Semantic Object Segmentation via Detection in Weakly Labeled Video

Yu Zhang[1], Xiaowu Chen[1]*, Jia Li[1,2], Chen Wang[1], Changqun Xia[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University
[2]International Research Institute for Multidisciplinary Science, Beihang University

Abstract

Semantic object segmentation in video is an important step for large-scale multimedia analysis. In many cases, however, semantic objects are only tagged at video-level, making them difficult to be located and segmented. To address this problem, this paper proposes an approach to segment semantic objects in weakly labeled video via object detection. In our approach, a novel video segmentation-by-detection framework is proposed, which first incorporates object and region detectors pre-trained on still images to generate a set of detection and segmentation proposals. Based on the noisy proposals, several object tracks are then initialized by solving a joint binary optimization problem with min-cost flow. As such tracks actually provide rough configurations of semantic objects, we thus refine the object segmentation while preserving the spatiotemporal consistency by inferring the shape likelihoods of pixels from the statistical information of tracks. Experimental results on Youtube-Objects dataset and SegTrack v2 dataset demonstrate that our method outperforms state-of-the-arts and shows impressive results.

1. Introduction

Semantic video object segmentation, which aims to detect and segment object-like regions in video according to predefined object labels, is an essential step in computer vision and multimedia analysis. In many scenarios, however, objects are only labeled at video-level, making them difficult to be located and segmented. As videos tagged with only semantic labels are growing explosively on the Internet, it is necessary to explore how to segment the desired semantic objects in such weakly labeled videos.

Recently, many approaches (e.g., [12, 25, 21]) have been proposed to address this problem through weakly supervised learning. In the learning process, object-relevant instances were usually selected among videos sharing the same semantic tags, while background instances were sampled from videos with irrelevant tags. These instances, which may be inadequately selected or inaccurately labeled, were then fed into the weakly supervised learning framework (e.g., multi-instance learning [12], negative mining [25] and label transfer [21]) to train a segment classifier. Although these approaches can achieve promising results on certain scenarios, the ambiguity of training instances may lead to unexpected segmentation results. Moreover, multiple videos are required during training, preventing the usage of these approaches in single video segmentation.

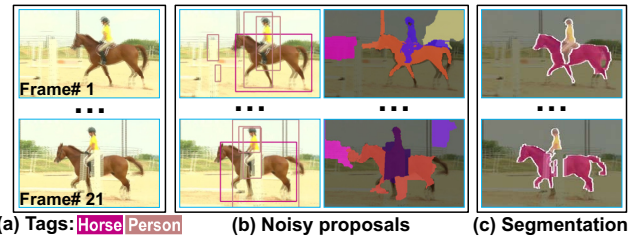

Figure 1. The motivation of our approach. (a) The input video is weakly labeled with semantic tags, making it difficult to locate and segment the desired objects (e.g., the horse is occluded by fence in frame #21); (b) Per-frame detection and segmentation proposals provide location information but are often very noisy; (c) The proposed segmentation-by-detection framework can generate consistent object segmentation results from noisy detection and segmentation proposals.

To address these problems, this paper proposes to segment semantic objects in video via detection. The proposed approach does not need segment-level training stage and thus avoids the selection of ambiguous instances. Instead, the image-based object detectors, which have demonstrated great successes in segmenting semantic objects in images [32, 17, 14, 26, 31], are employed into our segmentation-by-detection framework. In this framework, object and region detectors pre-trained on still images are first used to generate a set of detection and segmentation proposals on various video frames. As shown in Fig. 1, such image-based proposals often lack spatiotemporal consistency and may be inaccurate due to blurred boundaries generated from video compression, object occlusion and camera motion. Therefore, we propose to initialize several object tracks from

*Corresponding author (email: chen@buaa.edu.cn)

these noisy proposals by solving a joint assignment problem formulated as min-cost flow. As these tracks actually provide rough configurations of semantic objects, we thus propose to infer the shape likelihoods of pixels from the statistical information of tracks. In this manner, background noises can be suppressed and the segmentation results of desired objects can be refined while preserving the spatiotemporal consistency. To validate the effectiveness of the proposed approach, we have conducted extensive experiments on two public video benchmarks, including Youtube-Objects dataset and SegTrack v2 dataset. Experimental results show that our approach outperforms several state-of-the-art weakly-supervised and unsupervised approaches on challenging object categories. The main contributions of the proposed approach are summarized as follows:

1) A novel segmentation-by-detection framework is proposed for semantic object segmentation in weakly labeled video, and demonstrates impressive performance on two public video benchmarks.

2) We present an algorithm to initialize object tracks from image-based detection and segmentation proposals. By solving a joint assignment problem with min-cost flow, this algorithm is robust to noisy proposals.

3) We refine the object tracks by inferring shape likelihoods from the statistical information of tracks, which can effectively suppress background noise while preserving the spatiotemporal consistency of foreground objects.

2. Related Work

Video object segmentation approaches in the current literature can be grouped into the (semi-)supervised, unsupervised and weakly supervised ones.

Supervised and semi-supervised approaches typically act through training label classifiers [20, 27] or propagating user-annotated labels over time [13, 29, 28, 2]. Although being well studied in a long period, such methods are limited to a small range of applications for its extreme dependence on labor-intensive pixel annotations to train suitable models.

Unsupervised approaches generally focus on segmenting the most primal object [8, 18, 19, 22] in a single video and co-segmenting the common object among a video collection [9, 11, 30]. As several recent successes, Lee et al. [18] attempted to segment the foreground object through identifing the *key segments* of highly salient appearance and motion in the video. Dong et al. [8] proposed to densely extract object segments with high objectness and smooth evolvement based on *directed acyclic graph*. Papazoglou et al. [22] developed a fast object segmentation approach that quickly estimates rough object configurations through the use of *inside-outside maps*.

