

# Motion-Appearance Interactive Encoding for Object Segmentation in Unconstrained Videos

Chunchao Guo, Jianhuang Lai, and Xiaohua Xie  
Sun Yat-sen University, China

**Abstract**—We present a novel method of integrating motion and appearance cues for foreground object segmentation in unconstrained videos. Unlike conventional methods encoding motion and appearance patterns individually, our method puts particular emphasis on their mutual assistance. Specifically, we propose using an interactively constrained encoding (ICE) scheme to incorporate motion and appearance patterns into a graph that leads to a spatiotemporal energy optimization. The reason of utilizing ICE is that both motion and appearance cues for the same target share underlying correlative structure, thus can be exploited in a deeply collaborative manner. We perform ICE not only in the initialization but also in the refinement stage of a two-layer framework for object segmentation. This scheme allows our method to consistently capture structural patterns about object perceptions throughout the whole framework. Our method can be operated on superpixels instead of raw pixels to reduce the number of graph nodes by two orders of magnitude. Moreover, we propose to partially explore the multi-object localization problem with inter-occlusion by weighted bipartite graph matching. Comprehensive experiments on three benchmark datasets (i.e., SegTrack, MOVICS, and GaTech) demonstrate the effectiveness of our approach compared with extensive state-of-the-art methods.

**Index Terms**—video object segmentation, foreground detection, interactively constrained encoding.

## I. INTRODUCTION

THE purpose of video object segmentation is to acquire foreground moving objects in videos. Foreground object segmentation is greatly significant and has been leveraged for use in various vision tasks, including object appearance modeling [1], object tracking [2], video matting [3], activity recognition [4], and image retrieval [5]. Compared with early methods that only consider the case of static camera settings and address this problem through static background subtraction [6], [7], separating targets in an arbitrary background is inherently more difficult due to camera jittering, motion blur, and the fast and large displacement of targets. Recent years have witnessed much progress [8], [9], [10], [11] of handling unconstrained videos, but it remains an open issue that has not yet been adequately explored.

Motion features (e.g., optical flow) and appearance features (e.g., color segmentations) are both important cues for

addressing the object segmentation problem with an unconstrained background. However, optical flow generates inaccurate boundaries and it often diffuses under rapid motion, while appearance is severely hindered by cluttered or low-contrast backgrounds. Therefore, a natural idea is to integrate motion and appearance cues for object segmentation. In many related studies [10], [11], [12], the feature-level or decision-level fusion with regard to these two cues has been considered, yet motion and appearance patterns are separately extracted and not integrated in a deeply collaborative way. In other words, traditional manner neglects the intrinsic correlation between motion and appearance patterns. Actually, motion and appearance features for the same target to a certain extent are homologous, and share underlying correlative patterns about object perceptions, including semantic structure, shape, and movement. Therefore, for well detecting a moving target, it is better to exploit the motion and appearance features synergistically rather than to utilize them separately. Along this line, we develop an interactively constrained encoding (ICE) approach for integrating motion and appearance cues and incorporate it into a coarse-to-fine framework. The procedure of feature encoding is interactive between multi-type cues; that is, ICE imposes motion constraints during appearance feature encoding, and vice versa. Unlike many existing methods in which multiple-feature fusion only serves the initialization stage, our method performs ICE throughout all of the stages of the coarse-to-fine framework. Especially, ICE allows our method to capture the semantic structure of object perception while refining moving regions.

Figures 1(f) and 4(d) illustrate living examples of ICE and Figure 2 demonstrates our framework. In Figure 1, (b) and (c) are the segmentation cues extracted by [11] and [13], respectively. Figure 1(e) is the combination of (b) and (c), but the motion and appearance cues are encoded separately. By considering the interactively constrained encoding on two cues, our approach obviously gets more uniform values inside the target than (c) and (e), and preserves much more accurate boundaries than (b) and (e). Overall, ICE exhibits more potential in high-level object perception.

Besides putting forward ICE, we employ another two strategies in addressing the video object segmentation problem. First, the superpixel representation is used to reduce the computation complexity. Because our method performs ICE throughout the whole framework, the superpixel-level graph still works well and maintains the ability of perceiving targets in the video even with a cluttered background. Second, to handle the multi-object initialization with inter-occlusion, we

Xiaohua Xie is the corresponding author.

Chunchao Guo, Jianhuang Lai, and Xiaohua Xie are all with the School of Data and Computer Science, Sun Yat-sen University, and with the Guangdong Key Laboratory of Information Security Technology, Guangzhou, 510006, P. R. China. E-mail: chunchaoguo@gmail.com, stsljh@mail.sysu.edu.cn, xiexiaoh6@mail.sysu.edu.cn.

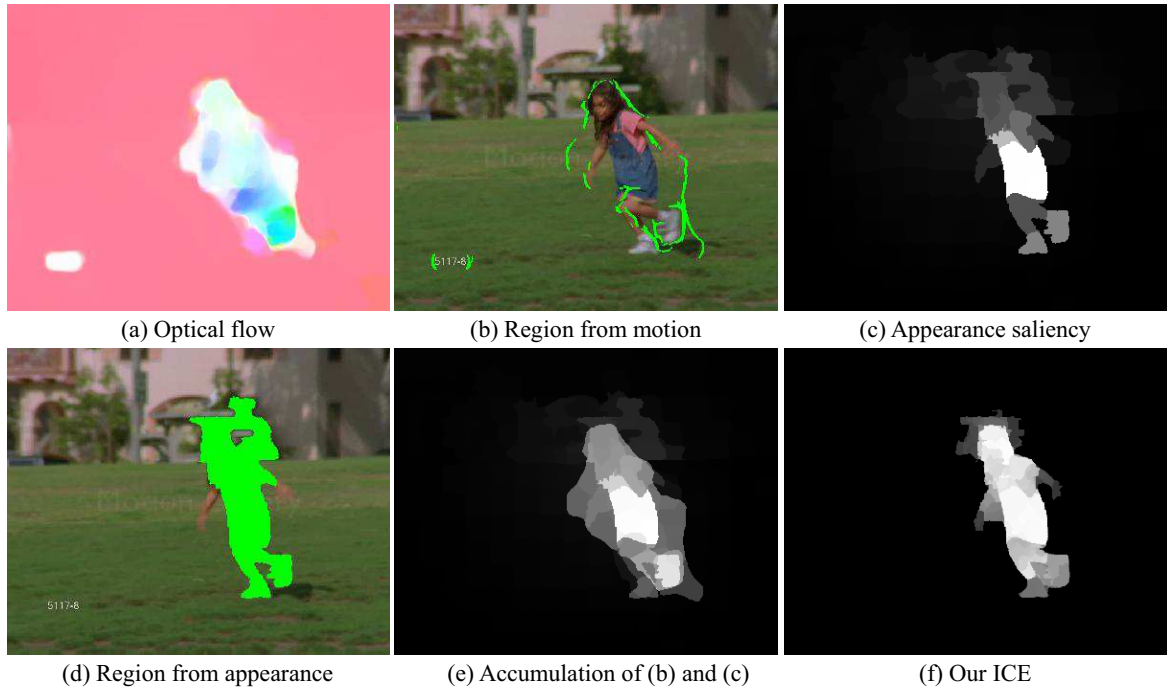


Fig. 1: Illustration of separate encoding and Interactively Constrained Encoding (ICE) of motion and appearance patterns for a moving target. Best view in color.

propose to activate maximum bipartite graph matching between adjacent frames at the proposal level, which re-assigns coarse IDs to different occluded objects. It should be noted that the inter-occlusion is often ignored by previous studies.

The remainder of this paper is structured as follows. Related works are reviewed in Section 2. Section 3 presents our approach. Section 4 demonstrates the results on three benchmark datasets, and our conclusions are drawn in Section 5.

## II. RELATED WORKS

This paper aims to detect and segment moving foreground objects in videos, a goal shared by previous works on background subtraction and video object segmentation. Related works are briefly introduced in this section.

