Exploring Weakly Labeled Images for Video Object Segmentation With Submodular Proposal Selection

Yu Zhang, *Student Member, IEEE*, Xiaowu Chen, *Senior Member, IEEE*,
Jia Li, *Senior Member, IEEE*, Wei Teng, and Haokun Song

***Abstract*— Video object segmentation (VOS) is important for various computer vision problems, and handling it with minimal human supervision is highly desired for the large-scale applications. To bring down the supervision, existing approaches largely follow a data mining perspective by assuming the availability of multiple videos sharing the same object categories. It, however, would be problematic for the tasks that consume a single video. To address this problem, this paper proposes a novel approach that explores weakly labeled images to solve video object segmentation. Given a video labeled with a target category, images labeled with the same category are collected, from which noisy object exemplars are automatically discovered. After that the proposed approach extracts a set of region proposals on various frames and efficiently matches them with massive noisy exemplars in terms of appearance and spatial context. We then jointly select the best proposals across the video by solving a novel submodular problem that combines region voting and global region matching. Finally, the localization results are leveraged as strong supervision to guide pixel-level segmentation. Extensive experiments are conducted on two challenging public databases: Youtube-Objects and DAVIS. The results suggest that the proposed approach improves over previous weakly supervised/unsupervised approaches significantly, showing a performance even comparable with the several approaches supervised by the costly manual segmentations.**

***Index Terms*— Semantic object segmentation, weakly labeled video, exemplar matching, submodular optimization.**

Manuscript received July 21, 2017; revised December 22, 2017; accepted February 3, 2018. Date of publication February 16, 2018; date of current version June 1, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61672072, Grant 61532003, and Grant 61421003, in part by the Beijing Nova Program under Grant Z181100006218063, and in part by the Academic Excellence Foundation of BUAA for Ph.D. Students. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yonggang Shi.

Y. Zhang, X. Chen, W. Teng, and H. Song are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: zhangyulb@gmail.com; chen@buaa.edu.cn; tengw@buaa.edu.cn; songhk@buaa.edu.cn).

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the International Research Institute for Multidisciplinary Science, Beihang University, Beijing 100191, China (e-mail: jiali@buaa.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a proof of Submodularity. The 2 videos include comparisons and results from the paper. The total size of the videos is 97.6 MB. Contact zhangyulb@gmail.com for further questions about this work.

I. Introduction

VIDEO object segmentation (VOS) aims to delineate the physical boundaries of object(s) of interest present in a video along space and time axes. It supports various computer vision applications, *e.g.*, class model training [52], [62], action recognition [37] and video editing [10]. In general, there are two tightly correlated subtasks to be addressed for VOS. The first is the *localization* problem, *i.e.*, the target object(s) should be effectively accessed by the system. The other one is object *segmentation*, which looks deeper into detailed object boundaries. The first problem is highly critical since it provides important initial object-level cues.

For user editing tasks, the localization part is (partially) addressed by users with low-cost interactions. They can take various forms: manual segmentations on key frame(s) [6], [48], [51], [65], bounding boxes [15], [57] and even strokes [42]. With these sparse but reliable localizations, the segmentation part of VOS can be effectively solved, either through heuristic spatiotemporal propagation [51], [65] or guided training [6], [48]. The latter has achieved high-quality results by making use of modern well-annotated datasets and learning architectures. However, requiring user interactions for novel videos could be expensive for large-scale or even web-scale applications.

Localizing the video objects automatically is not trivial with a variety of challenges. A major one is the creation of sufficient annotated videos for training the localization module, which is hardly to achieve at large scale. Consequently, the past decade has largely witnessed learning-free VOS models developed with low-level motion, saliency, and boundary cues [9], [26], [32], [47], [67]. However, these bottom-up solutions might be less effective for recognizing unsalient targets within complex scenes. Recent works proposed to employ image-based object detectors to help VOS [5], [66], [70]. However, learning such detectors still relies on massive annotations of the objects of interest, which are scarce for most real-world categories.

To handle this problem, a line of works [19], [35], [62], [71] proposed to address the localization part of VOS with video-level labels, which are much easier to collect. In these approaches, input videos are labeled so that the presence/absence status of the target object categories are known, rather than their exact positions. The collective cues of large amounts of such weakly labeled videos were then explored for localizing the common targets. Such a paradigm,

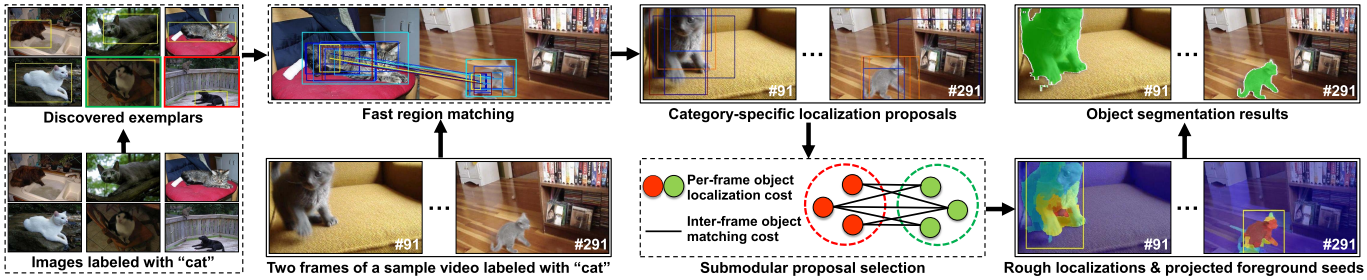


Fig. 1. The pipeline of the proposed approach. Noisy object exemplars are automatically discovered from the images weakly labeled with a category (*e.g.*, "cat"), which may have inaccurate or false localizations (marked in green and red boxes, respectively). Given an input video labeled with the same category, a set of region proposals are extracted on various frames and matched to the exemplars to get localization scores. We then select the best localization proposals on each frame simultaneously via the proposed submodular proposal selection algorithm, and project them onto superpixels to form multiple initial foreground seeds. We further propagate these seeds along the pixel graph of the video to generate consistent segmentations. (Best viewed in color with zoom.)

however, is usually not directly applicable on single-video applications.

This paper's objective is to provide a novel weakly supervised approach for VOS that is applicable to a single video. To this end, we propose to explore massive weakly labeled images to address the localization part of VOS. Given a category of interest, our approach firstly discovers object exemplars from the images labeled with the same category. Region proposals are then extracted from various video frames, and matched with the exemplars for proposal scoring. As shown in the left of Fig. 1, however, an issue that prohibits accurate matching is that automatic localizations of exemplars tend to be noisy. To handle this problem, we propose a fast algorithm that robustly matches the region proposals with massive exemplars in terms of appearance and spatial context. After proposal scoring, the best localization proposals are jointly selected on each frame by solving a submodular problem combining region voting and global matching. Finally, the localizations are projected onto pixels and refined to generate final segmentations. Extensive experiments on two public benchmarks Youtube-Objects [52] and DAVIS [50] show that our approach performs substantially better than existing weakly supervised and unsupervised approaches, even comparably with various supervised ones.

This paper makes the following contributions:

1) We propose a novel weakly supervised VOS approach by exploring massive weakly labeled images, which is applicable on handling a single video and achieves impressive results on large-scale public benchmarks;
2) We present a fast matching algorithm which jointly considers region appearance and spatial context, to robustly handle the noisy localizations of image exemplars;
3) We formulate a submodular problem for selecting localization proposals combining region voting and global matching, which could effectively remove false localizations to generate consistent segmentation results.

We organize this paper as follows. Sect. II briefly reviews the relevant works on VOS. Sect. III describes how to explore weakly labeled images for extracting the localization proposals in video and Sect. IV introduces the techniques of submodular proposal selection. Necessary details for implementation are summarized in Sect. V. Experimental results are presented in Sect. VI and the paper is concluded in Sect. VII.

II. Related Works

Existing models address VOS with various types of supervision, which can be roughly categorized into three groups. The *supervised* group is trained and/or initialized with manual segmentations of the target object categories, while the *unsupervised* group requires neither category-specific training nor user's interactions. The weakly supervised group allows using light-weight form of manual annotations during training and/or testing. We briefly review each of them, and make a discussion on highly relevant works that explores weakly labeled images for visual understanding problems.

A. Supervised VOS Models

There are two paradigms to introduce accurate supervisions into VOS. The first learns category-aware appearance models from massive segmented images/videos [22], [31], [33], [63], [66]. Typically, pretrained segmentation models were applied on the input video(s), whose output are further refined through intra-video consistency (*e.g.* temporal cues [31], [33]) or inter-video consistency [65]. However, since this paradigm solves a larger problem of assigning each pixel to a known category or background, model training is often costly and heavily affected by the availability of annotations from the target categories.

The second paradigm, as widely referred as *semi-supervised* VOS, propagates user's manual annotations on the sparse key frame(s) to the whole video. It could be performed by heuristically fusing appearance and motion cues [38], [38], [51], [65], or learned from existing image/video annotations [6], [48]. In general, semi-supervised VOS can achieve high-quality results due to the accurate prior knowledge of the objects of interest. However, they somewhat have difficulty scaling to large-scale applications that have to process a large amount of videos.