Weakly supervised approaches have received growing attention for its convenience in gathering video-level labels and the prospect in analyzing web-scale data. Existing algorithms employed variants on the learning techniques to predict the confidence of each pixel belonging to a given concept. Hartmann *et al.* [12] first addressed it by training large-scale segment classifiers. Tang *et al.* [25] compared the segments in positive videos with a large collection of negative examples and identified those of distinct appearance as foreground. The study was further pushed forward by Xiao *et al.* [21] for handling this problem in multi-class criterion as opposed to traditional binary classification.

A common issue affecting the performance of weakly supervised approaches is the learning procedure with ambiguous training labels (i.e. locations of target objects). Different from these methods, our approach addresses video segmentation with weak labels through leveraging image-based object detectors and avoids such a procedure. Detection-based approaches have been widely studied on image segmentation [16, 31, 14, 7, 17, 26, 32]. For example, Wei et al. [31] utilized detectors to guide semantic object segmentation in images without any pixel-level training stage. Inspired by these successes, this paper proposes to segment semantic objects in weakly labeled videos via object detection, which still receives less attention in the literature.

3. The Approach

As shown in Fig. 2, our framework consists of three components. The first one generates detection and segmentation proposals for each frame. In the second one, several object tracks are initialized from the proposals. These tracks are finally refined in the third component to obtain fine-grained segmentation. In the following subsections, we provide the technical details for each of the three components.

3.1. Detection/Segmentation Proposal Generation

For a video with T frames, we first extract a set of object detection and segmentation proposals for each video frame. Taking the tth frame for example, such proposals can be extracted as follows:

Detection Proposals. We use the part-based detector [10] trained on PASCAL VOC 2007 dataset to generate a set of detections $\mathbb{D}_t^+$. For each detection $\mathcal{D} \in \mathbb{D}_t^+$, we denote its original response as $\bar{r}(\mathcal{D})$, and the transformed response with $r(\mathcal{D}) = 1/(1+\exp(-\frac{1}{1.5\lambda_d}(\bar{r}(\mathcal{D})-\gamma)))$, where γ denotes the calibrated threshold using a few object bounding boxes manually annotated by [23], and λ_d is the average value of all $(\bar{r}(\mathcal{D})-\gamma)$. To account for missing detection, we additionally introduce a dummy detection variable $\mathcal{D}_\phi$ for the frame. Thus, the set of detection proposals on the tth frame is given by $\mathbb{D}_t = \mathbb{D}_t^+ \cup \{\mathcal{D}_\phi\}$.

Segmentation proposals. We use a motion-aware version [3] of the region detector [4] to generate segmentation

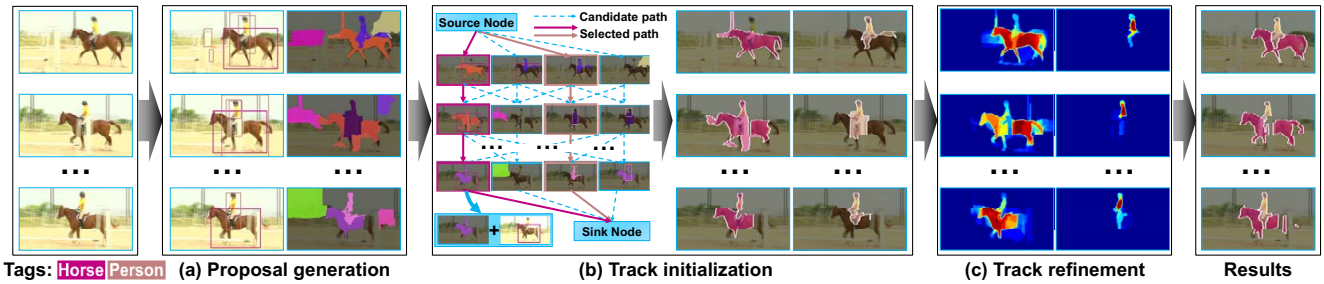

Figure 2. The system framework of our approach consists of three major components, including: **(a) Proposal generation:** For each frame, a set of detection and segmentation proposals are generated; **(b) Track initialization:** Given the noisy proposals, several object tracks are initialized; **(c) Track refinement:** Each of the track is further refined to obtain fine-grained segmentation.

proposals $\mathbb{S}_t$ for the tth frame. In this procedure, additional foreground seeds are placed on the image area where detections are densely overlapped using a low threshold. For each segmentation $\mathcal{S} \in \mathbb{S}_t$, we use $\mathbf{f}(\mathcal{S})$ and $o(\mathcal{S})$ to represent its feature vector and objectness score, respectively. The feature vector consists of Hue color histogram and texton histogram quantized into 64 bins. Objectness is defined as combination of the score produced by region detector and magnitude of boundary motions, as suggested by [8].

For the sake of simplicity, we give here the *intersection-over-union overlap* of two regions (detection or segmentation) $\mathcal{R}_1$ and $\mathcal{R}_2$ as $\pi(\mathcal{R}_1, \mathcal{R}_2) = \frac{|\mathcal{R}_1 \cap \mathcal{R}_2|}{|\mathcal{R}_1 \cup \mathcal{R}_2|}$, where $|\mathcal{R}|$ represents the number of pixels in $\mathcal{R}$. Moreover, if $\mathcal{R}_1$ or $\mathcal{R}_2$ equals $\mathcal{D}_\phi$, we define $\pi(\mathcal{R}_1, \mathcal{R}_2) = 1$.