### A. Background Subtraction

When detecting moving regions, a way off the shelf is to model pixel-wise backgrounds and then subtract it to find pixels whose differences exceed a threshold. Both parametric [6], [14], [15] and nonparametric [7] mechanisms can be adopted for constructing backgrounds. The single Gaussian model [14] is a simple method for fitting the distribution of pixel intensity, but it is insufficient in complex scenes. The elaborate mixture of Gaussian [6] provides a better distribution of background pixels and it is usually eligible for non-cluttered scenes. Chen et al. [16] proposed a hierarchical block-based approach that combines Mixture of Gaussian and a contrast histogram. With respect to nonparametric approaches, the background model in [7] is built upon a set of background samples and is updated by randomly selecting samples. [7] also considers neighbor pixels during updating, which is different from Gaussian mixture

models [6]. Nevertheless, the common drawback of traditional background modeling approaches is that they usually require static camera settings, and thus cannot cope well with moving backgrounds.

### B. Video Object Segmentation

Video object segmentation has emerged as a feasible solution for tackling arbitrary background. Conventional methods of video object segmentation can be categorized into the supervised mode and the unsupervised mode.

1) *Supervised Methods*: Supervised approaches usually require manual annotations in several key frames to explicitly indicate moving objects. Tsai et al. [9] proposed to track human-labeled regions and segment the remaining frames utilizing a multi-label Markov random field model. Chockalingam et al. [17] also requires a manual label that indicates the location of a moving target. Recently, some methods [18], [19] have focused on co-segmenting objects, given multiple videos where the same targets appear simultaneously. Those can be categorized as weakly supervised tasks, as the same video object is must be present in multiple videos and the co-occurrence hints at what the object looks like.

2) *Unsupervised Methods*: Unsupervised segmentation methods mainly build on motion trajectories or optimization in a graph with integrating multiple cues.

a) *Methods based on Motion Trajectory*: Trajectories provide a natural way to describe object movement. Those methods usually obtain trajectories by linking dense points or objects, and then the point trajectories are clustered and mapped into pixel labels. Brox et al. [8] exploited dense flow points to cluster long-term motion trajectories, and the temporally consistent clusters alleviated the influence of objects

that were sometimes static in the sequence. Fragkiadaki et al. [20] detected discontinuities for trajectory embedding to obtain motion boundaries, and thus segment objects from world scenes. The common underlying assumption that supports these methods is motion homogeneity, where the points within objects share a single similarity transformation. However, deformable objects apparently do not meet this requirement and may lead to poor performance. Moreover, these methods rely heavily on the reliability of optical flow and explore insufficient appearance clues.

*b) Methods based on Coarse-to-Fine Graph Optimization:* Beyond tracking dense points, an alternative that benefits more from both motion and appearance is the coarse-to-fine optimization scheme in a graph. Raw objects are initially localized with the help of optical flow [11] or generic object proposals [10], and then multiple clues are incorporated into an optimization procedure to refine the foreground labels. Lim et al. [21] employed block-wise density propagation to obtain likelihood maps, and then optimized those maps. However, this approach involves a large number of parameters and thus is not feasible for generalization. Wang et al. [22] designed a framework that incorporates robust geodesic measurement to segment video targets. [23] optimizes a weighted graph at the pixel level using the shortest path algorithm. Giordano et al. [12] incorporated properties of a compact geometrical structure and optimized the graph around each moving region based on appearance and perceptual organization.

Although motion and appearance patterns are both used in above-mentioned methods when optimizing a graph, the common drawback of these methods is that motion and appearance are treated as isolated components at a low level without considering their homologous property. Furthermore, existing methods often ignore the inter-occlusion problem in multiple object location. Our method tends to address all these shortcomings.

### III. THE PROPOSED APPROACH

#### A. Overview and Notations

Our framework comprises two stages: the label initialization stage and the refinement stage. The proposed interactively constrained encoding (ICE) will be used throughout both stages. Figure 2 briefly illustrates our approach.

For clarity, the notations used in our paper are provided in Table I.

#### B. Interactively Constrained Encoding (ICE)

*1) Overview of ICE:* To exploit the homologous properties of multimodal cues and capture semantic structural information on object perception, appearance restrictions are leveraged to induce the extraction of motion patterns, and vice versa.

The motion cues used in our approach mainly include optical flow [24] and its further derivative features, such as a color image and a gradient map of optical flow. In terms of appearance, object proposals [25] are the primary cues in our approach. Appearance saliency [13], trimap and color descriptors are also exploited as supplementary cues.

Symbol	Definition
$F^i \in \mathcal{X}^{W \times H}$	The $i$ th frame with width $W$ and height $H$
$\mathcal{F} = \{F^i\}_{i=1}^K$	A video sequence with $K$ frames
$(j, k)$	A pixel index
$O^i$	The two-channel optical flow image
$V^i$	The intensity magnitude of $O^i$
$E^i$	The gradient magnitude of $O^i$
$C^i$	The three-channel color optical flow image
$Y_{RGB}^i$	The appearance saliency map of $F^i$
$Y_C^i$	The appearance saliency map of $C^i$
$\mathcal{P}_{RGB}^i = \{P_{RGB}^{i,r}\}_{r=1}^{N_1}$	A set of object proposals in $F^i$
$\mathcal{P}_C^i = \{P_C^{i,r}\}_{r=1}^{N_2}$	A set of object proposals in $C^i$
$P^{i,r}$	The $r$ th proposal in $F^i$ or $C^i$
$D^{i,r}$	The binary mask of $P^{i,r}$
$B^{i,r}$	The binary boundary/edge map of $P^{i,r}$
$G$	The gradient/boundary strength for a proposal $P^{i,r}$
$I$	The intensity strength for a proposal $P^{i,r}$
$G^i$	The accumulated gradient/boundary strength of $\mathcal{P}$ in $F^i$ or $C^i$
$I^i$	The accumulated intensity strength of $\mathcal{P}^i$ in $F^i$ or $C^i$
$T^i$	The trimap for $C^i$
$\mathbf{M}^i$	The ICE map for $F^i$
$H_p$	Multiple concatenated histograms for a superpixel/node $p$
$\mathcal{D}$	The Bhattacharyya distance

TABLE I: The notations in this paper.

The gradient magnitude map  $E^i$  in the optical flow field  $O^i$  is calculated by:

$$E^i = \|\nabla O^i\|_2. \quad (1)$$

For  $P^{i,r}$ , its gradient or boundary strength  $G$  in the optical flow field is formulated as

$$G(P^{i,r}) = \frac{1}{Z_G^{i,r}} \sum_{j=1}^H \sum_{k=1}^W E_{j,k}^i \cdot \mathbb{1}(B_{j,k}^{i,r} = 1), \quad (2)$$

where  $(j, k)$  is a pixel index,  $\mathbb{1}$  is the indicator function, and  $Z_G^{i,r} = \sum_{j=1}^H \sum_{k=1}^W B_{j,k}^{i,r}$  is a normalization factor. Given  $P^{i,r}$ , Eq.(2) evaluates the strength of a proposal boundary that overlaps with the boundary of a moving object.

In the same way, intensity strength  $I$  in the optical flow field for a proposal  $P^{i,r}$  is

$$I(P^{i,r}) = \frac{1}{Z_I^{i,r}} \sum_{j=1}^H \sum_{k=1}^W V_{j,k}^i \cdot \mathbb{1}(D_{j,k}^{i,r} = 1), \quad (3)$$

where  $Z_I^{i,r} = \sum_{j=1}^H \sum_{k=1}^W D_{j,k}^{i,r}$  is also a normalization factor. Given  $P^{i,r}$ , Eq.(3) assesses its intensity strength weighted by the intensity of the optical field. This metric also indicates the likelihood of belonging to moving targets for  $P^{i,r}$ .