B. Unsupervised VOS Models

Unsupervised VOS models perform object localization and segmentation with intrinsic cues, which is often achieved by identifying the regions with salient motion and/or appearance. For example, several approaches suppose that the foreground objects move in distinct patterns which are different from the background [27], [45], or have closed motion boundaries [47]. Visual saliency was also explored to help localize the objects when motion cues fail [9], [26]. However, low-level visual or motion cues are usually unreliable in case

of complex scenes, which may lead to unexpected results in certain scenarios.

There was also a trend that incorporates mid-level cues for unsupervised VOS [8], [23], [32]. Typically, these approaches first apply the mid-level object proposals [30], [46] to extract a pool of segments in the video, then select a consistent subset of segments to represent the target objects. However, without prior knowledge, these approaches tend to emphasize on the primary object(s) present in a video, and may not work well on the videos where the targets are unsalient in space and time.

C. Weakly Supervised VOS Models

Weakly supervised approaches overcome the drawbacks of supervised and unsupervised ones by assuming low-cost human supervision during training and testing. For the VOS task it is considered to originate from Hartmann *et al.* [19], which learns object segmentations from a large corpus of web videos. There have been many following works afterwards [21], [35], [62], [71]. In these works a set of videos are firstly collected, which are labeled with the target categories or irrelevant categories. After that, they decompose the input videos into spatiotemporal segments and perform various forms of weakly supervised learning (*e.g.* negative mining [62], nearest neighbor classification [35] and representative segment selection [71]) to divide them into foreground and background. Recently, Hong *et al.* [21] has proposed the first deep learning approach that combines class-agnostic attentions and video-level propagation for weakly supervised SOS, and achieves impressive results.

The learning perspective of weakly supervised VOS is based on the availability of large amounts of relevant input videos. To handle a single video, recent works [5], [70] proposed to incorporate object detectors trained with weaker annotations, i.e., bounding boxes. They perform object detection on various frames to generate object tracks, which are further refined with motion cues to generate visually consistent results. Although these works have demonstrated strong performance, they still leave behind the problem that even bounding box annotations are available for rather limited number of real-world object categories. Therefore, bringing down the human supervision required for VOS to the minimal level is still an open problem.

D. Visual Understanding With Weakly Labeled Images

Exploring weakly labeled images for visual understanding is a promising direction that has been demonstrated by solving a wide range of tasks. For example, Sultani and Shah [60] proposes to explore the human actions automatically discovered from web images to guide action localization in videos. Aubry *et al.* [4] utilizes massive chair images rendered from 3D CAD models to detect object locations and 3D poses in real images. Khosla *et al.* [28] proposes to discover canonical object views from internet photos for large-scale video summarization.

A similar idea with ours is proposed by Ahmed *et al.* [3], which utilizes weakly labeled images for image segmentation. However, they assume that most of the collected images have clean background by searching on the internet with controlled key words. In this work, we explore much more general form of weakly labeled images. Our approach also differs with [3] in addressing object segmentation in spatiotemporal domain.

III. Exploring Weakly Labeled Images for Video Object Localization

Given the input video which is associated with a category label, images labeled with the same category are firstly collected, from which a set of exemplars are discovered. A novel algorithm is then proposed that matches the exemplars to the video content and generates category-specific localization proposals. In the rest of this section, we firstly describe the exemplar discovery algorithm used in this work for completeness, and then present the proposed matching algorithm in more details.

Note that instead of employing existing object detectors pretrained for large-scale object detection (*e.g.* Faster RCNN [54] and YOLO [53]), we propose, to the best of our knowledge, the first exemplar-driven approach to solve video object segmentation. Such approaches have been extensively studied for image recognition [3], [17], [43] for their practical advantage that handles newly acquired data/categories without expensive retraining/fine-tuning. Experiments suggest that our approach can achieve even better results than previous works that employ pretrained detectors and tracking-by-detection methods.

A. Exemplar Discovery From Weakly Labeled Images

Although there are many existing approaches for exemplar discovery from images (*e.g.* [7], [56]), they are often inefficient since the ultimate goal of them is to accurately co-localize the common objects in all the images. As we are only interested to collect a set of good exemplars, we can adopt a simpler and more efficient approach to achieve this purpose.

In details, we apply a state-of-the-art salient object proposal generator [69] to extract up to 30 proposals per image. Similar with many mid-level object proposals [13], [34], [39], [61], salient object proposals [34], [69] are pretrained with a small amount of annotations to learn general object-level knowledge. However, instead of localizing all the objects present in the scene with a large pool of proposals to achieve high recall, they aim to detect a few objects of salient appearance with high precision and are thus suitable for quickly discovering a set of high-quality exemplars. Without category-specific knowledge, the extracted proposals may contain many irrelevant objects. However, it is reasonable to assume that objects from the target category appear most frequently. Therefore, we can construct a pairwise affinity matrix $\mathbf{S}$ to model the consensus of category distributions between a pair of proposals $\mathcal{E}_i$ and $\mathcal{E}_j$:

$$\mathbf{S}_{ij} = \frac{e^{-\|\mathbf{c}(\mathcal{E}_i)-\mathbf{c}(\mathcal{E}_j)\|^2}}{\sum_{\mathcal{E}\in\mathbb{E}} e^{-\|\mathbf{c}(\mathcal{E}_i)-\mathbf{c}(\mathcal{E})\|^2}}. \quad (1)$$

The proposal representation $\mathbf{c}(\mathcal{E})$ is the 1000-class probability vector predicted by the VGG-16 network [58] pretrained for image classification. If two proposals are from the same category, they have similar class distributions and the consensus is thus larger. Following [60], we perform the *manifold ranking*

Fig. 2. Discovered exemplars on the Pascal VOC 2012 dataset. Each column shows results from a different object category. Row 1st: successfully discovered exemplars. Row 2nd: false positive exemplars, which mainly come from frequently co-occurred categories (e.g. *person* appear frequently with many categories). Row 3rd: representative exemplars removed by manifold ranking. The removed exemplars include irrelevant object categories, meaningful but incomplete object parts, and false localizations.

algorithm to rank the proposals. Initially, the scores of all the proposals (denoted with $\mathbf{y}_0$) are the same. Manifold ranking iteratively smoothes each exemplar's score by gradually taking the votes from their neighbors, namely,

$$\mathbf{y}^{(t+1)} = \alpha \mathbf{S}\mathbf{y}^{(t)} + (1-\alpha)\mathbf{y}_0, \tag{2}$$

where α is set to 0.99, and $\mathbf{y}^{(t)}$ is the vector of exemplar scores at the tth iteration. In this manner, windows with irregular class distributions are pushed far away from others and scored lower. We refer the readers to [72] for more details of the algorithm.

After scoring the proposals, we retain the top N proposals as exemplars. However, the retained exemplars are still noisy and may contain co-occurred but actually irrelevant categories (*e.g.* persons appear frequently along with many other categories, see Fig. 2). In the rest of this section, we propose a novel algorithm to efficiently and robustly handle this issue.

B. Object Localization in Video

For each frame of the input video, the proposed localization algorithm starts with extracting a set of region proposals. For the sake of clarity, we concentrate on the principle of algorithm design here and put the necessary details for implementation in Sect. V. For each region $\mathcal{R}$, its appearance is represented as $\mathbf{f}(\mathcal{R})$ in some feature space. We find its K nearest exemplars in this space and denote them with $\mathcal{O}_k(\mathcal{R})$, $k \in \{1, 2, \cdots, K\}$. With the matched exemplars, the region $\mathcal{R}$ can be scored by accumulating its affinities with the nearest exemplars

$$s(\mathcal{R}) = \sum_{k=1}^{K} a\left(\mathcal{R}, \mathcal{O}_k(\mathcal{R})\right) = \sum_{k=1}^{K} e^{-\frac{1}{\gamma_a}\|\mathbf{f}(\mathcal{R})-\mathbf{f}(\mathcal{O}_k(\mathcal{R})\|^2}, \tag{3}$$

where γ_a is the bandwidth of the Gaussian kernel. Note that (3) can be deemed as a soft K-Nearest-Neighbor (kNN) classifier. In ideal case, the number of nearest exemplars K is usually set small to reduce the approximation error of a kNN model. However, since the discovered exemplars often contain noisy ones from irrelevant categories, small K may introduce large estimation error. Fig. 3 illustrates an example. Since persons frequently appear in the *train* category, the person in this video receives good matches from the exemplar set and is scored higher than the target train. Therefore, K should be necessarily large to effectively bypass this negative effect.

Fig. 3. The highest scored region proposal using different numbers of nearest exemplars on a *train* video. Few neighbors makes the matching sensitive to the noisy exemplars from irrelevant categories (*e.g.*, persons appear in many images labeled with other categories, see Fig. 2).