3.2. Track Initialization with Min-cost Flow

Given the generated noisy proposals, we wish to find K object tracks passing through reliable detections and segmentations that best cover the semantic objects. This can be achieved by jointly assigning detections and segmentations to tracks. To this end, we define the binary variables

$$\mathbb{A} = \{a_{\mathcal{D}}^k | \forall k, t, \mathcal{D} \in \mathbb{D}_t\}, \quad \mathbb{B} = \{b_{\mathcal{S}}^k | \forall k, t, \mathcal{S} \in \mathbb{S}_t\}, \quad (1)$$

where $a_{\mathcal{D}}^k, b_{\mathcal{S}}^k \in \{0, 1\}$ represent the assignment of detection $\mathcal{D}$ and segmentation $\mathcal{S}$ to the kth track, respectively. For a track, we assume that 1) it selects at most one segmentation or detection on a frame, 2) its selected segmentation should be coupled with a detection (or the dummy detection), 3) it cannot be overlapped with other tracks. With these assumptions in mind, we optimize $\mathbb{A}$ and $\mathbb{B}$ by solving

$$\begin{aligned} &\min_{\mathbb{A},\mathbb{B}} \mathcal{L}(\mathbb{A}, \mathbb{B}) + \lambda_1 \Omega_1(\mathbb{A}, \mathbb{B}) + \lambda_2 \Omega_2(\mathbb{B}), \\ &\text{s.t. } a_{\mathcal{D}}^k, b_{\mathcal{S}}^k \in \{0, 1\}, \ \forall k, t, \mathcal{D} \in \mathbb{D}_t, \mathcal{S} \in \mathbb{S}_t, \\ &\sum_{\mathcal{D} \in \mathbb{D}_t} a_{\mathcal{D}}^k = \sum_{\mathcal{S} \in \mathbb{S}_t} b_{\mathcal{S}}^k \leq 1, \ \forall k, t, \\ &\sum_k a_{\mathcal{D}}^k \leq 1, \sum_k b_{\mathcal{S}}^k \leq 1, \ \forall t, \mathcal{D} \in \mathbb{D}_t, \mathcal{S} \in \mathbb{S}_t, \end{aligned} \quad (2)$$

where $\mathcal{L}$ is the loss function, Ω_1 and Ω_2 penalize the tracking inconsistency and track interactions, respectively. We set the weights $\lambda_1 = 10$ and $\lambda_2 = 10^3$. Moreover, we assume that a track is consecutive and cannot be empty, leading to the additional constraints that $\forall k, t, p, q$,

$$\left(\sum_{\mathcal{D} \in \mathbb{D}_{t-p}} a_{\mathcal{D}}^k\right)\left(1 - \sum_{\mathcal{D} \in \mathbb{D}_t} a_{\mathcal{D}}^k\right)\left(\sum_{\mathcal{D} \in \mathbb{D}_{t+q}} a_{\mathcal{D}}^k\right) = 0, \quad (3)$$

$$\sum_t \sum_{\mathcal{D} \in \mathbb{D}_t} a_{\mathcal{D}}^k = \sum_t \sum_{\mathcal{S} \in \mathbb{S}_t} b_{\mathcal{S}}^k \geq 1. \quad (4)$$

Note that for simplicity, we omit the range of indices k, t, p and q in the formulation.

The loss term $\mathcal{L}$ sums over the per-proposal losses:

$$\begin{aligned} &\mathcal{L}(\mathbb{A}, \mathbb{B}) = -\sum_{t,k} \sum_{\mathcal{D} \in \mathbb{D}_t} \sum_{\mathcal{S} \in \mathbb{S}_t} \xi(\mathcal{D}, \mathcal{S}) a_{\mathcal{D}}^k b_{\mathcal{S}}^k, \text{ where} \\ &\xi(\mathcal{D}, \mathcal{S}) = \begin{cases} \pi(\mathcal{D}, \mathcal{S}) \log \frac{r(\mathcal{D}) o(\mathcal{S})}{1 - r(\mathcal{D}) o(\mathcal{S})}, & \text{if } \mathcal{D} \neq \mathcal{D}_\phi \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (5)$$

Intuitively, it favors choosing detection (segmentation) with high response (objectness). Moreover, the negative log-likelihood is scaled by the overlap, enforcing that the selected segmentation be spatially close to the detection.

The tracking cost Ω_1 constrains assignments in neighboring frames by imposing temporal smoothness:

$$\Omega_1(\mathbb{A}, \mathbb{B}) = \sum_{t,k} \sum_{\substack{\mathcal{D} \in \mathbb{D}_t \\ \mathcal{S} \in \mathbb{S}_t}} \sum_{\substack{\mathcal{D}_0 \in \mathbb{D}_{t+1} \\ \mathcal{S}_0 \in \mathbb{S}_{t+1}}} a_{\mathcal{D}}^k b_{\mathcal{S}}^k a_{\mathcal{D}_0}^k b_{\mathcal{S}_0}^k \eta(\mathcal{D}, \mathcal{D}_0, \mathcal{S}, \mathcal{S}_0),$$

$$\text{where } \eta(\mathcal{D}, \mathcal{D}_0, \mathcal{S}, \mathcal{S}_0) = \frac{\chi^2(\mathbf{f}(\mathcal{S}), \mathbf{f}(\mathcal{S}_0))}{1 + \pi(\bar{\mathcal{D}}, \mathcal{D}_0) \cdot \pi(\bar{\mathcal{S}}, \mathcal{S}_0)}, \quad (6)$$

where $\chi^2(\cdot)$ is the *chi-square* distance, $\bar{\mathcal{D}}$ and $\bar{\mathcal{S}}$ are the warped area of $\mathcal{D}$ and $\mathcal{S}$ by optical flows [6], respectively. The formulation (6) considers both appearance and shape similarity to measure the evolvement of objects over time.

The interaction cost Ω_2 models pairwise relationships among tracks by summing per-segment mutual overlaps:

$$\Omega_2(\mathbb{B}) = \sum_t \sum_{k \neq k_0} \sum_{\mathcal{S}, \mathcal{S}_0 \in \mathbb{S}_t} b_{\mathcal{S}}^k b_{\mathcal{S}_0}^{k_0} \pi(\mathcal{S}, \mathcal{S}_0), \quad (7)$$

Incorporating (3)-(7) into (2), we obtain a constrained combinatorial problem which is very hard to optimize directly. However, after substituting the variables $\delta_x(\mathcal{D},\mathcal{S}) = \sum_k a^k_{\mathcal{D}} b^k_{\mathcal{S}}$, $\delta_y(\mathcal{D},\mathcal{D}_0,\mathcal{S},\mathcal{S}_0) = \sum_k a^k_{\mathcal{D}} a^k_{\mathcal{D}_0} b^k_{\mathcal{S}} b^k_{\mathcal{S}_0}$ into the objective, we can rewrite (2) with a compact form

$$-\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\delta}_x + \lambda_1 \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\delta}_y + \lambda_2 \boldsymbol{\delta}_x^{\mathrm{T}}\boldsymbol{\Pi}\boldsymbol{\delta}_x, \tag{8}$$

where $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, $\boldsymbol{\delta}_x$ and $\boldsymbol{\delta}_y$ are column vectors that concatenate variables ξ, η, δ_x and δ_y, respectively. The matrix $\boldsymbol{\Pi}$ records segment overlaps with its element $\boldsymbol{\Pi}_{ij}$ denoting the overlap between segmentations determined by the ith and jth component of $\boldsymbol{\delta}_x$.