*2) Appearance-Constrained Motion:* We adapt appearance restrictions to the motion field; that is, we aim to encode patterns in  $O^i$ , by following the manner of extracting appearance features. Given  $C^i$ , we calculate its appearance saliency map [13]  $Y_C^i$ , its color name descriptors [26] and its trimap  $T^i$ . The construction of  $T^i$  is individually introduced later. Color name

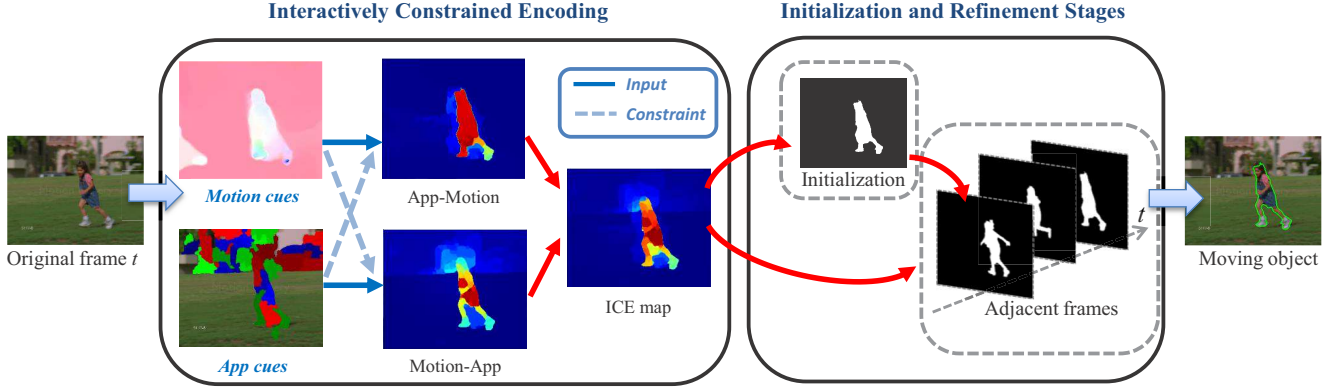


Fig. 2: Flowchart of the framework.

descriptors categorize each pixel into the eleven semantic color names and thus produce a histogram with eleven dimensions.

Then, we rank and accumulate  $\mathcal{P}_C^i = \left\{ P_C^{i,r} \right\}_{r=1}^{N_2}$  in  $C^i$  based on Eqs.(2) and (3). The accumulated boundary or gradient strength of  $\mathcal{P}_C^i$  is

$$G_C^i = \sum_{r=1}^{N_2} B_C^{i,r} \cdot G \left( P_C^{i,r} \right) \quad (4)$$

Given  $B_C^{i,r}$ , as a mask for  $P_C^{i,r}$ , Eq.(4) assigns a uniform value  $G \left( P_C^{i,r} \right)$  for each pixel belonging to  $P_C^{i,r}$ . Considering that those proposals are generated based on similarity in  $C^i$  and are thus compact in motion, the settings in Eq.(4) can preserve spatial layouts of targets when many interior boundaries occur inside them.

In the same way, the accumulated intensity strength of  $\mathcal{P}_C^i$  is

$$I_C^i = \sum_{r=1}^{N_2} B_C^{i,r} \cdot I \left( P_C^{i,r} \right), \quad (5)$$

Note that, while appearance constraints were imposed during the above feature encoding, the whole procedures are conducted in optical flow field  $O^i$  or  $C^i$ . Thus,  $Y_C^i$ ,  $G_C^i$ ,  $I_C^i$ ,  $T^i$  and the color descriptors of  $C^i$  are categorized into the appearance-constrained motion.

a) *Trimap*: A trimap  $T^i$  denotes the division of definite foreground, definite background, and ambiguous regions in  $F^i$ . During model optimization,  $T^i$  reveals the correlations of regions both in a local and global view, and thus can narrow down our focus to only ambiguous regions.

Unlike [27], where the goal is to find salient objects in appearance, we aim for localizing moving areas. Hence, our moving trimap  $T^i$  is built upon  $Y_C^i$ .  $Y_C^i$  is first subdivided into equal-sized blocks under three different spatial scales, with the number of blocks in each scale being  $2 \times 2$ ,  $3 \times 3$ , and  $4 \times 4$ , respectively. Similar to the settings in [27], the Otsu's algorithm [28] with seven-level threshold is individually applied to each block in each scale. Then, all three maps are summed to global map  $Y^{i'}$ . Hence,  $Y^{i'}$  is a map with 21 levels.  $T^i$  is then constructed by globally thresholding  $Y^{i'}$

using Eq.(6), through a conservative scheme to ensure the purity of the definite foreground:

$$T^i(j, k) = \begin{cases} 1, & \text{if } Y^{i'}(j, k) \geq \theta_1, \\ 0, & \text{if } Y^{i'}(j, k) \leq \theta_2, \\ 0.5, & \text{otherwise} \end{cases} \quad (6)$$

where  $(j, k)$  is a pixel index in frame  $F^i$ . In this work,  $\theta_1$  and  $\theta_2$  are set as 18 and 6, respectively.

An example of trimap in Figure 3 indicates that it benefits for localizing objects and narrowing down our attention region.

Thus, an appearance-constrained motion map  $\mathbf{M}_C^i$  is expressed as

$$\mathbf{M}_C^i = G_C^i + I_C^i + \alpha \cdot Y_C^i + \beta \cdot T^i, \quad (7)$$

where  $\alpha$  and  $\beta$  are set as 0.9 and 0.5 in our experiments, respectively. Each term in Eq.(7) is restricted by the object-level appearance relations in  $C^i$ .

3) *Motion-Constrained Appearance*: In contrast to  $\mathbf{M}_C^i$ , which has the aid of appearance constraints, motion restrictions can also be leveraged to encode appearance patterns. Object proposals in color space uncover the potential of maintaining the semantic structure even in cases of optical flow failure. We first rank and accumulate  $\mathcal{P}_{RGB}^i$  based on Eqs.(2) and (3).

The accumulated gradient strength of  $\mathcal{P}_{RGB}^i$  is

$$G_{RGB}^i = \sum_{r=1}^{N_1} B_{RGB}^{i,r} \cdot G \left( P_{RGB}^{i,r} \right). \quad (8)$$

Given  $B_{RGB}^{i,r}$ , Eq.(8) assigns a uniform value  $G \left( P_{RGB}^{i,r} \right)$  for each pixel belonging to  $P_{RGB}^{i,r}$ . Here optical flow boundaries  $E^i$  must be pre-processed before using Eq.(8) is required to avoid many zero values. Given that the true image boundaries in  $F^i$  are slightly misaligned with  $E^i$  in  $O^i$ , we first produce an expansion map  $E^{i'}$  by dilating  $E^i$  to make it overlap with the boundaries of  $F^i$  and  $\mathcal{P}_{RGB}^i$ .

In the same way, the accumulated intensity strength of  $\mathcal{P}_{RGB}^i$  is

$$I_{RGB}^i = \sum_{r=1}^{N_1} B_{RGB}^{i,r} \cdot I \left( P_{RGB}^{i,r} \right). \quad (9)$$

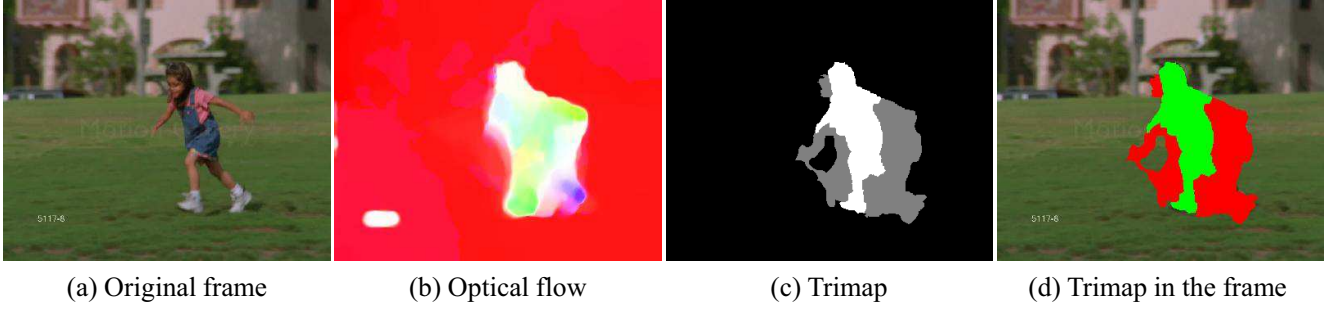


Fig. 3: Illustration of a trimap. In (d), the green middle region represents definite foreground, the red color marks ambiguous regions, and the remaining region is the background.