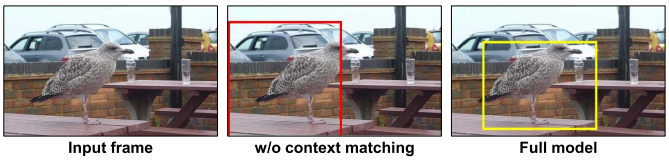


Fig. 4. The highest scored region proposal with and without context matching on a frame of a video labeled with *bird*. Localization by appearance matching with a large number of exemplars is biased to include additional background area to adapt to the diverse appearance of exemplars.

However, since the target object is usually similar with only a few exemplars, simply aggregating the appearance affinities with a large number of dissimilar exemplars may unexpectedly "over-smooth" the signal of the target object. As a result, the localization is biased to include addtional background to better adapt to the diverse appearance of the exemplars (see Fig. 4). To address this issue, we propose to weight the contributions of different exemplars by their spatial context, namely

$$s'(\mathcal{R}) = \sum_{k=1}^{K} c\left(\mathcal{R}, \mathcal{O}_k(\mathcal{R})\right) a\left(\mathcal{R}, \mathcal{O}_k(\mathcal{R})\right), \tag{4}$$

where $c\left(\mathcal{R}, \mathcal{O}_k(\mathcal{R})\right)$ measures the similarity of spatial configurations of the nearby regions between the input proposal and the exemplar neighbor. The intuition behind (4) is that if the spatial context of an exemplar is similar with the input region, this match should be more confident than the others and thus weighted larger. To measure the context-based similarity, we sample a set of regions $\mathbb{N}(\mathcal{R})$ arount $\mathcal{R}$ with sufficient spatial overlap (*i.e.*, the Jaccard index is greater than 0.5). For each nearby region $\mathcal{R}' \in \mathbb{N}(\mathcal{R})$, we match it to the image

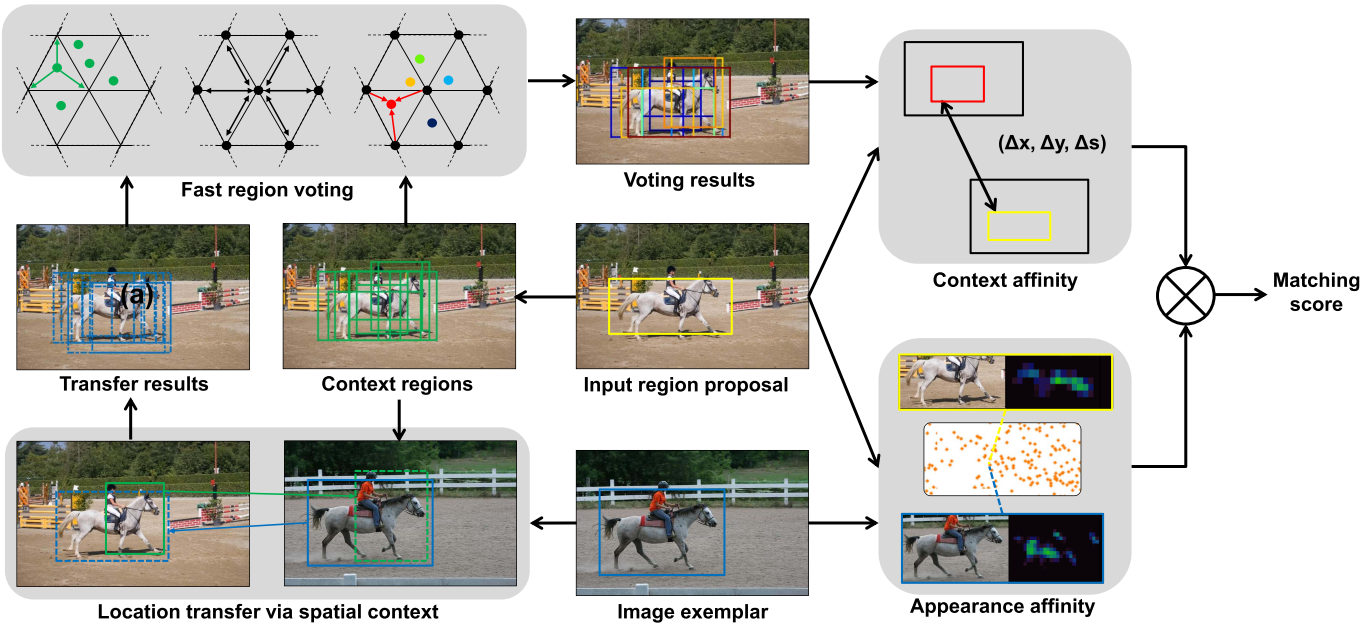


Fig. 5. The pipeline of the proposed exemplar matching algorithm. The algorithm output combines two kinds of affinities. The appearance affinity is computed by matching region appearance in feature space. To compute the context similarity, a set of nearby regions is sampled around the input proposal and matched to the exemplar image. Each match transfers back a predicted object position, which gives votes to all the proposals on the frame. The context affinity quantifies the similarity between the position of the input proposal and and the one that recieves maximal votes. (Best viewed in color.)

containing $\mathcal{O}_k(\mathcal{R})$ to obtain the matched region $\mathcal{P}_k(\mathcal{R}')$, based on their appearance features. If the context configurations are similar, the offset between $\mathcal{P}_k(\mathcal{R}')$ and $\mathcal{O}_k(\mathcal{R})$ and that between $\mathcal{R}'$ and the underlying object should be consistent. Therefore, each region $\mathcal{R}'$ can generate a guessed object location $\mathcal{T}_k(\mathcal{R}')$ by preserving such offset. This lead to the following formula:

$$\mathbf{x}\left(\mathcal{T}_k(\mathcal{R}')\right) = \mathbf{x}(\mathcal{R}) \oplus \left(\mathbf{x}\left(\mathcal{O}_k(\mathcal{R})\right) \ominus \mathbf{x}\left(\mathcal{P}_k(\mathcal{R}')\right)\right), \quad (5)$$

where $\mathbf{x}(\cdot)$ is a vector concatenating the x and y coordinates and scale (*i.e.*, the square root of area) of the input region, the symbols $\oplus$ and $\ominus$ represent element-wise plus and minus operations for vectors, respectively. In this manner, we obtain a set of transfer results for each region in $\mathbb{N}(\mathcal{R})$.

The transferred regions often have inconsistent positions due to the noise in region matching. Thus, we let them to softly assign votes to each input proposal, and select the one with the maximal consensus as the final prediction. For each proposal $\mathcal{R}_0 \in \mathbb{R}$, its vote is accumulated as follows

$$v_k(\mathcal{R}_0) = \sum_{\mathcal{N} \in \mathbb{N}_{\mathcal{R}}} a\left(\mathcal{N}, \mathcal{P}_k(\mathcal{N})\right) e^{-\frac{1}{\gamma_s}\|\mathbf{x}(\mathcal{T}(\mathcal{N})) - \mathbf{x}(\mathcal{R}_0)\|^2}, \quad (6)$$

which jointly considers the transfer quality and spatial affinity. Afterwards, the final prediction is chosen by

$$\mathcal{R}^* = \arg\max_{\mathcal{R}_0 \in \mathbb{R}} v_k(\mathcal{R}_0). \quad (7)$$

We set the context-based weights as the agreement between the position of the input proposal $\mathcal{R}$ and the result predicted by its spatial context:

$$c\left(\mathcal{R}, \mathcal{O}_k(\mathcal{R})\right) = e^{-\frac{1}{\gamma_s}\|\mathbf{x}(\mathcal{R}) - \mathbf{x}(\mathcal{R}^*)\|^2}. \quad (8)$$

Although the proposed spatial matching algorithm improves the performance significantly in our experiments, a drawback is that computing votes for every input proposal using (6) is expensive since it has time complexity $O(|\mathbb{R}||\mathbb{N}_{\mathcal{R}}|)$, where $|\cdot|$ is the size of an input set. Preliminary experiments suggest that a naive implementation will take 30 seconds to process a frame with $K_{\mathcal{E}} = 200$ and $|\mathbb{R}| \approx 700$, which cannot be affordable in practice. However, an interesting property of the proposed formulation is that the time-consuming step (6) is actually an instance of Gaussian filtering. This property makes it possible to apply an efficient solver [2] to compute (6) via fast convolution on a specialized data structure. With respect to our case, it consists of three steps: 1) *splatting*, which maps the region coordinates to a set of pre-defined feature nodes; 2) *blurring*, approximating the filtering steps using the fast solver; and 3) *slicing*, inversely reconstructing the region votes as the linear combinations of the blurred values on the feature nodes. The time complexity of this algorithm is only $O\left(|\mathbb{R}| + |\mathbb{N}_{\mathcal{R}}|\right)$, which reduces the time cost from quadratic to linear. Due to the page limit, we omit further details and refer the interested readers to [2]. After scoring the proposals by (4), we retain the top 30 proposals on each frame as the candidate localization proposals of the target object in the video.

IV. Submodular Proposal Selection

Note that the above matching process operates at per-frame basis and ignores video-level consistency. Thus, it is somewhat sensitive to the imperfect conditions on certain frames, where occlusion and fast camera/object transition occurs. To address this problem, this section proposes an efficient algorithm that globally selects localizations on all the frames.