The objective (8) implies a *k-individual-flow* whose nodes represent segmentation-detection pairs, with flow variables $\boldsymbol{\delta}_x$ and $\boldsymbol{\delta}_y$ denoting activation of nodes and edges. The second set of constraints in (2) together with the constraints (3) and (4) indicate standard *flow conservation* and *requirements* constraints [1]. However, the last set of constraints in (2) require each segmentation or detection be claimed by at most one track, which in turn break the *total unimodularity* [1] of the constraint matrix and make the general solvers inapplicable. Intuitively, Lagrangian relaxation can be used to address this problem by relaxing the side constraints and solving a sequence of pairwise flow sub-problems [5]. However, the computational cost will become very high due to the existence of massive variables that need to be optimized in (8). Therefore, we develop a two-step heuristic algorithm to minimize (8) that considers both performance and efficiency. Several independent tracks are efficiently initialized in the first step, followed by a slower refinement step that takes the mutual overlaps of tracks into consideration.

Step 1 (initialization). In this step, several tracks with high confidence are initialized by dropping the quadratic term and minimizing $\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\delta}_x + \lambda_1 \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\delta}_y$ while satisfying all constraints. When the track number $K = 1$, it degenerates to finding the minimum-cost flow and is solvable in polynomial time. Thus, we greedily establish a flow at one time, while preserving previously obtained flows. It repeats until inclusion of any new flows breaks the constraint or increases the objective value. Flows passing over less than 5 frames are considered unreliable and omitted.

Step 2 (local search). The initial solution results in a set of tracks with high detection response, while may have poor relative region structure for neighboring objects. For each initial flow, we further sample a set of low-cost candidate flows passing through the identical detections but various segmentations by *k-shortest-path* algorithm. After that, we select one flow from each of the K candidate set to minimize (8). The interaction cost of two flows sharing the same segmentation or detection is set to $+\infty$ to satisfy the constraints. This is a classic multi-labeling problem and can be handled by binary quadratic programming.

The above steps assume that the track number K is already given. To determine the optimal K^*, we first estimate the largest K_{max} via the first step, and then perform local search for each possible K. The optimal K^* is chosen as the one with the minimal objective value.

Note that in some extreme cases, reliable detections cannot be obtained, i.e., the first step fails to get a feasible solution. To handle it, we set the values of the terms r, π defined on detections to one, and build a *directed acyclic graph* proposed in [8] to initialize a single track for the video.

3.3. Track Refinement using Shape Likelihoods

The initial tracks can roughly locate the semantic objects in the video, but may have inconsistent appearance in different frames. For each track, we further improve its visual coherence by estimating spatiotemporally consistent shape likelihoods. Instead of pixel-level inference, the shape likelihoods are inferred by optimizing object-level statistics of multiple segment tracks that overlap with the initial track, through which long-range object properties such as recurrence along time can be leveraged.

Given an initial track, we first retain the pre-generated segments with more than 50% area overlapped with it on each frame. After that, a series of N individual tracks are extracted by linking the retained segments across frames through greedy tracking [19, 9]. Denote the i-th track with length L_i as $\{\mathcal{S}_{i,l}\}_{l=1}^{L_i}$, its initial score α_i^0 is computed using

$$\alpha_i^0 = \sum_{l=1}^{L_i} o(\mathcal{S}_{i,l}) + \sum_{l=1}^{L_i-1} e^{-\frac{1}{\lambda_f}\chi^2(\mathbf{f}(\mathcal{S}_{i,l}),\mathbf{f}(\mathcal{S}_{i,l+1}))}, \tag{9}$$

and normalized, where o, $\mathbf{f}$, and χ^2 are defined in Sect. 3.2, λ_f is the mean of the *chi-square* distances of all samples. Since these scores may be trapped to background noises, we propose to adjust them with a set of new scores $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^N$ through solving the following objective

$$\min_{0\preceq\boldsymbol{\alpha}\preceq 1} \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^0\right)^2 + \theta_1 \mathcal{C}_1(\boldsymbol{\alpha}) + \theta_2 \mathcal{C}_2(\boldsymbol{\alpha}), \tag{10}$$

where the first term penalizes deviation of the adjusted scores from initial estimation, the other two terms account for appearance and temporal consistency, respectively. Parameters $\theta_1 = \theta_2 = 10$ control the relative weights.

Appearance consistency. This term allows shape confidence of a track to be passed to its visually similar neighbors. The neighbor set of the ith track is denoted as $\mathbb{R}_i$, which indexes the tracks that have distance below a threshold, and do not share co-existing time with the ith track. The *Hausdorff Distance* in *chi-square* space of segment features is used to compute track distance. With this setup, we have

$$\mathcal{C}_1(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \sum_{j\in\mathbb{R}_i} w_{i,j} \left(\alpha_i - \alpha_j\right)^2, \tag{11}$$