Note that, while optical flow cues aid feature encoding here, the whole procedures are still operated on  $\mathcal{P}_{RGB}^i$ .  $\mathcal{P}_{RGB}^i$  belongs to appearance patterns, as they are generated from RGB frames and the generation is uninfluenced by optical flow. Thus, the feature maps in this part are categorized into the motion-constrained appearance.

The motion-constrained appearance map  $\mathbf{M}_{RGB}^i$  is then formulated as

$$\mathbf{M}_{RGB}^i = G_{RGB}^i + I_{RGB}^i + \alpha \cdot Y_{RGB}^i. \quad (10)$$

4) *The Eventual ICE Map*: Having achieved the motion-constrained appearance and the appearance-constrained motion, we assign equal weights to the two terms. An ICE map is then calculated by

$$\mathbf{M}^i = \mathbf{M}_C^i + \mathbf{M}_{RGB}^i. \quad (11)$$

We then normalize  $\mathbf{M}^i$  to  $[0, 1]$ . Eventually,  $\mathbf{M}^i$  indicates the probabilities of being a moving foreground for the pixels in  $F^i$ .

Our ICE scheme considers mutual restrictions from multi-modal cues for the same target, and it tends to improve the robustness and accuracy of perceiving moving targets. Figure 4 provides a living example, which shows that our approach can work effectively in variant environments, even in which the optical flow method fails.

### C. Label Initialization

Having obtained an ICE map  $\mathbf{M}^i$ , an adaptive threshold  $t^i$  is computed to binarize  $\mathbf{M}^i$ .  $t^i = 0.5 \cdot (\mu(\mathbf{M}^i) + \rho(\mathbf{M}^i))$ , where  $\mu(\cdot)$  is the average value of  $\mathbf{M}^i$  and  $\rho(\cdot)$  denotes the threshold computed on  $\mathbf{M}^i$  using Otsu's algorithm [28]. The initial foreground or background map  $X^i$  is obtained by

$$X^i(j, k) = \begin{cases} 1, & \text{if } \mathbf{M}^i(j, k) \geq t^i \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where  $(j, k)$  is a pixel index in  $F^i$ . Eq.(12) usually works well in cases of single-object segmentation and multi-object segmentation without occlusions.

Nevertheless, for a sequence  $\mathcal{F} = \{F^i\}_{i=1}^K$  containing multiple object movement, there may exist inter-occlusion in certain frames  $\{F^{o+1}, \dots, F^{o+r}\}$  after initialization. To partially alleviate this problem, we introduce a conservative procedure to activate the inter-occlusion handling. A decrease

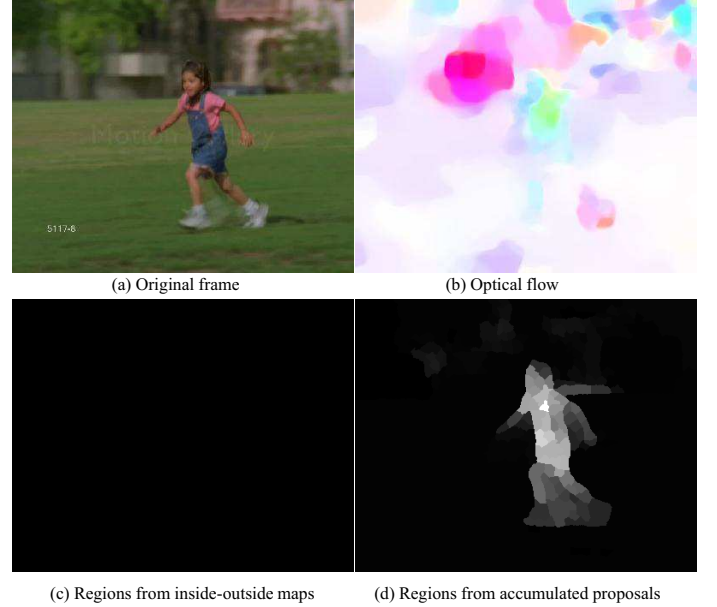


Fig. 4: Illustration of moving object perception by different method. Optical flow in (b) is inaccurate due to rapid motion. The initial moving region in (c) is calculated using inside-outside maps according to [11], which only relies on optical flow and fails in this case. (d) is our accumulated object proposals induced by optical flow. Motion(optical flow) with appearance (RGB proposals) release potentials of object perception due to the homologous property.

in the number of detected targets suggests that certain targets leave the scene or inter-occlusion happens. Hence, our decision criterion is that the number of initial targets decreases from  $n$  to  $m$  and then returns to  $n$ . Here,  $n$  is the number of targets before decreasing, and  $m$  is the number of detected occluded blobs, s.t.  $m < n$ . This criterion indicates that inter-occlusion occurs from  $F^{o+1}$  and the occluded blobs re-split to  $n$  isolated targets after  $F^{o+r}$ . Thus, the  $m$  occluded blobs in  $\{F^{o+1}, \dots, F^{o+r}\}$  can be split into proposals and re-assigned IDs.

We construct a graph  $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$  for every two adjacent frames  $F^i, F^{i+1}$ , where  $\mathcal{V}_1 = \mathcal{P}_{RGB}^i \cup \mathcal{P}_{RGB}^{i+1}$  and  $\mathcal{P}_{RGB}^{i+1}$  is the set of proposals inside the occluded blobs. Given that  $\mathcal{P}_{RGB}^i$  and  $\mathcal{P}_{RGB}^{i+1}$  are disjoint sets of proposals from adjacent



frames  $F^i$  and  $F^{i+1}$ , respectively,  $\mathcal{G}_1$  naturally consists of a bipartite graph. Every edge  $e \in \mathcal{E}_1$  indicates the cost between two nodes/proposals from  $P_{RGB}^{i,j}$  and  $P_{RGB}^{i+1,k}$ , respectively. The cost of  $e$  is calculated by three terms, including the Bhattacharyya distance between concatenated of simple RGB and LAB histograms, the normalized difference of the sizes of boundingboxes, and the normalized centroid distance between  $P_{RGB}^{i,j}$  and  $P_{RGB}^{i+1,k}$ . The maximum matching of bipartite graph  $\mathcal{G}_1$  is solved by the Hungarian algorithm.

#### D. Label Refinement

The purpose of label refinement is to reduce the misclassifications generated by local nature during foreground initialization, particularly near the edges of moving targets. An energy minimization formulation is thus introduced to enforce spatial consistency of targets. Given that superpixels [29] are usually achieved through a conservative strategy to ensure highly local compactness, our framework advocates them as basic units in the spatiotemporal graph  $\mathcal{G}$ . This setting, an operation at a middle level, may sacrifice some accuracy compared with pixel-by-pixel optimization, but it is still feasible and faster than pixel-wise labeling because the number of nodes in  $\mathcal{G}$  has been reduced by two orders of magnitude. Here, our ICE is still used in unary potentials to maintain the ability to capture more structural information about object perception.

Given a video sequences  $\mathcal{F} = \{F^i\}_{i=1}^K$ , we formulate  $\mathcal{F}$  as a spatiotemporal graph  $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ , with an initial node label set  $\mathcal{L} = \{l_p\}_{p=1}^N$ ,  $l_p \in \{0, 1\}$ . Here,  $\mathcal{S}$  is the collection of superpixels/nodes and  $\mathcal{E}$  is the set of edges.

Then the energy in the spatiotemporal graph  $\mathcal{G}$  is defined as

$$\begin{aligned} \mathbf{E}(\mathcal{L}) = & \sum_{p \in \mathcal{G}} \mathcal{U}_p(l_p) + \lambda_1 \sum_{p \in \mathcal{G}} \sum_{q \in \mathcal{N}_s(p)} \mathcal{V}_{pq}(l_p, l_q) \\ & + \lambda_2 \sum_{p \in \mathcal{G}} \sum_{r \in \mathcal{N}_t(p)} \mathcal{V}_{pr}(l_p, l_r) \quad (13) \\ \text{s.t. } & l_p, l_q, l_r \in \{0, 1\}, \end{aligned}$$

where  $\mathcal{N}_s(p)$  and  $\mathcal{N}_t(p)$  represent the spatial neighbors and temporal neighbors of node  $p$ , respectively. All of the superpixels in the same frame that spatially connected to node  $p$  consist of  $\mathcal{N}_s(p)$ , whereas all of the superpixels in the forward or backward adjacent frames that overlap with  $p$  comprise  $\mathcal{N}_t(p)$ .