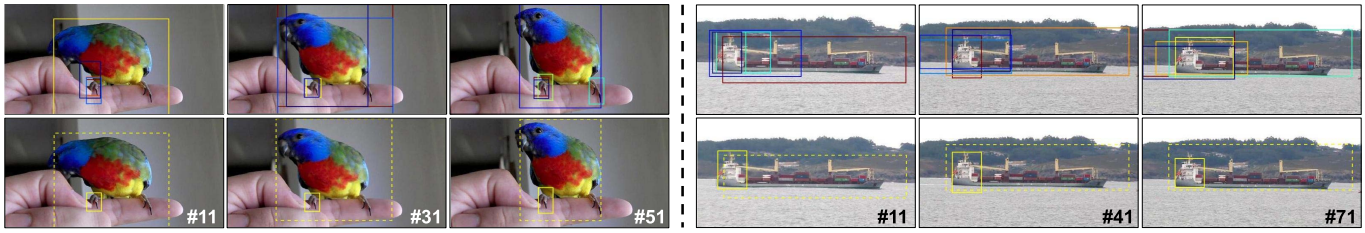


Fig. 6. Proposal selection results for two sample videos. For each video, we show several localization proposals with the highest scores for each frame in the top row and the selection results in the bottom. In the top row, colors represent localization scores (red is highest and dark blue is lowest). Not considering the mutual dependence of proposal scores only localizes the discriminative parts of the target objects instead of their whole extent (solid boxes). The proposed context-aware algorithm correctly localizes the objects (dashed boxes). (Best viewed in color.)

A. Problem Formulation

Given the input video $\mathcal{V}$, and a set of localization proposals $\mathbb{L}_t$ on the tth frame, $t \in \mathbb{T} = \{1, 2, \cdots, T\}$, our task is to select one proposal on each frame, *i.e.* $\mathbb{L}^* = \{\mathcal{L}_t^*\}_{t\in\mathbb{T}, \mathcal{L}_t^*\in\mathbb{L}_t}$, to represent the target object. We formulate this task as finding a subgraph from a fully connected graph, where the localization proposals on all the frames form the nodes and edges exist between any pair of nodes. With the redundant connections that can span across long-range time windows, fully connected graphs are effective to handle segmentation-related challenges such as occlusion and fast camera/object transition [29], [51].

We follow the *Maximum A Posteriori* (MAP) framework to address the proposal selection task, *i.e.* to maximize the posterior $P(\mathbb{L}^*|\mathcal{V})$. We start with the following *Gibbs* distribution:

$$P_{gibbs}(\mathbb{L}^*|\mathcal{V}) \propto \prod_{\mathcal{L}\in\mathbb{L}^*} e^{-E_d(\mathcal{L}|\mathcal{V})} \cdot \prod_{\mathcal{L},\mathcal{L}_0\in\mathbb{L}^*, \mathcal{L}\neq\mathcal{L}_0} e^{-\lambda E_s(\mathcal{L},\mathcal{L}_0|\mathcal{V})}, \quad (9)$$

where the *data* term $P_d(\mathbb{L}^*|\mathcal{V}) = \prod_{\mathcal{L}\in\mathbb{L}^*} e^{-E_d(\mathcal{L}|\mathcal{V})}$ factorizes across each individual proposal modeling their relevance to the category of interest, while the *smooth* term $P_s(\mathbb{L}^*|\mathcal{V}) = \prod_{\mathcal{L},\mathcal{L}_0\in\mathbb{L}^*, \mathcal{L}\neq\mathcal{L}_0} e^{-E_s(\mathcal{L},\mathcal{L}_0|\mathcal{V})}$ enforces matching consistency among the selected proposals, and λ is a non-negative weight parameter. Formulations like (9) have been widely adopted to solve proposal selection tasks [20], [25].

A drawback of (9), however, is that the data term considers the proposals independently, which means that goodness of a proposal only depends on its score without accessing to other context information. In our task, this leads to failures if a small part of the target object receives higher score than the whole object region. Such issues can be observed for objects with non-rigid or irregular shapes, since their discriminative parts may appear more consistently than the whole bodies across both the image exemplars and the input video (see Fig. 6).

To address this issue we rely on two observations. First, the top scored proposals, although may correspond to small object parts, still lie inside the groundtruth object region. Second, the proposals with high scores are mostly distributed on the object body, and for those spanning over additional background area their scores decrease sharply. Such discontinuity of proposal scores is a side-effect of the proposed context matching algorithm as large background area makes the spatial configuration of context regions differ greatly with that of exemplars.

The observations and analysis above inspire us to propose a formulation that explores the second-order information of proposal scores implemented as region voting. For each proposal $\mathcal{L}_t \in \mathbb{L}_t$, we introduce a binary auxiliary variable $u(\mathcal{L}_t, \mathcal{L}_t^*)$ which takes 1 if $\mathcal{L}_t$ serves as a "supporter" of the selected proposal $\mathcal{L}_t^*$ on the same frame, and 0 otherwise. Assuming that $u(\cdot,\cdot)$ is independent in terms of different proposals, we apply the following conditional probabilities

$$P_d\left(u(\mathcal{L}_t, \mathcal{L}_t^*) = 1|\mathcal{V}\right) = \begin{cases} \dfrac{\bar{s}(\mathcal{L}_t) + \bar{s}(\mathcal{L}_t^*)}{2Z(\mathcal{L}_t)}, & \mathcal{L}_t \sqsubset \mathcal{L}_t^* \quad (10) \\ \dfrac{\left(s(\mathcal{L}_t^*) - s(\mathcal{L}_t)\right)_+}{Z(\mathcal{L}_t)}, & \mathcal{L}_t \sqsupset \mathcal{L}_t^* \quad (11) \end{cases}$$

and $P_d\left(u(\mathcal{L}_t, \mathcal{L}_t^*) = 0|\mathcal{V}\right) = \frac{\tau}{Z(\mathcal{L}_t)}$. In these equations, $\bar{s}(\cdot)$ is computed by normalizing the proposal score (4) into [0, 1], and the symbol $\sqsupset$ or $\sqsubset$ means that the left region contains or is contained by the right region, respectively. The function $(\cdot)_+$ is short for $\max(\cdot, 0)$, and $Z(\cdot)$ is the normalization factor of the probabilities. Note that (10) and (11) correspond to the two observations of proposal score distributions. The parameter τ controls the probability that the proposal $\mathcal{L}_t$ is drawn from the background and does not contribute to the voting.

Regarding the values of voting variables $\mathbf{u}$ as hidden states in defining $P_d(\mathbb{L}^*|\mathcal{V})$, now we have

$$P_d(\mathbb{L}^*|\mathcal{V}) = \sum_{\mathbf{u}} P_d(\mathbb{L}^*, \mathbf{u}|\mathcal{V}) \propto \sum_{\mathbf{u}} P_d(\mathbb{L}^*|\mathbf{u}, \mathcal{V}) P_d(\mathbf{u}|\mathcal{V}).$$

For fixed voting variables $\mathbf{u}$, the selected proposals $\mathbb{L}^*$ could be uniquely determined since proposals that are voted should be also in the selection results. It means that the probability $P_d(\mathbb{L}^*|\mathbf{u}, \mathcal{V})$ equals zero anywhere except at certain $\mathbb{L}^*$ indicated by $\mathbf{u}$. Thus, maximizing $P_{gibbs}(\mathbb{L}^*|\mathcal{V})$ is equivalent to

$$\begin{aligned} \max_{\mathbb{L}^*} P_{gibbs}(\mathbb{L}^*|\mathcal{V}) &= \max_{\mathbb{L}^*} P_d(\mathbb{L}^*|\mathcal{V}) P_s(\mathbb{L}^*|\mathcal{V}) \\ &\propto \max_{\mathbb{L}^*, \mathbf{u}} P_d(\mathbb{L}^*|\mathbf{u}, \mathcal{V}) P_d(\mathbf{u}|\mathcal{V}) P_s(\mathbb{L}^*|\mathcal{V}). \quad (12) \end{aligned}$$

Since the relationship between $\mathbf{u}$ and $\mathbb{L}^*$ does not depend on the specific form of the input video,

$$P_d(\mathbb{L}^*|\mathbf{u}, \mathcal{V}) = P_d(\mathbb{L}^*|\mathbf{u}) = \frac{P_d(\mathbb{L}^*, \mathbf{u})}{P_d(\mathbf{u})}. \quad (13)$$

We define the joint distribution of $\mathbb{L}^*$ and $\mathbf{u}$ as

$$P_d(\mathbb{L}^*, \mathbf{u}) = P_d(\mathbb{L}^*) P_d(\mathbf{u}) C_1(\mathbb{L}^*, \mathbf{u}) C_2(\mathbb{L}^*), \quad (14)$$

where $C_1(\cdot,\cdot)$ is an indicator function that takes 1 if constraints between $\mathbb{L}^*$ and $\mathbf{u}$ are satisfied, and 0 otherwise. The function $C_2(\cdot)$ complies the constraint that one proposal is selected on each frame. Substituting (13) and (14) into (12), we have

$$\max_{\mathbb{L}^*} P_{gibbs}(\mathcal{L}^*|\mathcal{V}) \propto \max_{\mathbb{L}^*,\mathbf{u}} P_d(\mathbf{u}|\mathcal{V}) P_s(\mathbb{L}^*|\mathcal{V}) C_1(\mathbb{L}^*,\mathbf{u}) C_2(\mathbb{L}^*), \tag{15}$$

which is the objective we need to optimize. Note that $P_d(\mathbf{u}|\mathcal{V})$ has been previously defined, and we assume uniform prior of $\mathbb{L}^*$ and thus omit the term $P_d(\mathbb{L}^*)$ in (15). To compute the smooth term $P_s(\mathbb{L}^*|\mathcal{V})$, we define the cost as $E_s(\mathcal{L}_1^*,\mathcal{L}_2^*|\mathcal{V}) = 1 - c(\mathcal{L}_1^*,\mathcal{L}_2^*)a(\mathcal{L}_1^*,\mathcal{L}_2^*)$, *i.e.*, reusing the appearance and context matching proposed in (4).