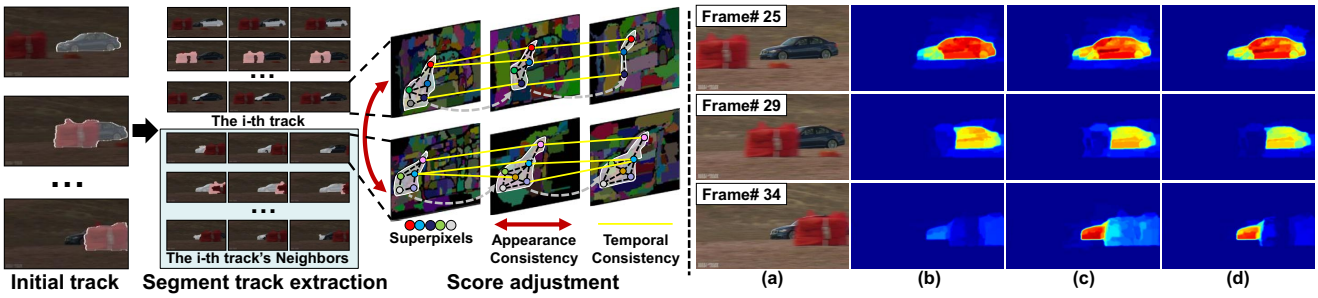

Figure 3. Shape likelihood estimation. Left: The non-local appearance consistency passes shape confidence among visually similar tracks, while the local temporal consistency forces the likelihoods between adjacent frames to change smoothly.mutli instance segmentation Right: Estimated shape likelihoods of (a) a car sequence with (b) initially obtained scores, (c) appearance consistency and (d) both appearance and temporal consistency.

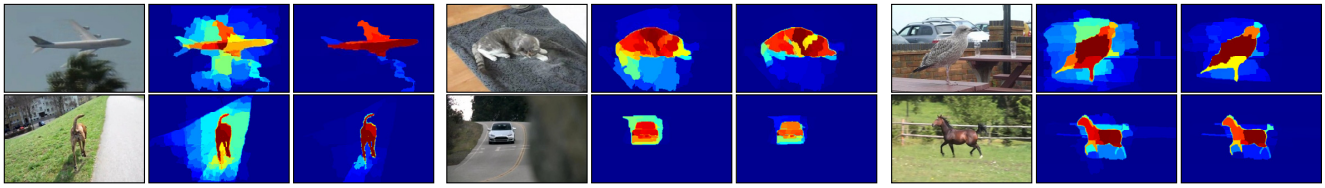

Figure 4. Estimated shape likelihood maps (best viewed in color). For each example, the original image, shape likelihoods before and after score adjustment are illustrated from left to right, respectively.

where $w_{i,j} = e^{-dist(i,j)}$ weights the similarity of the ith and jth track, with $dist(\cdot)$ denoting the distance function.

Temporal consistency. We enforce temporally adjacent pixels to have similar likelihoods. To do that, we first break down the initial track into a set of P non-overlapping superpixels $\{A_p\}_{p=1}^{P}$ by computing *multi-intersections* [19] of pre-generated segmentations on each frame. The shape likelihood β_p of a superpixel A_p is then estimated by average voting, i.e. $\beta_p = \frac{1}{N}\sum_{i\in\mathbb{I}_p}\alpha_i$, where $\mathbb{I}_p$ denotes the indices of tracks that cover A_p. Directly optimizing on superpixel likelihoods, we have

$$\mathcal{C}_2(\boldsymbol{\alpha}) = \sum_{d\in\{1,2\}}\sum_{p=1}^{P}(|A_p|\beta_p - \sum_{q\in\mathbb{M}_p^d}\omega_{q,p}^d|A_q|\beta_q)^2, \quad (12)$$

where $\mathbb{M}_p^1$ ($\mathbb{M}_p^2$) indexes the superpixels that are mapped to A_p by forward (backward) flows, $\omega_{q,p}^1$ ($\omega_{q,p}^2$) denotes the proportion of a superpixel $A_q \in \mathbb{M}_p^1$ ($\mathbb{M}_p^2$) that is mapped to A_p, and $|A_p|$ is the area of A_p. Inspired by [19], we force the likelihood of a superpixel be close to the sum of that of the mapped superpixels in different proportions.

The effects of appearance and temporal consistency are illustrated in Fig. 3(a)-(d). Likelihoods of the car drop dramatically after strong occlusion, but are recovered through passing the scores of previously appeared tracks. Incorporating temporal consistency further suppresses background confidence and lead to more consistent estimation.

The objective (10) is smooth and convex, and thus can be efficiently solved using *L-BFGS* algorithm. After obtaining the optimal $\boldsymbol{\alpha}^*$, pixel-level likelihoods are computed through average voting of track scores and normalized to $[0, 1]$ (see Fig. 4 for some estimated shape likelihoods).

Finally, a refinement procedure similar with that of [31] via graph-cut is adopted for each frame. In the energy function, the unary term combines the shape likelihoods and the color-induced evidence produced by GMMs learned on the initial tracks, both weighted by 0.5. The details are omitted here and referred to to [31] due to lack of space.

The procedure works individually for each track, which may result in some intersecting regions. A simple solution is adopted to compare the feature histogram of the intersecting region with that of a small neighboring area belonging to each track using *chi-square* distance, and assign the label of the track with the smallest distance to the given area.

4. Experiments

We perform experiments on two challenging datasets: the large-scale Youtube-Objects dataset [13, 25] originally introduced by [23], and the SegTrack v2 dataset [19, 28]. The Youtube-Objects dataset consists of 126 videos from 10 object classes with pixel-level ground truth [13] for 1 in every 10 frames. On this dataset, we test on the first 100[1] frames (sampled every other frame) for each video, resulting in more than 5500 frames in total. We use the standard *intersection-over-union overlap* as accuracy metric. The

[1]Note that some videos in the Youtube Objects dataset are insufficiently annotated by [13]. For instance, some videos only have the first or the last several frames labeled, while some others may have very sparse annotations. Therefore, we uniformly select the first 100 frames from each video and manually labeled the missed frames by [13] for 1 in every 10 frames.

Table 1. Intersection-over-union overlap accuracies on Youtube-Objects Dataset.