There exist three terms in Eq.(13). Unary potential  $\mathcal{U}_p$ , also called a data term, represents the likelihood of node  $p$  belonging to label 0 or 1. Pairwise potential  $\mathcal{V}$ , also called a smooth term, includes spatial pairwise potential  $\mathcal{V}_{pq}$  in the same frame and temporal pairwise potential  $\mathcal{V}_{pr}$  across adjacent frames. Below we detail the definitions of  $\mathcal{U}$  and  $\mathcal{V}$ .

1) *Unary Potential  $\mathcal{U}$* : The first term in unary potentials is calculated using the ICE maps  $\mathcal{M} = \{\mathbf{M}^i\}_{i=1}^K$ .

The ICE map for node  $p$  is transformed into unary values by

$$\mathcal{U}_1(l_p) = \begin{cases} -\log(1 - \mathbf{M}^i(p)) & \text{if } l_p = 0, \\ -\log(\mathbf{M}^i(p)) & \text{otherwise} \end{cases} \quad (14)$$

where  $\mathbf{M}^i(p)$  is the normalized summation value of node  $p$  in  $\mathbf{M}^i$ . The second data term, Eq.(14), aims to penalize the

initial background superpixels that own large likelihood values in ICE, and the initial foreground nodes with small values in ICE.

As complementary, the weighted histograms  $\mathcal{H} = \{H_p\}_{p=1}^N$  from each superpixel/node  $p$  are also employed.  $H_p$  is a feature pool that concatenates three types of histograms: the Bag of Words (BoW) histograms respectively projected by a dense SIFT dictionary and an RGB value dictionary, color name descriptors as described in the appearance-constrained motion, and the concatenation of simple histograms in four color spaces including RGB, HSV, LAB, and YCbCr. Given that histograms of node  $p$  may be sparse, we employ the  $k$  nearest spatial neighbors of  $p$  to weight it through a Gaussian kernel and achieve a weighted  $H_p$ .

$$\mathcal{U}_2(l_p) = \begin{cases} 1 - \mathcal{D}(H_p, H(Q_{fg})) & \text{if } l_p = 0, \\ 1 - \mathcal{D}(H_p, H(Q_{bg})) & \text{otherwise} \end{cases} \quad (15)$$

where  $Q_{bg}$  and  $Q_{fg}$  are sets of initial background and foreground superpixels, respectively.  $\mathcal{D}$  represents the Bhattacharyya distance of the feature vectors between node  $p$  and  $Q_{fg}$  or  $Q_{bg}$ . The purpose of formulating the first data term as Eq.(15) is to assign a higher unary value to background superpixels that are close to foreground objects after feature embedding, and vice versa.

Thus, the unary potential for node  $p$  is

$$\mathcal{U}_p(l_p) = \mathcal{U}_1(l_p) + \mathcal{U}_2(l_p). \quad (16)$$

2) *Pairwise Potential  $\mathcal{V}$* : To calculate the degree of agreement between two spatial or temporal adjacent nodes, we need to define feature representation and metrics for nodes. Two types of cues are used here to compute the pairwise potential: histograms  $H$  and boundary connectivity  $\mathcal{C}$ .  $H$  demonstrates the appearance and motion similarity, and  $\mathcal{C}$  depicts the spatial closeness of adjacent nodes.

According to visual perception, two superpixels/nodes  $p$  and  $q$  are likely to be intimate and compact if they connect much with each other. In this case, a large percentage of boundary pixels for  $p$  and  $q$  are overlapped. The boundary connectivity value  $\mathcal{C}$  is thus defined as

$$\mathcal{C}(p, q) = \begin{cases} 0, & \text{if } l_p = l_q \\ \frac{Len(p) \cap Len(q)}{\min(Len(p), Len(q))} & \text{if } l_p \neq l_q \end{cases} \quad (17)$$

where  $Len(p)$  denotes the perimeter of superpixel  $p$  and  $Len(p) \cap Len(q)$  is the overlapped length of their perimeters.

Given two nodes  $p$  and  $q$ , s.t.  $q \in \mathcal{N}_s(p)$ , spatial pairwise potential  $\mathcal{V}_{pq}$  is written as

$$\mathcal{V}_{pq}(l_p, l_q) = \begin{cases} 0, & \text{if } l_p = l_q \\ 1 - \mathcal{D}(H_p, H_q) + \mathcal{C}(p, q), & \text{if } l_p \neq l_q \end{cases} \quad (18)$$

According to node  $p$  and  $r \in \mathcal{N}_t(p)$ , temporal pairwise potential  $\mathcal{V}_{pr}$  can be expressed as

$$\mathcal{V}_{pr}(l_p, l_r) = \begin{cases} 0, & \text{if } l_p = l_r \\ 1 - \mathcal{D}(H_p, H_r), & \text{if } l_p \neq l_r \end{cases} \quad (19)$$

Eqs.(18) and (19) aim to penalize adjacent nodes that are assigned with different initial labels.

Based on Eqs.(13), (16), (18), and (19), the refined foreground labels are achieved by minimizing the objective function:

$$\mathcal{L}^* = \underset{\mathcal{L}}{\operatorname{argmin}} \mathbf{E}(\mathcal{L}). \quad (20)$$

Given that the unary and pairwise potentials in our approach are submodular, this task can be done via the graph cut algorithm.

#### IV. EXPERIMENTS

##### A. Experimental Settings

a) **Datasets:** Three benchmark datasets were employed to evaluate our method: SegTrack [9], MOVICS [30], and GaTech [31]. **SegTrack** [9] is a commonly used dataset for video object segmentation. It contains six video sequences named *birdfall*, *cheetah*, *girl*, *monkeydog*, *parachute*, and *penguin*. A pixel-level ground-truth is provided for the primary foreground object in each video. We follow the same criterion as in [10], [11], [32], where the penguin video is discarded because only one penguin is annotated among a group of penguins. **MOVICS** [30] was initially proposed for video object co-segmentation. It is a weakly supervised pipeline for segmenting objects in multiple relevant videos. The dataset has four video sets: *chickensAll*, *lionsAll*, *giraffesAll* and *tigersAll*. Each video set contains two to four videos. In each video, the authors provide the ground truth of object class labeling for five frames that are equidistantly sampled from the video. The **GaTech** video segmentation dataset [31] consists of 15 sequences, and the video length ranges from 1 second to 28 seconds.

b) **Evaluation Metrics:** With respect to quantitative analysis, the popular average per-frame pixel error [9] was used as a basic metric. The metric is defined as

$$\text{error} = \frac{\text{XOR}(FG, GT)}{K}, \quad (21)$$

where  $FG$  is the labeling results for all frames output by the segmentation approaches,  $GT$  is the ground-truth labels, and  $K$  is the total number of frames for a sequence. Given that the average per-pixel error is an absolute quantitative metric and can vary in a wide range influenced by video resolution, we supplement a normalized metric named average labeling precision

$$\text{precision} = 1 - \frac{\text{XOR}(FG, GT)}{K \cdot N_0}, \quad (22)$$

where  $N_0$  is the total number of pixels in each frame.

c) **Implementation Details:** All of the experiments were run on a computer with Intel Core i7(3.4GHz) and 8GB RAM. In our model, SLIC superpixels [29] were used with a regularizer 0.1 and a regionsize 20. The GOP algorithm [25] with default parameters was used to generating proposals due to its fast computation. The  $\lambda_1$  and  $\lambda_2$  in Eq.(13) were set to 3 and 2, respectively. Regarding the BoW dictionaries used in our experiments, the one for dense SIFT was 200-dimensional, while another for RGB values was 150-dimensional. The two dictionaries were learned using natural scene images from PASCAL VOC 2012. With respect to the concatenated

Methods	birdfall	cheetah	girl	monkeydog	parachute	Avg.
SE	447	1626	4217	1576	627	1450
ICE	379	1381	2432	951	396	942

TABLE II: Comparison of interactively constrained encoding (ICE) and separate encoding (SE) on SegTrack. The metric is average per-frame pixel error.

color histograms, 16 bins for RGB, HSV, LAB, and YCbCr were used in each channel, and thus a 192-dimensional color histogram was obtained for each node.