B. Optimization

The logarithm of the objective of (15) takes the form

$$\sum_t \sum_{\substack{\mathcal{L}_t\in\mathbb{L}_t,\\ \mathcal{L}_t^*\in\mathbb{L}^*\cap\mathbb{L}_t}} \log P\left(u(\mathcal{L}_t,\mathcal{L}_t^*)|\mathcal{V}\right) - \lambda \sum_{\mathcal{L}_1^*,\mathcal{L}_2^*\in\mathbb{L}^*} E_s(\mathcal{L}_1^*,\mathcal{L}_2^*|\mathcal{V}),$$
$$\text{subject to } \forall t\in\mathbb{T},\ |\mathbb{L}^*\cap\mathbb{L}_t| = 1. \tag{16}$$

We omit $C_1(\cdot,\cdot)$ and $C_2(\cdot)$ as the constraints are now explicitly considered. Maximizing (16) jointly over $\mathbb{L}^*$ and $\mathbf{u}$ is NP-hard. However, by regarding it as a set optimization problem defined over $\mathbb{L}^*$ (*i.e.*, firstly maximizing over $\mathbf{u}$ then over $\mathbb{L}^*$), two good properties could be shown. First, the objective function now becomes submodular (the proof is referred to the supplementary material). Second, the disjoint sets $\{\mathbb{L}_t\}_{t=1}^T$ form a *partition matroid w.r.t.* the solution $\mathbb{L}^*$ since $\forall t, |\mathbb{L}^*\cup\mathbb{L}_t| \le 1$.

These facts suggest that (16) is connected to the constrained submodular maximization problem defined on partition matroids [12], [68]. For such family of problems, good local optimality could be guaranteed with a simple greedy algorithm. Thus, we propose a two-stage algorithm that firstly solves the relaxed submodular problem to obtain the selected results on several frames. These proposals are then regarded as reliable seeds to guide the selection on the remaining frames.

Stage I: Relaxing the original equality constraints to inequalities $\forall t, |\mathbb{L}^*\cup\mathbb{L}_t| \le 1$ leads to a constrained submodular maximization problem. Starting with an initially selected proposal, the greedy algorithm performs a sequence of local update operations: *Add Operation*, which introduces a new proposal into the solution set; *Swap Operation*, which replaces a selected proposal with a currently unselected one; *Delete Operation*, which removes an already selected proposal. Each operation aims to increase the value of the objective function. We perform the algorithm with multiple iterations, each time starting with the highest scored proposal on a different frame. The algorithm initialized on the tth frame is summarized in Alg. 1, where $\Psi(\mathbb{L}^*)$ represents the objective value under the solution $\mathbb{L}^*$. The final solution is chosen as one that maximizes the objective value among all the iterations.

Stage II: The relaxed constraints do not ensure that each frame has a selected proposal. Thus we greedily select proposals for the remaining frames, starting with the current solution set $\mathbb{L}^*$, $\mathbb{T}^*$. We iteratively perform the following operations: 1) solve $(\hat{t},\hat{\mathcal{L}}) = \arg\max_{t\in\mathbb{T}\setminus\mathbb{T}^*,\mathcal{L}\in\mathbb{L}_t} \Psi(\mathbb{L}^*\cup\mathcal{L})$; 2) update $\mathbb{L}^* = \mathbb{L}^*\cup\{\hat{\mathcal{L}}\}$, $\mathbb{T}^* = \mathbb{T}^*\cup\{\hat{t}\}$, until $|\mathbb{L}^*| = |\mathbb{T}^*| = T$, *i.e.* all frames are processed.

Algorithm 1 The Algorithm of State I Initialized by the Highest Scored Proposal on the tth Frame

1: **Input** $\mathbb{L}^* = \{\arg\max_{\mathcal{L}\in\mathbb{L}_t} \bar{s}(\mathcal{L})\}, \mathbb{T}^* = \{t\}$;
2: **While** $|\mathbb{L}^*| \le T$ and $\mathbb{L}^*$ is updated by any operation
3: **Add operation:** if $\exists t\in\mathbb{T}\setminus\mathbb{T}^*, \mathcal{L}\in\mathbb{L}_t$, s.t. $\Psi(\mathbb{L}^*\cup\{\mathcal{L}\}) > \Psi(\mathbb{L}^*)$, then $\mathbb{L}^* = \mathbb{L}^*\cup\{\mathcal{L}\}, \mathbb{T}^* = \mathbb{T}^*\cup\{t\}$;
4: **Swap operation:** if $\exists t_0\in\mathbb{T}^*, \mathcal{L}_0\in\mathbb{L}^*\cap\mathbb{L}_{t_0}, t\in\mathbb{T}\setminus\mathbb{T}^*, \mathcal{L}\in\mathbb{L}_t$, s.t. $\Psi(\mathbb{L}^*\setminus\{\mathcal{L}_0\}\cup\{\mathcal{L}\}) > \Psi(\mathbb{L}^*)$, then $\mathbb{L}^* = \mathbb{L}^*\setminus\{\mathcal{L}_0\}\cup\{\mathcal{L}\}, \mathbb{T}^* = \mathbb{T}^*\setminus\{t_0\}\cup\{t\}$;
5: **Delete operation:** if $\exists t\in\mathbb{T}^*, \mathcal{L}\in\mathbb{L}^*\cap\mathbb{L}_t$, s.t. $\Psi(\mathbb{L}^*\setminus\{\mathcal{L}\}) > \Psi(\mathbb{L}^*)$, then $\mathbb{L}^* = \mathbb{L}^*\setminus\{\mathcal{L}\}$, $\mathbb{T}^* = \mathbb{T}^*\setminus\{t\}$;
6: **End While**
7: **return** $\mathbb{L}^*$.

C. Integration for Video Object Segmentation

So far, the proposed approach generates a set of localization bounding-boxes for the input video. In the rest of this section, we show how to produce pixel-level segmentation masks from bounding-box initializations. Our main idea follows the consensus voting approach [9], which generates initial foreground saliency maps individually on each frame and then iteratively refines them on a superpixel graph.

We first derive a global saliency map for each frame using the segmentation masks associated with the region proposals. Such segmentations are available along with bounding-boxes for many concurrent proposals [30], [40], [46]. Denote $\mathbb{M}(\mathcal{L})$ as the set of foreground pixels inside some region proposal $\mathcal{L}$. For each pixel location $\mathcal{N}$ on the tth frame, its saliency value is given as the frequency that it is covered by foreground masks:

$$S_t(\mathcal{N}) = \frac{\sum_{\mathcal{L}\in\mathbb{L}_t} \mathbb{1}\left(\mathcal{N}\in\mathbb{M}(\mathcal{L})\right) F(\mathcal{L})}{\sum_{\mathcal{L}\in\mathbb{L}_t} \mathbb{1}\left(\mathcal{N}\in\mathbb{M}(\mathcal{L})\right)}, \tag{17}$$

where $\mathbb{1}(\cdot)$ is the characteristic function which takes 1 if the input condition holds, and 0 otherwise. To account for salient motions, each proposal $\mathcal{L}$ is reweighted by the average magnitude of optical flows $F(\mathcal{L})$ computed along the segmentation boundary, following [8].

The saliency map obtained by (17) is then modulated by the localizations provided by the proposal selection algorithm. Given the localized bounding box $\mathcal{L}_t^*$ on the tth frame, we define a spatial Gaussian centered at $\mathcal{L}_t^*$ to generate a position-sensitive foreground map:

$$G_t(\mathcal{N}) = e^{-\frac{1}{\gamma_s}\left(\|\frac{x(\mathcal{N})-x(\mathcal{L}_t^*)}{w(\mathcal{L}_t^*)}\|^2 + \|\frac{y(\mathcal{N})-y(\mathcal{L}_t^*)}{h(\mathcal{L}_t^*)}\|^2\right)}, \tag{18}$$

where $x(\cdot)$ and $y(\cdot)$ define the image coordinates of a pixel (or region center), and $w(\cdot)$, $h(\cdot)$ represent the width and height of a region, respectively. The final foreground saliencies combine (17) and (18) by pixel-wise multiplication.