Method	Plane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Cls. Avg.	Vid. Avg.
Tang et al. [25]	0.178	0.198	0.225	0.383	0.236	0.268	0.237	0.140	0.125	0.404	0.239	0.228
Papazoglou et al. [22]	0.674	**0.625**	0.378	0.670	0.435	0.327	0.489	0.313	0.331	**0.434**	0.468	0.432
Baseline-INT	0.641	0.472	0.321	0.511	0.369	0.455	0.472	0.520	0.349	0.279	0.439	0.434
Baseline-GC	0.733	0.609	0.438	0.619	0.420	**0.550**	0.534	0.537	0.419	0.304	0.516	0.507
Baseline-SG [31]	0.673	0.621	**0.488**	0.695	0.441	0.497	0.519	0.457	0.386	0.314	0.509	0.494
Ours	**0.758**	0.608	0.437	**0.711**	**0.465**	0.546	**0.555**	**0.549**	**0.424**	0.358	**0.541**	**0.526**

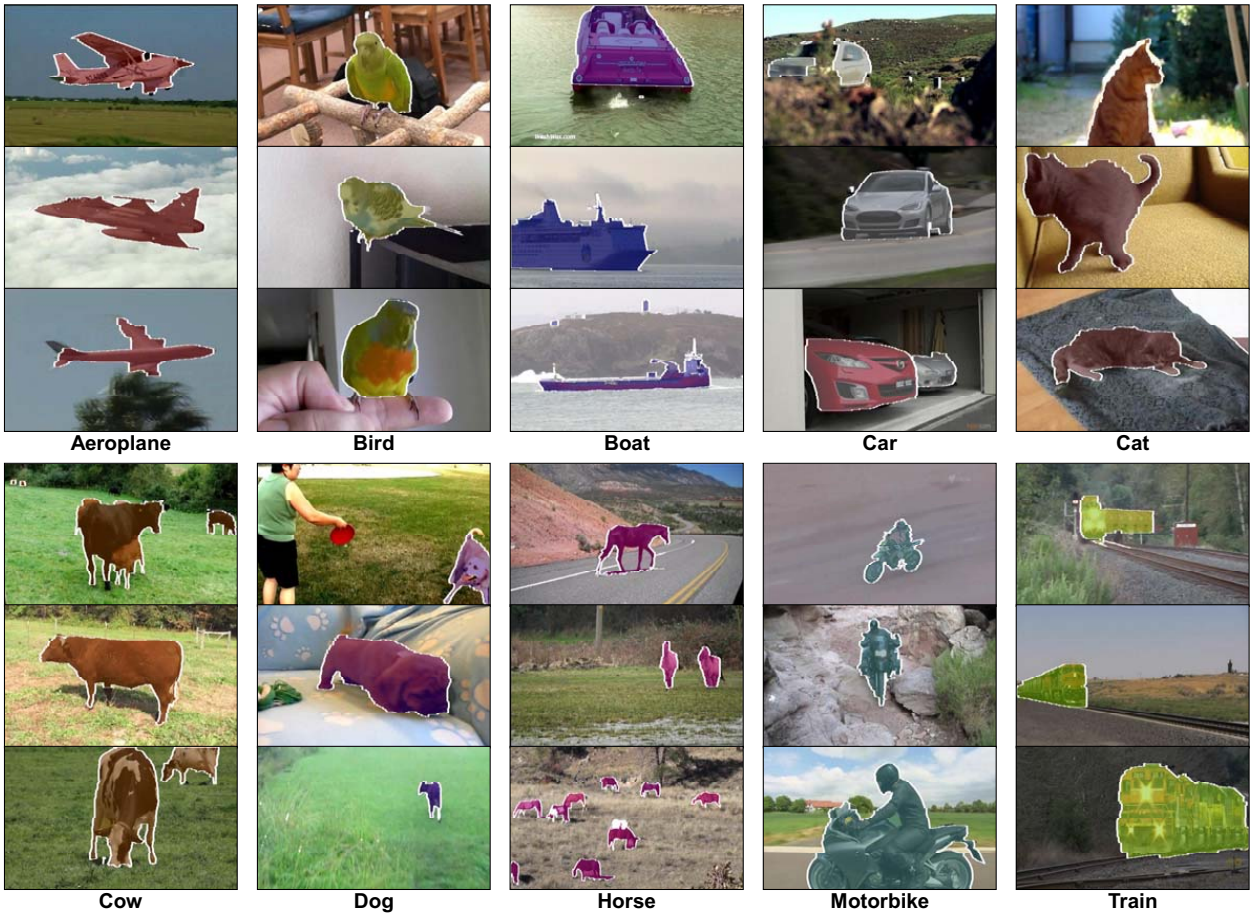

Figure 5. Representative successful examples generated by our approach on Youtube-Objects dataset. Results are overlaid on images with colored regions with white boundaries. Different colors represent different semantic categories. Best viewed in color.

Segtrack v2 dataset consists of 14 videos with single object or interacting objects presented in each video. We use 8 videos containing a single object to test the performance of our approach on unsupervised segmentation. The *per-frame pixel error rate* [28] is adopted as the evaluation criteria.

4.1. Results on Youtube-Objects Dataset

Settings. We compare our approach with two state-of-the-arts on this dataset, including the weakly supervised approach [25] and the unsupervised approach [22]. Several baselines are also implemented to evaluate different parts of the proposed approach: Baseline-INT refers to the initialization procedure in Sect. 3.2 without further optimization, Baseline-GC refers to GrabCut [24] bootstrapped by bounding boxes of the initial segments to isolate the shape likelihoods. To compare with a per-frame-basis counterpart for shape estimation, we replace the proposed spatiotemporal shape likelihoods with the shape guidance in [31], which shows promising results on PASCAL VOC Challenges. The shape guidance is estimated individually for each frame using bounding boxes of the segments selected by our track initialization procedure. Note that it is unfair to directly compare to the image-based approach [31] since the low-quality detections on videos will decrease the performance.

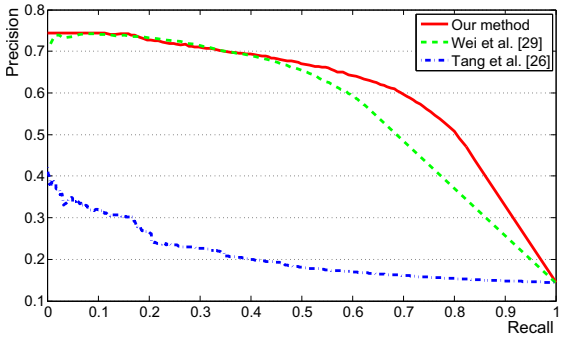

Figure 6. P/R curves (averaged on videos) computed by binarizing the shape likelihoods of different methods.