##### B. ICE vs. Separate Encoding

In this experiment part, we compare our ICE against the conventional separate encoding. The quantitative experimental results for five SegTrack videos are shown in Table II.

Here, we extracted the optical flow cues based on [11] and appearance saliency based on [13], and then combined them in the final stage as a likelihood map denoted as separate encoding (SE). We compare SE with ICE using the same energy model and the same parameters. In the simplified energy minimization model, the pre- and post-processing procedures are discarded, and only ICE and SE maps are respectively employed to build unary potentials. With respect to pairwise potentials, it is impossible to measure the pairwise cost of two adjacent nodes using just the two likelihood maps. For simplicity, the absolute difference between two adjacent nodes is used as their pairwise potential.

As Table II shows, ICE obviously outperforms SE under the same condition. The improvement is mainly caused by exploiting the homologous properties of multimodal cues for the same target, and use them throughout the whole processing framework.

##### C. Results and Analysis on SegTrack

Comprehensive comparisons of object segmentation results on SegTrack are demonstrated in Table III, which includes our approach and nine state-of-the-art methods. In addition to the average per-frame pixel error for each video and the whole video set, we enumerate what the basic unit is when optimizing the model. Similar to [11], we process objects at the superpixel level, whereas [10], [22] and other approaches operate at the pixel level. [9] and [17] are conducted in a supervised manner that requires manual annotation in the first frame, whereas others are not. Our approach achieves a competitive result, especially comparing with the method presented in [11], which also uses superpixels as basic units. Although sacrificing some accuracy compared with pixel-by-pixel optimization [10], [22], our approach is very feasible and efficient, as the number of units in the model is reduced by two orders of magnitude. To investigate related algorithms under a normalized metric, we also calculate the average precision using our newly defined metric in Eq.(22), which measures the percentage of correctly labeled pixels. As Table IV shows, our approach obtains an average precision of 99.4%. Under this normalized metric, we can see that performances of most approaches are very close and eligible. The qualitative results for the five sequences are also given in Figure 5.

Video	Ours	Papazoglou's[11]	Wang's[22]	Varas's[33]	Ochs's[34]	Zhang's[10]	Lee's[32]	Brox's[8]	Tsai's[9]	Chockalingam's[17]
birdfall	219	217	209	243	468	155	288	468	252	454
cheetah	834	890	796	391	1175	633	905	1968	1142	1217
girl	1512	3859	1040	1935	5683	1488	1785	7595	1304	1755
monkeydog	536	284	562	497	1434	365	521	1434	563	683
parachute	353	855	207	187	1595	220	201	1113	235	502
Avg.	587	877	503	515	1736	452	592	1926	594	791
Basic Unit	sp	sp	p	p	p	p	p	p	p	p
Supervised?	×	×	×	×	×	×	×	×	✓	✓

TABLE III: Average per-frame pixel error on SegTrack

Video	Ours	Papazoglou's[11]	Wang's[22]	Varas's[33]	Ochs's[34]	Zhang's[10]	Lee's[32]	Brox's[8]	Tsai's[9]	Chockalingam's[17]
birdfall	99.7	99.7	99.8	99.7	99.4	99.8	99.7	99.4	99.7	99.5
cheetah	98.9	98.8	99.0	99.5	98.5	99.2	98.8	97.4	98.5	98.4
girl	98.8	97.0	99.2	98.5	95.6	98.8	98.6	94.1	99.0	98.6
monkeydog	99.3	99.6	99.3	99.4	98.1	99.5	99.3	98.1	99.3	99.1
parachute	99.8	99.4	99.9	99.9	98.9	99.8	99.9	99.2	99.8	99.7
Avg.	99.4	99.1	99.5	99.5	98.3	99.6	99.4	98.1	99.4	99.2

TABLE IV: Average precision (%) on SegTrack

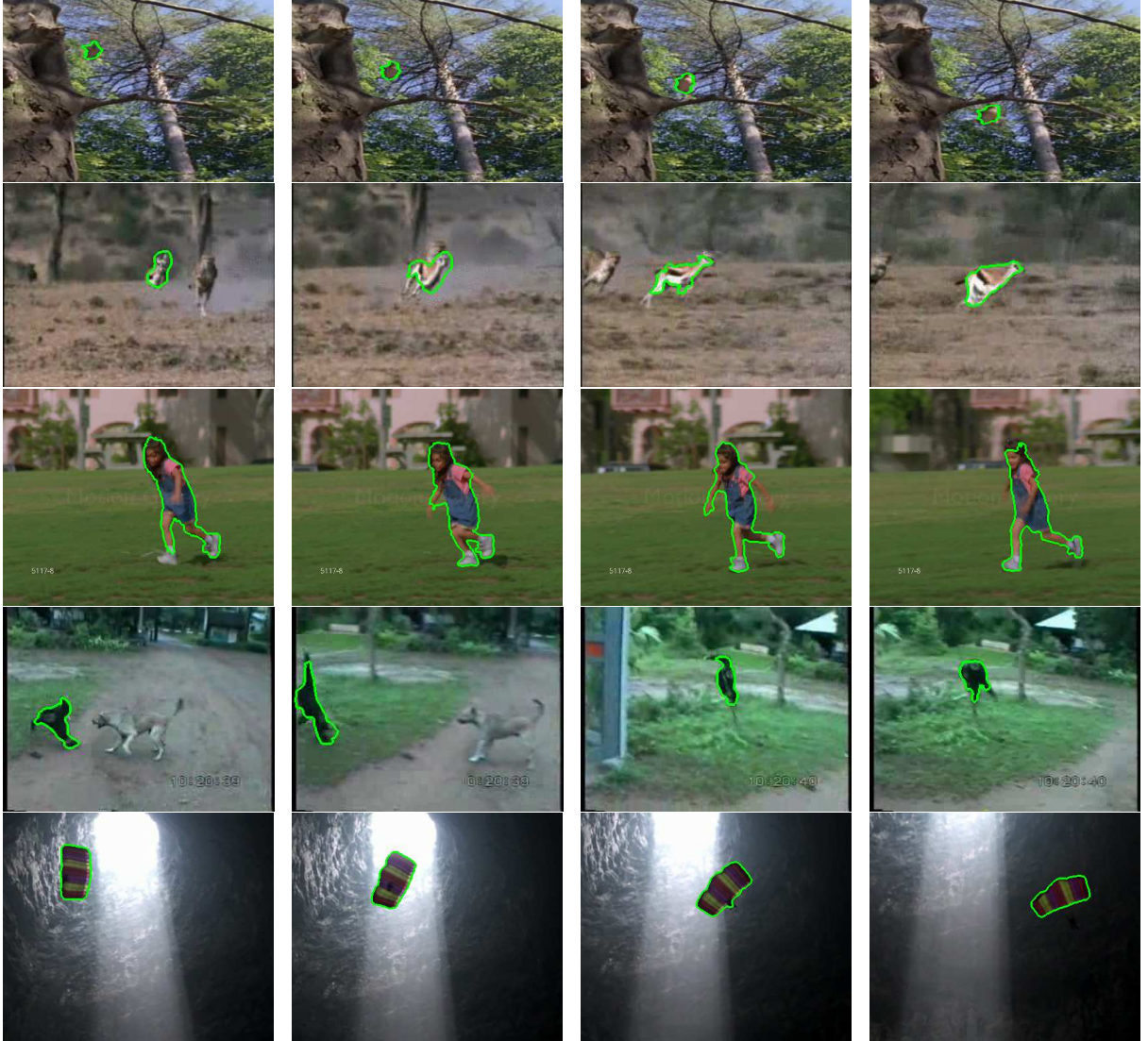


Fig. 5: Illustration of object segmentation results by our method on SegTrack. Results from top to bottom rows are respectively from videos: *birdfall*, *cheetah*, *girl*, *monkeydog*, and *parachute*. Segmented objects are delimited by green curves. Best view in color.