For refining the initial foreground maps we follow the first stage of the consensus voting algorithm [9] precisely. Briefly speaking, a kNN superpixel graph is firstly constructed for the video by finding the nearest neighbors of each superpixel in

appearance feature space. On this graph, the initial saliencies are regarded as seeding segmentations and iteratively improved with the random-walk algorithm [18] to generate final segmentations. We choose this refinement procedure for its simplicity, fast speed and superior performance, while other commonly adopted routines (*e.g.* GrabCut [55]) could be applied either.

V. Implementation Details

A. Object Proposals

We adopt the fast mode of MCG proposals [46], using its unsupervised part only without supervised ranking process. It generates 700 proposals for each image on average. Such proposals are extracted on both the input video frames and the database images.

B. Proposal Appearance Features

We represent each proposal with DeepPyramid features [14]. To this end, the input image is warped to fit into 9 discretized aspect ratios ranging from 0.25 to 4 with the factor $\sqrt{2}$. It is also up/down-sampled to construct 7-level image pyramids with the largest dimension of the original scale resized to 512 pixels. This generates $7 \times 9 = 63$ combinations, which are passed separately to the network to generate the *conv5* feature maps. In this way, each proposal can be aligned with a 8×8 feature template at a proper location, aspect ratio and scale. Flatting the template leads to 16384-dimensional features. On the GTX1080 platform it takes 2.4 seconds to process a frame on average.

C. Fast Distance Computation

The high-dimensional features are further embedded into a 256-bit Hamming space via [16] for speed-up. In this way the distance computation is extremely fast: with a specialized kd-tree [41] for hamming space kNN search, it costs only a fraction of a second to process a frame.

VI. Experiments

This section aims to evaluate the proposed approach through several experiments. Across all the evaluations, object exemplars are discovered from the Pascal VOC 2012 [1] database, which comprises various unfiltered Internet photos and represents real-world difficulties well. No provided annotations are accessed except the image-level labels. For each category, we retain up to $N = 800$ exemplars.

Two challenging public benchmarks are employed to evaluate the proposed approach:

1) **Youtube-Objects.** This dataset is originated from Prest *et al.* [52], which comprises various internet videos. Jain and Grauman [24] and Tang *et al.* [62] have provided groundtruth object segmentations for different subsets of the original dataset. For the sake of convenience we refer the two subsets as YTO-Jain and YTO-Tang, respectively. Both subsets have 10 object categories from the Pascal VOC classes, while for each video a single class is annotated. YTO-Jain is composed of 126 videos, where the objects are accurately labeled for 1 in every 10 frames. YTO-Tang consists of 151 densely annotated videos, while the annotations are roughly labeled on supervoxels. Both subsets have more than 20000 frames (up to 400 frames for each video), which are among the largest benchmarks for video object segmentation nowadays.
2) **DAVIS.** This benchmark is originated from [50], comprising 50 high-quality videos with 3455 frames in total. Each video contains a primary foreground object, which is densely and accurately annotated. Various real-world object categories are present spanning humans, animals, vehicles, etc. Note that most categories have consistent counterparts using the Pascal VOC class definition. For those not, we assign them with the closest category (*e.g.*, *camel*, *elephant* and *rhino* are assigned to *cow* category).

We comprehensively compare the proposed approach (denoted as SPS) with 16 existing automatic video object segmentation approaches. According to the type of annotations they use, we categorize them into 4 groups:

1) **UN group.** This group contains 9 approaches which rely on bottom-up saliency and/or motion cues while do not require annotated data. These approaches are KEY [32], MSG [44], TRC [11], LTV [45], FST [47], NLC [9], SAG [67], CVOS [64] and ACO [26].
2) **SL group.** This group contains 2 approaches that assume models learned with manual segmentations. It includes FCN [36] and SCV [66], which are both trained with the Pascal VOC 2012 dataset.
3) **LL group.** This group contains 3 approaches that assume image/video-level labels as supervision. It includes CRANE [62], MWS [35] and WCV [21]. The first two do not assume additional data, while WCV is also trained with the weakly labeled Pascal VOC images.
4) **BL group.** This group is composed of 2 approaches that require detectors learned with bounding-box annotations, including DET [70] and DTM [5]. These annotations come from the Pascal VOC 2012 dataset. SPS falls into this group as it adopts salient object proposals pre-trained with bounding-box annotations. However, we emphasize that SPS only assumes image/video labels when applied to a novel category.

On the Youtube-Objects dataset two evaluation metrics are employed, following previous works: 1) mean Intersection-over-Union scores (mIoU), also known as the Jaccard Index, computed as the number of intersected pixels divided by that of the union pixels between the predicted and the groundtruth segmentation masks for each video, and averaged across all the videos; 2) mean Average Precision (mAP), computed from the soft foreground segmentation probabilities before thresholding. On the DAVIS dataset we directly apply the provided metrics $\mathcal{J}$, $\mathcal{F}$ and $\mathcal{T}$ that evaluate the segmented regions, contours and the temporal stability across frames [50].

Unless specifically explained, the number of nearest exemplars $K_{\mathcal{E}}$ is set to 200 among the experiments. We also provide evaluations to analyze the performance of our approach with other values. The kernel bandwidths γ_a and γ_s are empirically set to 80 and 10, respectively. Parameters for proposal

TABLE I

PERFORMANCE ON YTO-JAIN DATASET IN MIOU. ALONG EACH COLUMN, BOLD HIGHLIGHTS THE TOP PLACE WHILE UNDERLINE THE SECOND

Group	Model	Aeroplane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Motorbike	Train	Average
UN	LTV	0.137	0.122	0.108	0.237	0.186	0.163	0.180	0.115	0.106	0.196	0.155
	FST	0.709	0.706	0.425	0.652	0.521	0.445	0.653	0.535	0.442	0.296	0.538
	ACO	0.630	0.690	0.400	0.610	0.480	0.460	**0.670**	0.530	0.470	0.380	0.530
SL	FCN	0.635	0.698	0.464	0.699	0.557	0.549	0.595	0.515	0.445	0.563	0.572
	SCV	0.693	**0.761**	0.572	0.704	**0.677**	0.597	0.642	**0.571**	0.441	**0.579**	**0.623**
LL	WCV	-	-	-	-	-	-	-	-	-	-	0.586
BL	DET	0.698	0.677	0.515	0.695	0.408	0.599	0.614	0.512	0.435	0.525	0.568
	DTM	0.744	0.721	**0.585**	0.600	0.457	**0.612**	0.552	0.566	0.421	0.367	0.562
	SPS	**0.784**	0.712	0.562	**0.772**	0.559	0.589	0.618	0.565	**0.515**	0.549	0.622

TABLE II

PERFORMANCE ON YTO-TANG SUBSET IN MAP

	CRANE	MWS	SPS
mAP	0.425	0.461	0.712

selection λ and τ are set to 3 and 0.5. Their impacts are also systematically analyzed with additional experiments.

A. Comparisons With State-of-the-Arts

In the first experiment, we compare the proposed approach with the other 11 approaches with available codes or results on the Youtube-Objects dataset. The results are summarized in Table I and Table II, and several representative examples are shown in Fig. 7. From Table I, we find that the UL group performs very well although both the object localization and segmentation are heuristically designed. FST and ACO achieve high-quality results on the categories with salient motions, *e.g.*, animals. However, their performance may significantly drop if no such intrinsic information could be utilized, *e.g.* the objects are nearly static and do not have obvious motion.

The previous BL models improve over the UN group by at least 0.03, notably on vehicle categories but slightly or even negatively on animals. While the former is explained by the power of object detections, the latter indicates that previous BL models somewhat have difficulty handling non-rigid objects well due to the imperfect detections on these categories.

The SL group demonstrates high-quality results as expected. With the detailed annotations and strong learning techniques, this group performs well on various categories. SCV achieves the leading results among the comparisons. Despite the superior performance, however, SCV still suffers several practical limitations, such as the dependence on segmented training data for each category, and the requirement of multiple relelvant input videos to be simultaneously processed.

The performance of our approach SPS closely follows SCV, with a small gap within 0.1%. Moreover, SPS assumes only weakly annotated data in practice, and is applicable to a single video. Compared with the other approaches in the BL group, the number of SPS is substantially higher. Remarkably, the proposed approach does not seem to have severe performance degeneration on non-rigid categories. We suspect that while fitting a global template for non-rigid objects is challenging, the proposed matching algorithm finds locally similar matches, which effectively avoids missing detection in several videos.

Looking into in the LL group, we find that WCV performs impressively well, even surpassing the strong FCN baseline. With slight additional supervisions SPS improves over WCV by a large margin. Compared with two previous approaches CRANE and MWC that do not assume additional data, SPS has at least 54% relative improvements.