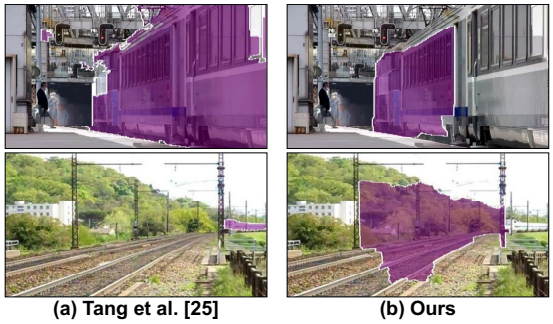

Figure 7. Some failures of our approach on *train* class. The method [25] can better handle objects under rare views while detection-based methods do not due to inaccurate bounding box (top) and missing detection (bottom).

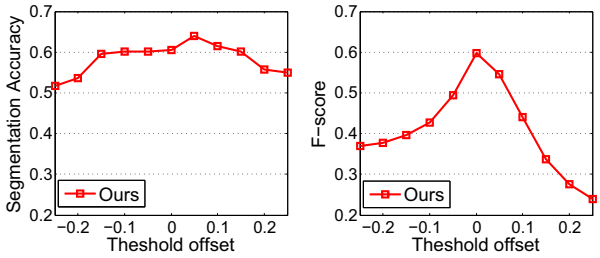

Figure 8. Change of segmentation accuracy with different detection thresholds (left) and corresponding F-score curve (right).

This variant of the framework is referred to Baseline-SG.

Performance comparisons. We summarize the comparisons of our method with other approaches in Table 1, and present some representative results generated by our approach in Fig. 5. Since [25] is based on segment ranking, we partition the ranked list with different threshold settings and report the best accuracy for each object class. Table 1 suggests that the proposed approach outperforms other methods on most challenging categories. However, on categories *bird*, *boat*, and *train*, in which the backgrounds in videos are relatively clean and changing slowly, the method [22] can better distinguish moving foreground objects and the shape guidance in [31] is more capable to align real object boundaries, thus leading to higher performance. Interestingly, the weakly supervised approach [25] performs impressively well on *train* by successfully capturing objects in rare view, while detection-based methods perform poorly due to inaccuracy of object bounding boxes and missing detections (see Fig. 7). However, it is noticeable that in other cases our performance doubles that of [25]. We owe it to the additional detection information that complements the weak knowledge of video tags. To make more comprehensive analysis, we treat the estimated likelihoods (note that [25] produces scores for spatial-temporal segments and thus for pixels) as soft segmentation masks and binarize them with different thresholds to generate precision-recall curves shown in Fig. 6. The curves show that our method significantly surpasses [25] by restricting shape estimation in localized area and thus reducing ambiguities. Nevertheless, it is reasonable to expect that learning based approaches [12, 25, 21] could yield improved performance with the increase of training samples. However, the proposed framework reveals a potential way by exploring well-trained, image-based models for segmentation with video-level labels, which inherently deviates from the existing approaches built on weakly supervised learning and demonstrates great competitiveness. Fig. 6 also suggests the relative superiority of the proposed shape likelihoods than the shape guidance in [31], since the enforced inter-frame consistency is effective to suppress the background and highlight the real foreground objects.

Influence of detection accuracy. We analyse the performance of our method under different detection qualities on a subset containing 30 videos. This is achieved by running the algorithm under different threshold settings. However, since the detection threshold for each object class is calibrated individually, we cannot use a common threshold for all categories. Instead a constant is added to these thresholds and varied uniformly to generate an accuracy curve shown in Fig. 8. We have also manually labeled the object bounding boxes for each video to produce the corresponding F-scores of detectors. From Fig. 8 we see that the accuracy does not lose too much while the F-scores drops dramatically. It contributes to the mechanism to handle missing detections and the interaction modeling which eliminates the tracks that have poor relative structure with others.

4.2. Results on Segtrack v2 Dataset

To compare with a series of the state-of-the-art unsupervised primal object segmentation algorithms [8, 18, 22], we set all the scores relevant to detections to one as explained in Sect. 3.2. In this case, our modified object extraction framework is very similar to the *directed acyclic graph* proposed in [8], which makes [8] a comparable baseline to test the efficacy of our shape likelihoods. To make fair comparisons, all methods use the same procedure with that of [8] to

