.

[17] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid, "Unsupervised object discovery and tracking in video collections," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3173–3181.

[18] J. Yang, et al., "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. Circuits Syst. for Video Tech.*, vol. 26, no. 6, pp. 1070–7083, Jun. 2016.

[19] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Dec. 2010.

[20] 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.

[21] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 284–302, 2010.

[22] H. Wang, G. Zhao, and J. Yuan, "Visual pattern discovery in image and video data: a brief survey," *WIREs: Data Mining Knowl. Discovery*, vol. 4, no. 1, pp. 24–37, 2014.

[23] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video SnapCut: Robust video object cutout using localized classifiers," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 70.

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

[25] S. A. Ramakanth and R. V. Babu, "SeamSeg: Video object segmentation using patch seams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 376–383.

[26] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.

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

[28] 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.

[29] 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.

[30] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1777–1784.

[31] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1092–1106, Jun. 2014.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 465–479.

[38] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3169–3176.

[39] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1943–1950.

[40] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2217–2224.

[41] M. Rubinstein, C. Liu, and W. T. Freeman, "Annotation propagation in large image databases via dense image correspondence," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 85–99.

[42] J. Dai, Y. N. Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1305–1312.

[43] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas, "Unsupervised multi-class joint image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3142–3149.

[44] W. Tao, K. Li, and K. Sun, "SaCoseg: Object cosegmentation by shape conformability," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 943–955, Mar. 2015.

[45] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, "Multiple random walkers and their application to image cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3837–3845.

[46] G. Kim and E. P. Xing, "On multiple foreground cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 837–844.

[47] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 542–549.

[48] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1939–1946.

[49] 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, Sep. 2014.

[50] 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, Nov. 2012.

[51] 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.

[52] L. Wang, J. Xue, N. Zheng, and G. Hua, "Concurrent segmentation of categorized objects from an image collection," in *Proc. Int. Conf. Pattern Recognit.*, 2012, pp. 3309–3312.

[53] 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 Recognit.*, vol. 46, no. 11, pp. 2874–2889, 2013.

[54] 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, Sep. 2012.

[55] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 73–80.

[56] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[57] P. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proc. Advances Neural Inf. Process. Syst.*, 2005, pp. 1417–1424.

[58] S. Avidan, "SpatialBoost: Adding spatial reasoning to adaBoost," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 386–396.

[59] P. Ochs and T. Brox, "Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1583–1590.

[60] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 1124–1131.

[61] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 705–718.

[62] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1511–1519.

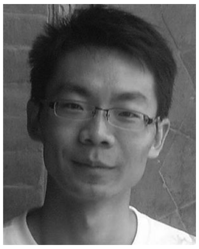
[63] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

Le Wang received the BS and PhD degrees in control science and engineering from Xi'an Jiaotong University, in 2008 and 2014, respectively. From 2013 to 2014, he was a visiting PhD student with Stevens Institute of Technology. From 2016 to 2017, he is a visiting scholar with Northwestern University. He is currently an assistant professor in the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, machine learning, and their application for web images and videos. He is a member of the IEEE.

Gang Hua was enrolled in the special class for the gifted young from Xi'an Jiaotong University (XJTU), in 1994, the BS degree in automatic control engineering from XJTU, in 1999, the MS degree in control science and engineering from XJTU, in 2002, and the PhD degree in electrical engineering and computer science from Northwestern University, in 2006. He is currently a senior research manager with Microsoft Research Asia. Before that, he was an associate professor of computer science with Stevens Institute of Technology. He also held an academic advisor position with IBM T. J. Watson Research Center between 2011 and 2014. He was a research staff member with IBM Research T. J. Watson Center from 2010 to 2011, a senior researcher with Nokia Research Center, Hollywood from 2009 to 2010, and a scientist with Microsoft Live Labs Research from 2006 to 2009. He is an associate editor of the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *Computer Vision and Image Understanding*, the *IEEE Multimedia*, the *The Visual Computer Journal* and the *Machine Vision and Applications*. He also served as the lead guest editor on two special issues in the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *International Journal of Computer Vision*, respectively. He is a program chair of CVPR'2019. He is an area chair of CVPR'2017, CVPR'2015, ICCV'2011, ICIP'2012&2013, ICASSP'2012&2013, and ACM MM 2011&2012&2015. He is the author of more than 120 peer reviewed publications in prestigious international journals and conferences. He holds 18 US patents and has 9 more US patents pending. He received the 2015 IAPR Young Biometrics Investigator Award for his contribution on Unconstrained Face Recognition from Images and Videos, and the 2013 Google Research Faculty Award. He is a Fellow of the IAPR, a Senior Member of the IEEE, and a Distinguished Scientists of the ACM.

Rahul Sukthankar received the BSE in computer science from Princeton, in 1991 and the PhD degree in robotics from Carnegie Mellon, in 1997. He leads research efforts in computer vision, machine learning and robotics with Google. He is also an adjunct research professor in the Robotics Institute at Carnegie Mellon and courtesy faculty with the University of Central Florida. He was previously a senior principal researcher with Intel Labs, a senior researcher with HP/Compaq Labs and a research scientist with Just Research. He has organized several workshops and conferences and currently serves as editor in chief of the *Machine Vision and Applications* journal. He is a member of the IEEE.

Jianru Xue received the BS degree from Xi'an University of Technology, in 1994 and the MS and PhD degrees from Xi'an Jiaotong University, in 1999 and 2003, respectively. From 2002 to 2003, he was with Fuji Xerox. In 2008 to 2009, he visited the University of California, Los Angeles. He is currently a professor in the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research field includes computer vision, visual navigation, and video coding based on analysis. He served as a Coorganization chair for ACCV'2009 and ICVSM'2006. He also served as a program committee member for ICPR'2012, ACCV'2010&2012, and ICME'2014. He is a member of the IEEE.



Zhenxing Niu received the PhD degree in control science and engineering from Xidian University, in 2012. From 2013 to 2014, he was a visiting scholar with the University of Texas at San Antonio. He is currently an associate professor in the School of Electronic Engineering, Xidian University. His research interests include computer vision, machine learning, and their application in object discovery and localization. He served as PC member of CVPR'2014, ACM MM'2014&2015. He is a member of the IEEE.

Nanning Zheng received the graduate degree from the Department of Electrical Engineering, Xi'an Jiaotong University (XJTU), in 1975, the ME degree in information and control engineering from Xi'an Jiaotong University, in 1981, and the PhD degree in electrical engineering from Keio University, in 1985. He is currently a professor and the director of the Institute of Artificial Intelligence and Robotics, 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.