In the second experiment, we provide additional benchmarking results on a subset of the original frames of the YTO-Jain dataset, which comprises the annotated frames only. It leads to 10x downsampling for each video, which simulates fast transitions of scales/viewpoints well. We denote this smaller dataset as YTO-Jain-Sub and summarize the results in Table III. In this setting, the performance of previous approaches degenerates greatly. Actually, many previous approaches assume that the scenes evolve smoothly, which may not hold in case of such fast motion. In contrast, SPS does not make such assumption but instead relies on global reasoning, thus is not very sensitive to this issue (*i.e.*, from 0.622 to 0.603). Later we show that SPS can indeed perform reliably under various difficulties.

In the third experiment, we compare the proposed approach with 7 automatic approaches and 7 interactive ones on the DAVIS datset. The quantative and qualitive comparisons are shown in Table IV and Fig. 8, respectively. Table IV illustrates that the proposed approach outperforms existing unsupervised approaches by a large margin in terms of region and boundary accuracies. We find that the proposed approach handles fast transitions well on this dataset (*e.g.*, see the results in Fig. 8). Capable of segmenting the objects in these cases leads to a significantly improved recall (*i.e.*, 0.845 versus the previous best performance 0.800, in Jaccard Recall).

The proposed approach performs better than or comparably with several interactive approaches, although it is automatic and does not need to be trained with costly annotations. We notice that many interactive approaches may have difficulty propagating the segmentations to the frames where viewpoint or scale greatly changes (*e.g.*, the *drift* and *motocross* videos). SPS is robust to these cases by utilizing the information provided by multiple frames jointly instead of sequentially.

TABLE III

PERFORMANCE ON YTO-JAIN-SUB DATASET IN MIOU. ALONG EACH COLUMN, BOLD HIGHLIGHTS THE TOP PLACE WHILE UNDERLINE THE SECOND

Group	Model	Aeroplane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Motorbike	Train	Average
UN	FST	0.485	0.641	0.315	0.365	0.309	0.337	0.386	0.323	0.170	0.341	0.367
	NLC	0.666	0.599	0.263	0.342	0.275	0.330	0.453	0.375	0.314	0.470	0.409
	SAG	0.543	0.551	0.339	0.505	0.344	0.413	0.387	0.371	0.311	0.275	0.404
	ACO	0.575	0.607	0.374	0.311	0.360	0.312	0.458	0.407	0.217	0.342	0.396
SL	SCV	0.655	0.634	0.358	0.495	0.424	0.408	0.452	0.349	0.462	0.365	0.478
BL	DET	0.580	0.679	0.457	0.608	0.367	**0.588**	0.539	0.469	0.443	0.485	0.521
	SPS	**0.781**	**0.698**	**0.521**	**0.675**	**0.528**	0.584	**0.631**	**0.528**	**0.544**	**0.532**	**0.602**

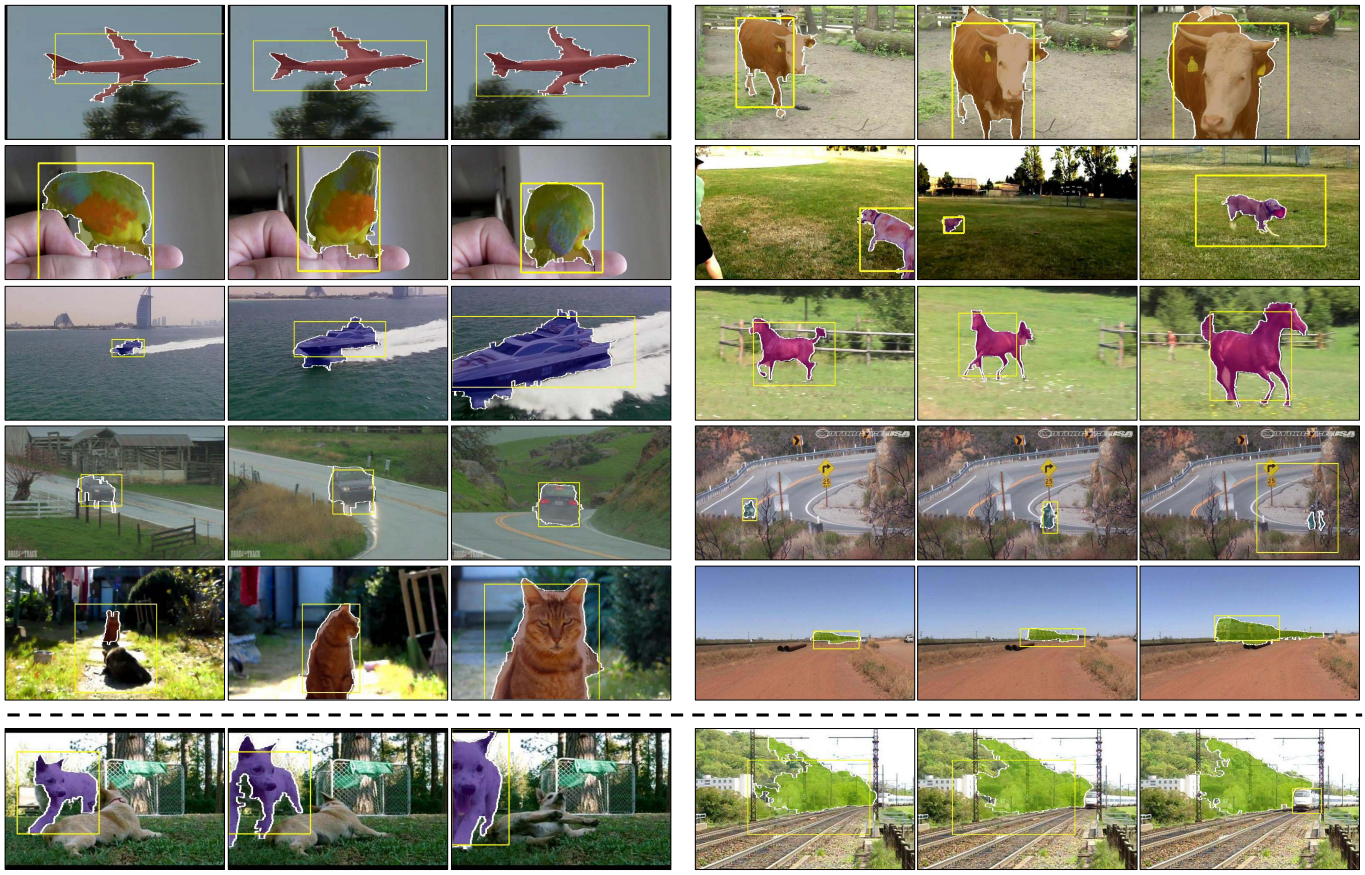


Fig. 7. Representative segmentation results generated by the proposed approach on the Youtube-Objects dataset. Object segmentations are indicated as colored regions, and the selected localization proposals are marked with yellow boxes. In the last row, we show two typical failure modes of our approach, including multi-object instances and detection failures. More results are included in the supplementary material. (Best viewed in color.)

TABLE IV

PERFORMANCE ON THE DAVIS DATASET, FOLLOWING THE CONVENTIONAL METRICS. THE ARROW ↑ (↓) INDICATES THAT RESULTS ARE BETTER IF THE NUMBER IS HIGHER (LOWER). FOR EACH ROW, BOLD HIGHLIGHTS THE TOP PLACE WHILE UNDERLINE THE SECOND

Model	SPS	FST	NLC	MSG	KEY	CVOS	TRC	SAL	OFL	BVS	FCP	JMP	HVS	SEA	TSP
Interaction?	N	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y
$\mathcal{J}$ Mean ↑	**0.679**	0.575	0.641	0.543	0.569	0.514	0.501	0.426	**0.711**	0.665	0.631	0.607	0.596	0.556	0.358
$\mathcal{J}$ Recall ↑	**0.845**	0.652	0.731	0.636	0.671	0.581	0.560	0.386	**0.800**	0.764	0.778	0.693	0.698	0.606	0.388
$\mathcal{J}$ Decay ↓	0.052	0.044	0.086	**0.028**	0.075	0.127	0.050	0.084	0.227	0.260	**0.031**	0.372	0.197	0.355	0.385
$\mathcal{F}$ Mean ↑	**0.642**	0.536	0.593	0.525	0.503	0.490	0.478	0.383	**0.679**	0.656	0.546	0.586	0.576	0.533	0.346
$\mathcal{F}$ Recall ↑	**0.759**	0.579	0.658	0.613	0.534	0.578	0.519	0.264	**0.780**	0.774	0.604	0.656	0.712	0.559	0.329
$\mathcal{F}$ Decay ↓	0.070	0.065	0.086	**0.057**	0.079	0.138	0.066	0.072	0.240	0.236	**0.039**	0.373	0.202	0.339	0.388
$\mathcal{T}$ (GT=9.5) ↓	0.480	0.293	0.366	0.263	**0.210**	0.256	0.345	0.616	0.224	0.317	0.294	**0.136**	0.305	0.141	0.333

The metric that our approach fails to ascend the leading places is the temporal stability (T). To keep a clear focus, we adopt a rather simplified segmentation pipeline that does not enforce temporal smoothness of the results. This, however, is straightforward to improve via constraints established between adjacent frames, *e.g.*, following [32], [51], and [65].