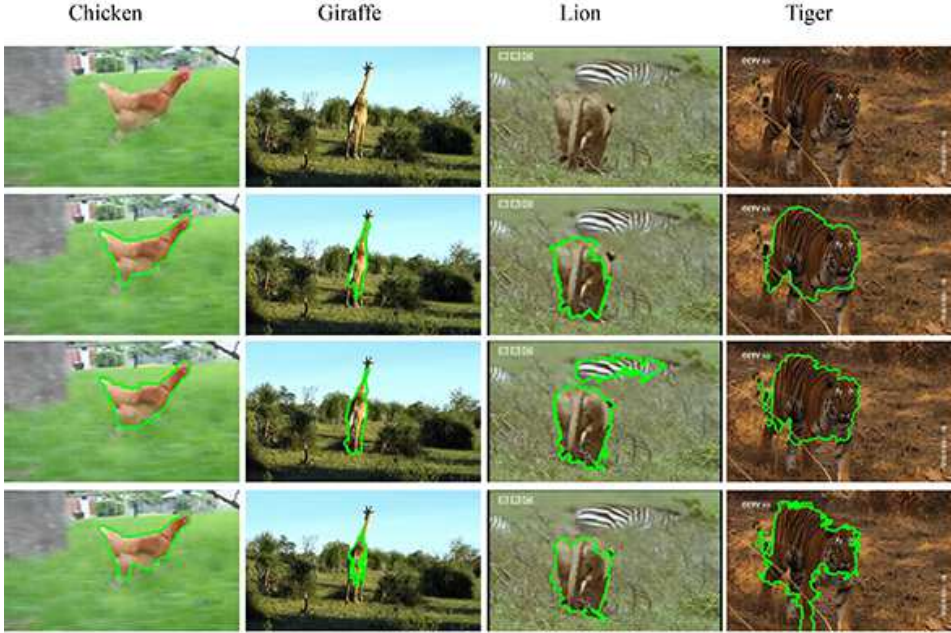


Fig. 6: Comparison of different methods on MOVICS. Four rows from top to bottom show *original frames*, *our results*, *segmentations from [11]*, and *results from [30]*, respectively. Best view in color.

Methods	Chicken	Giraffe	Lion	Tiger	Avg.	Supervised?
Ours	1437	2153	2605	14960	5289	×
Papazoglou's[11]	1762	2225	2670	17480	6034	×
Chiu's[30]	1271	1352	3250	6158	3008	✓

TABLE V: Average per-frame pixel error on MOVICS.

Methods	Chicken	Giraffe	Lion	Tiger	Avg.
Ours	98.9	98.3	96.6	93.5	96.3
Papazoglou's[11]	98.6	98.3	96.5	92.4	95.7
Chiu's[30]	99.0	99.0	95.8	97.3	97.9

TABLE VI: Average precision (%) on MOVICS.

#### D. Results and Analysis on MOVICS

There are four sets in MOVICS, each of which contains 2 to 4 relevant videos. As MOVICS [30] is originally collected for object co-segmentation, the identical object occurs repeatedly in relevant videos. However, in some videos, target objects are always nearly static, which is beyond the scope of this paper, which aims to segment moving objects. Thus, we only use one video sequence containing moving targets for each set, and four sequences in total are eventually used as test data. We compare our method against [11] and [30], running the codes provided by the authors on MOVICS. The parameter values are also the same as those set by the codes. Table V demonstrates the average per-frame pixel errors of different methods on MOVICS, and Table VI shows the average precision of the three methods. Additionally, visual comparisons of segmentation are also depicted in Figure 6.

Note that [30] is weakly supervised and requires relevant videos containing the identical targets as input. Hence, during the experiments, we still fed all of the relevant videos in MOVICS to the model in [30], but only selected the four videos that were both used by us and by [11] to report the performance. Additionally, the approach developed by [30] outputs a set of segments containing both background and foreground for each frame, and those segments do not specify which one belongs to the foreground moving target. Hence, we compare each segment to the ground-truth foreground one by one and adopt the segment with the highest accuracy as the

foreground. That means that unlike the settings in [11] and in our study, the performance of [30] in Table V is actually aided by both the ground truth and extra relevant videos.

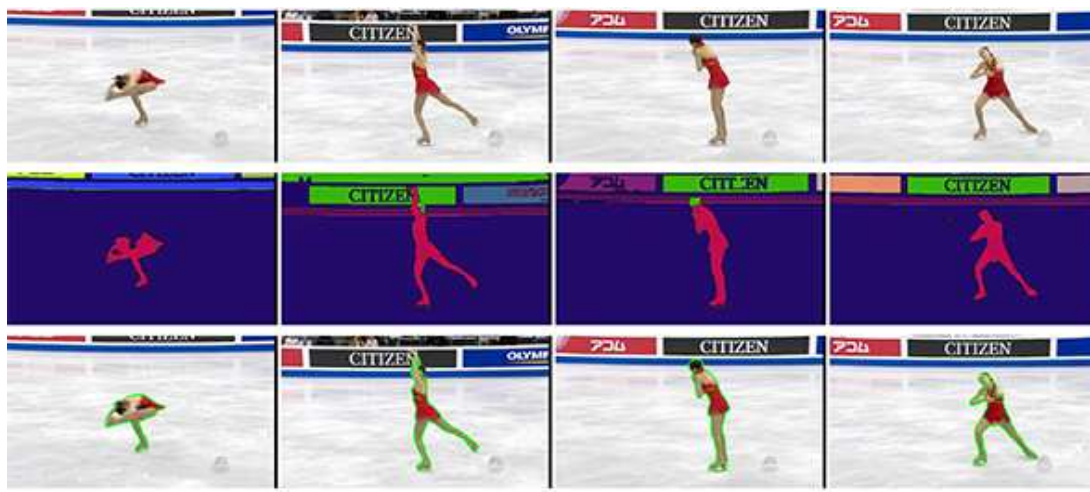
#### E. Results and Analysis on GaTech

Given that pixel-level ground truth is not offered in GaTech [9], we qualitatively compare our method with [9] on this dataset. The visual comparisons are shown in Figure 7, where our segmentations are delimited by green curves.

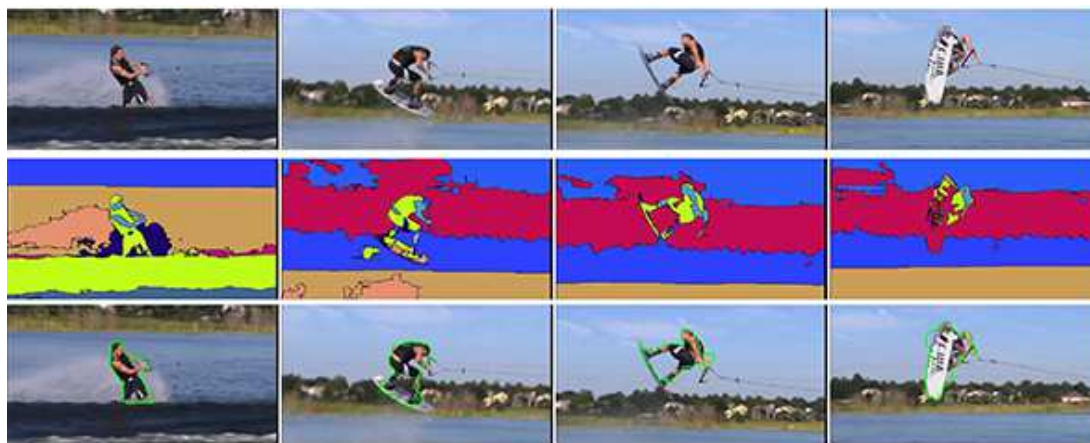
Different from our method that is unsupervised, [9] is a supervised approach that usually works well under uniform backgrounds (e.g., Yuna Kim), but can introduce a host of object fragments due to scene clutter (e.g., waterski). This is mainly because [9] does not integrate object-level cues, whereas our approach provides more insights into object-oriented information, such as the saliency and accumulated proposal maps in our ICE.