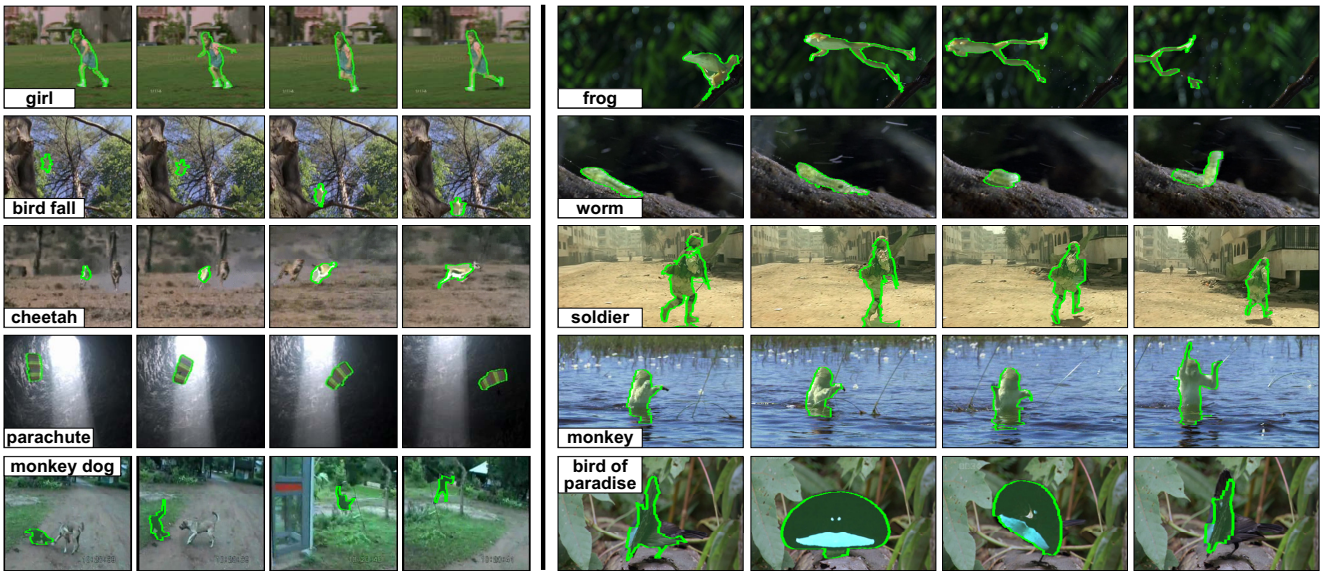

Figure 9. Qualitative results (outlined by green boundaries) of our method on Segtrack v2 dataset. Best viewed in color.

Table 2. Per-frame pixel errors on SegTrack v2 Dataset.

Method Supervised?	Ours N	[8] N	[18] N	[22] N	[13] Y
Girl	1459	1497	**1407**	3859	2883
Birdfall	339	**150**	258	890	1170
Parachute	**196**	219	202	855	228
Cheetah	803	618	930	217	**189**
Monkeydog	365	357	533	**284**	333
Frog	**1272**	3920	121695*	10996	-
Worm	1497	3987	**937**	5082	-
Soldier	**1879**	3233	58140*	5640	-
Monkey	**2526**	2615	3460	7244	-
Bird of Paradise	**1764**	4853	2122	9557	-

* The method [18] may fail on *frog* and *soldier* using the top ranked segment hypothesis (see [18] for more details).

generate segment proposals and default parameters without further tuning. To avoid confusion of segmentation results, the sequences with more than one objects are excluded since all competitors only capture the most prominent object.

The per-frame pixel errors (the smaller the better) shown in Table 2 reflect that the proposed spatio-temporal shape likelihoods consistently surpasses the simple location-based prior used in [8] on *frog*, *worm*, *soldier*, *monkey* and *bird of paradise*, but not perform equivalently well on other sequences. The reason lies in that the target objects of these sequences are too small to obtain enough segment tracks that are crucial for reliable estimation (e.g. the shape estimation stage for *cheetah* and *birdfall* even cannot converge). Our method also slightly outperforms [18], the clustering-driven approach with BPLR [15] based location prior, which shows impressive results but may fail in discovering the target objects when handling object with slow motion (*frog*) and object that has similar appearance with background (*soldier*). The recent approach [22] obtains high-quality segmentation results on this dataset with considerably reduced time cost, but is sometimes sensitive to the moving background regions with similar motion with the foreground objects. We visualize the segmentation results of our approach in Fig. 9.

5. Conclusion and Discussion

This paper proposes a segmentation-by-detection framework for semantic object segmentation in weakly labeled video. The framework starts from per-frame generation of detection and segmentation proposals by utilizing image-based object and region detectors. After that, object tracks are initialized through solving a joint assignment problem, which is shown to be robust to the noisy proposals. Finally, the initial tracks are refined by inferring spatiotemporally consistent shape likelihoods based on statistical information of tracks. Experiments on public datasets show that the proposed method boosts segmentation performance in both weakly labeled and unlabeled videos.

In our future work, we will try to improve the proposed approach mainly from two aspects. To further improve the overall performance, we will seek a tighter upper bound of the track initialization objective and find a more efficient algorithm to solve the optimization problem. Moreover, the proposed approach may be somehow slow when processing long videos since the segmentation is conducted "globally." To speed up our approach, we will try to split the long videos into short overlapping clips first and then perform the segmentation "locally" under a parallel framework.

Acknowledgement. We would like to thank the reviewers for their valuable feedback. This work is supported in part by grants from NSFC (61325011), 863 program (2013AA013801), and SRFDP (20131102130002).

References

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[2] V. Badrinarayanan, I. Budvytis, and R. Cipolla. Mixture of trees probabilistic graphical model for video segmentation. *IJCV*, 110(1):14–29, 2014.

[3] D. Banica, A. Agape, A. Ion, and C. Sminchisescu. Video object segmentation by salient segment chain composition. In *ICCV Workshop*, 2013.

[4] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 34(7):1312–1328, 2012.

[5] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic. On Pairwise Cost for Multi-Object Network Flow Tracking. *ArXiv e-prints*, Aug. 2014.

[6] S. Deqing, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010.

[7] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and semantic segmentation. In *ECCV*, 2014.

[8] Z. Dong, J. Omar, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013.

[9] Z. Dong, J. Omar, and M. Shah. Video object co-segmentation by regulated maximum weight cliques. In *ECCV*. 2014.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[11] H. Fu, D. Xu, B. Zhang, and S. Lin. Object-based multiple foreground video co-segmentation. In *CVPR*, 2014.

[12] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. K. nad O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV Workshop*, 2012.

[13] S.-D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *ECCV*, 2014.

[14] Y. Jian, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012.

[15] J. Kim and K. Grauman. Boundary preserving dense local regions. In *CVPR*, 2011.

[16] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012.

[17] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010.

[18] Y. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[19] F. Li, T.-Y. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

[20] Q. Li, X. Chen, Y. Song, Y. Zhang, X. Jin, and Q. Zhao. Geodesic propagation for semantic labeling. *TIP*, 23(11):4812–4825, 2014.

[21] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu. Nearest neighbor-based label transfer for weakly supervised multi class video segmentation. In *CVPR*, 2014.

[22] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[23] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[24] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[25] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013.

[26] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[27] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*, 101:329–349, 2013.

[28] D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC*, 2010.

[29] S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *ECCV*, 2012.

[30] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *ECCV*, 2014.

[31] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. Semantic segmentation without annotating segments. In *ICCV*, 2013.

[32] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010.