#### V. CONCLUSION

In this paper, we present an unsupervised approach for moving object segmentation in unconstrained videos. The interactively constrained encoding (ICE) is proposed to exploit the homologous properties of multimodal cues for the same object. Due to preserving interactive restrictions throughout both the initialization and refinement stages, our approach can well perceive and refine moving targets in variant environments, even under a failure of appearance saliency or optical



(a) Yuna Kim



(b) waterski

Fig. 7: Comparisons between our method and [9] on GaTech. In each subfigure, three rows from top to bottom, show *original frames*, *segmentations from [9]*, and *our results*, respectively. Best view in color.

flow. We also partially tackle the inter-occlusion problem through a conservative proposal-wise maximum bipartite graph matching. Furthermore, the lightweight superpixel-level graph optimization is developed to reduce the computation complexity. Future work involves incorporating more geometric patterns and perceptual information into the approach, and using multiple adjacent frames to generate temporal ICE.

#### ACKNOWLEDGEMENT

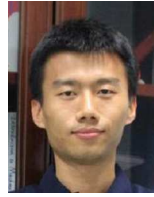
This work was supported by the NSFC (No. U1611461, 61573387, 61672544) and the Guangdong Program (No. 2015B010105005, 2015A030311047, 201604046018).

#### REFERENCES

- [1] H. Wang and D. Suter, “A consensus-based method for tracking: Modelling background scenario and foreground appearance,” *Pattern Recognition*, vol. 40, no. 3, pp. 1091–1105, 2007.
- [2] A. Colombari, A. Fusiello, and V. Murino, “Segmentation and tracking of multiple video objects,” *Pattern Recognition*, vol. 40, no. 4, pp. 1307–1317, 2007.
- [3] W.-C. Hu and J.-F. Hsu, “Automatic spectral video matting,” *Pattern Recognition*, vol. 46, no. 4, pp. 1183–1194, 2013.
- [4] M. Brand and V. Kettner, “Discovery and segmentation of activities in video,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, 2000.
- [5] C.-H. Chung, S.-C. Cheng, and C.-C. Chang, “Adaptive image segmentation for region-based object retrieval using generalized hough transform,” *Pattern recognition*, vol. 43, no. 10, pp. 3219–3232, 2010.
- [6] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999, pp. 2246–2252.
- [7] O. Barnich and M. Van Droogenbroeck, “ViBe: A universal background subtraction algorithm for video sequences,” *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.
- [8] T. Brox and J. Malik, “Object segmentation by long term analysis of point trajectories,” in *European Conference on Computer Vision (ECCV)*, 2010, pp. 282–295.
- [9] D. Tsai, M. Flagg, and J. M. Rehg, “Motion Coherent Tracking with Multi-label MRF optimization,” in *British Machine Vision Conference (BMVC)*, 2010, pp. 1–11.
- [10] D. Zhang, O. Javed, and M. Shah, “Video object segmentation through spatially accurate and temporally dense extraction of primary object regions,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 628–635.
- [11] A. Papazoglou and V. Ferrari, “Fast object segmentation in unconstrained video,” in *International Conference on Computer Vision (ICCV)*, 2013, pp. 1777–1784.
- [12] D. Giordano, F. Murabito, S. Palazzo, and C. Spampinato, “Superpixel-based video object segmentation using perceptual organization and lo-



- cation prior,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4814–4822.
- [13] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3166–3173.
- [14] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, 1997.
- [15] Y. Sheikh and M. Shah, “Bayesian object detection in dynamic scenes,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 74–79.
- [16] Y.-T. Chen, C.-S. Chen, C.-R. Huang, and Y.-P. Hung, “Efficient hierarchical method for background subtraction,” *Pattern Recognition*, vol. 40, pp. 2706–2715, 2007.
- [17] P. Chockalingam, N. Pradeep, and S. Birchfield, “Adaptive fragments-based tracking of non-rigid objects using level sets,” in *International Conference on Computer Vision (ICCV)*, 2009, pp. 1530–1537.
- [18] H. Fu, D. Xu, B. Zhang, and S. Lin, “Object-based multiple foreground video co-segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3166–3173.
- [19] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng, “Video object discovery and co-segmentation with extremely weak supervision,” in *European Conference on Computer Vision (ECCV)*, 2014, pp. 640–655.
- [20] K. Fragkiadaki, G. Zhang, and J. Shi, “Video segmentation by tracing discontinuities in a trajectory embedding,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1846–1853.
- [21] T. Lim, B. Han, and J. H. Han, “Modeling and segmentation of floating foreground and background in videos,” *Pattern Recognition*, vol. 45, no. 4, pp. 1696–1706, 2012.
- [22] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3395–3402.
- [23] X. Cao, F. Wang, B. Zhang, H. Fu, and C. Li, “Unsupervised pixel-level video foreground object segmentation via shortest path algorithm,” *Neurocomputing*, vol. 172, pp. 235 – 243, 2016.
- [24] T. Brox and J. Malik, “Large displacement optical flow: descriptor matching in variational motion estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 500–513, 2011.
- [25] P. Krähenbühl and V. Koltun, “Geodesic object proposals,” in *European Conference on Computer Vision (ECCV)*, 2014, pp. 725–739.
- [26] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, “Learning color names for real-world applications,” *IEEE Transactions on Image Processing*, vol. 18, pp. 1512–1523, 2009.
- [27] J. Kim, D. Han, Y.-W. Tai, and J. Kim, “Salient region detection via high-dimensional color transform,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 883–890.
- [28] N. Otsu, “A threshold selection method from gray-level histograms,” *Automatica*, vol. 11, pp. 23–27, 1975.
- [29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [30] W.-C. Chiu and M. Fritz, “Multi-class video co-segmentation with a generative multi-video model,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 321–328.
- [31] M. Grundmann, V. Kwatra, M. Han, and I. Essa, “Efficient hierarchical graph-based video segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2141–2148.
- [32] Y. J. Lee, J. Kim, and K. Grauman, “Key-segments for video object segmentation,” in *International Conference on Computer Vision (ICCV)*, 2011, pp. 1995–2002.
- [33] D. Varas and F. Marques, “Region-based particle filter for video object segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3470–3477.
- [34] P. Ochs and T. Brox, “Higher order motion models and spectral clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 614–621.

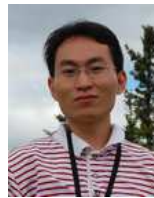


2015 National Graduate Contest on Smart-City Technology, and he was one of the winners in the 2014 Bocom Cup Contest on Video Analysis. He is a student member of CCF and IEEE.

**Chun-Chao Guo** received the B.E. degree in communication engineering with honors from Lanzhou University, China, in 2010. He is currently pursuing the Ph.D. degree in computer science at Sun Yat-sen University, China. His research interests are in computer vision and pattern recognition, with a focus on human identity recognition, object tracking, object detection and visual surveillance. Chun-Chao Guo is a recipient of the Excellent Paper Award at the 2014 National Conference on Image and Graphics. He won the first prize in the 2014 and



**Jianhuang Lai** received his M.Sc. degree in applied mathematics in 1989 and his Ph.D. in mathematics in 1999 from SUN YAT-SEN University, China. He joined Sun Yat-sen University in 1989 as an Assistant Professor, where currently, he is a Professor in School of Data and Computer Science. His current research interests are in the areas of computer vision, pattern recognition and its applications. He has published over 250 scientific papers in the international journals and conferences on image processing and pattern recognition, e.g. IEEE TPAMI, IEEE TNN, IEEE TIP, IEEE TSMC (Part B), Pattern Recognition, ICCV, CVPR and ICDM. Prof. Lai serves as a deputy director of the Image and Graphics Association of China and also serves as a standing director of the Image and Graphics Association of Guangdong. He is also the deputy director of Computer Vision Committee, China Computer Federation (CCF).



**Xiaohua Xie** is currently a Research Professor at Sun Yat-Sen University. Prior to joining SYSU, Xiaohua Xie was an Associate Professor at Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences. He received the B.Sc. in Mathematics and Applied Mathematics (2005) from Shantou University, the M.Sc. in Information of Computing Science (2007) and the Ph.D. in Applied Mathematics (2010) from Sun Yat-sen University in China (jointly supervised by Concordia University in Canada). His current research fields cover image processing, computer vision, pattern recognition, and computer graphics, especially focusing on image understanding and object modeling. He has published more than a dozen papers in the prestigious international journals and conferences. He is recognized as Overseas High-Caliber Personnel (Level B) in Shenzhen, China.