Fig. 8. Visual comparisons of the proposed approach with several leading approaches on two videos from the DAVIS dataset. Compared with other approaches, the proposed approach benefits from high-quality localizations and is robust to viewpoint/scale transitions and fast motion.

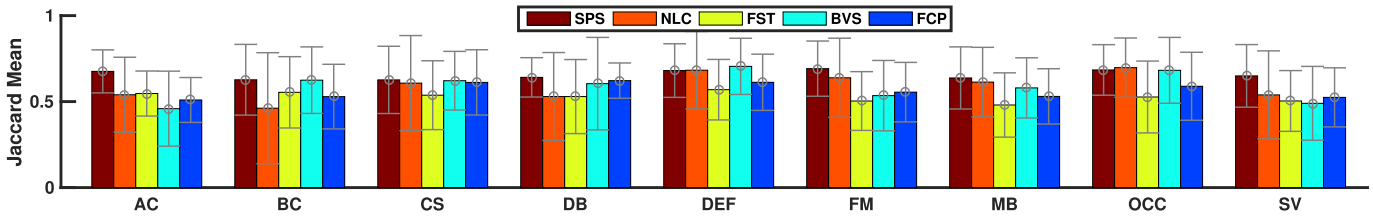


Fig. 9. Attribute-based evaluations on the DAVIS dataset. We consider 9 attributes: AC (Appearance Change), BC (Background Clutter), CS (Camera Shake), DB (dynamic background), DEF (non-linear deformation), FM (fast motion), MB (motion blur), OCC (occlusions) and SV (scale variation). For each attribute, we show the mean and variance of Jaccard scores of the proposed approach and state-of-the-art unsupervised/interactive approaches. (Best viewed in color.)

TABLE V

STEP-WISE PERFORMANCE OF THE PROPOSED APPROACH ON YTO-JAIN-SUB DATASET, IN MIOU. ALONG EACH COLUMN, BOLD HIGHLIGHTS THE TOP PLACE WHILE UNDERLINE THE SECOND

Step	Aeroplane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Motorbike	Train	Average
Best proposal	0.519	0.516	0.341	0.555	0.373	0.460	0.459	0.424	0.430	0.431	0.451
After region voting	0.535	0.530	0.345	0.566	0.383	0.482	0.465	0.424	0.461	0.440	0.463
After proposal selection	0.534	0.648	0.387	0.568	0.449	0.493	0.516	0.441	0.465	0.447	0.495
After refinement	**0.781**	**0.698**	**0.521**	**0.675**	**0.528**	**0.584**	**0.631**	**0.528**	**0.544**	**0.532**	**0.602**

B. Performance Analysis

Beyond the comparisons with the state-of-the-arts, we conduct additional experiments to show how our approach works under various scenarios. At first, we report the performance on the videos of various difficulty attributes provided by the DAVIS dataset, and summarize the results in Fig. 9.

From these results, we conclude that the proposed approach performs the best *w.r.t.* most attributes, and at least comparably *w.r.t.* all. Particularly, our approach is most effective to handle Appearance Change (AC), Fast Motion (FM) and Scale Variation (SV), which matches the observations on the YTO-Jain-Sub dataset. We find that many previous interactive approaches are less effective when the appearance of foreground objects change significantly, while many unsupervised approaches fail to localize the objects when the background also has salient motion. Our approach shows stable performance in both cases.

In the second experiment, we conduct a step-by-step ablation study to analyze the contributions of different components of our approach on the YTO-Jain-Sub dataset, as shown in Table V. The *Best proposal* baseline naively selects the highest scored proposal on each frame as the localization results. After incorporating the proposed region voting algorithm, the results are consistently improved, which generalizes across almost all the categories. The full model which further combines consistent matching improves the results by 0.032.

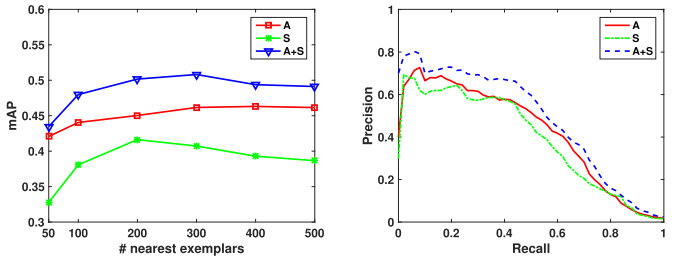


Fig. 10. Performance of the localization proposal generation module evaluated on the YTO-Jain-Sub dataset. "A" and "S" represent baseline algorithms which separately use appearance and spatial context, while "A+S" is the full model. Left: the mean Average Precisions (mAPs) of the generated localization proposals as a function of the number of neighbor exemplars K. Right: the precision-recall curves at $K = 200$.

Proposals	mAP	mIoU
SOD	0.381	0.587
(+MR)	0.430	0.602
COB	0.187	0.562
(+MR)	0.230	0.571
UDOL	0.109	0.561
(+MR)	0.137	0.579

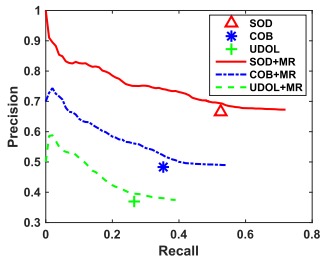


Fig. 11. Evaluating the impact of proposal generators. Left: Performances of using different proposals without and with manifold ranking (+MR). We show localization mAPs on the *trainval* set of the Pascal VOC 2012 dataset and the segmentation mIoUs on the YTO-Jain-Sub dataset. Right: the precision-recall curves for exemplar localization on the Pascal VOC2012 dataset.

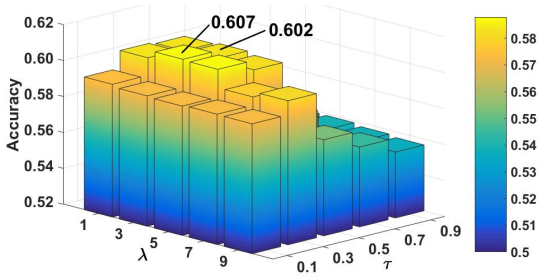


Fig. 12. Performances on the YTO-Jain-Sub dataset as a function of different settings of the proposal selection parameters, measured in mIoU. The settings that correspond to the best performance and the performance reported in our benchmarking (see Table III) are marked with text. (Best viewed in color.)

The refinement step has a significant impact on the final performance since it produces much more accurate segmentation boundaries.

Beyond the ablation study, we look in greater details into the proposed localization proposal generation module. To this end we fit bounding boxes for each ground-truth object from the YTO-Jain dataset, and evaluate the localization performance. The results are summarized in Fig. 10. When the number of exemplars increases, the performance of appearance matching baseline consistently improves. This matches our assumption that a large number of exemplars are necessary to handle noisy exemplars. The combined results (*i.e.*, "A+S") are much better than those generated by individual pipelines. However, using too many exemplars may unexpectedly harm the performance of spatial context matching. We suspect that when matching dissimilar exemplars, the predictions made by spatial context matching becomes unreliable due to ambiguous region correspondences. This, however, only leads to slight degeneration while the overall performance does not actively response to it.

The success of the proposed approach partly relies on the state-of-the-art SOD object proposals [69]. We thus investigate two other state-of-the-art proposal generators: COB [40] and UDOL [7]. COB is pretrained with annotated training data, while UDOL is an unsupervised co-localization approach. We retain the top 1 proposal from each image using the rankings output by both approaches, to generate results of high precision. As these approaches emphasize on higher recall rather than precision, retaining more proposals may introduce many noisy proposals and greatly harm the discovered exemplars.

The results are shown in Fig. 11. Among the comparisons, SOD achieves the best precision since it is designed to extract a few salient objects. This is in contrast to COB and UDOL which aim to localize all the target objects. As good precision is more important than recall for exemplar discovery, SOD achieves the best final results. Manifold ranking consistently improves the localization mAP for each approach, as well as the final segmentation results. Interestingly, SPS achieves similar segmentation performance even using unsupervised proposals. As suggested by the previous experiment, the proposed localization module is robust and shows stable performance even with a large number of noisy exemplars.

The remaining important question is how the parameters λ and τ introduced by our proposal selection algorithm impacts the final results. To this end, we vary their values and report the performances on the YTO-Jain-Sub dataset in Fig. 12. We observe that the proposed approach is not very sensitive to the choice of λ, which controls the weight of global matching. The choice of the probability that a proposal is inactive for voting, namely τ, is better around 0.5. When it is small, false positive votes will be introduced that make the selection unreliable. Large value close to 1 removes the effect of region voting. In this way, the algorithm turns to selecting the static background as they match more consistently than the foreground objects.

We report the time consumption of each individual stage of the proposed approach in Table VI. To process a frame, our approach takes 16 seconds on average using a workstation with 32GB memory, 3.4GHz CPU and a GTX1080 graphics card. On the same platform, it runs much faster than previous top-performing approaches SCV ($\approx$ half a minute) and DET ($>$ 2 minutes). Such performance is even comparable with several recent interactive approaches, although our approach localizes the object automatically. For example, MaskTrack [48] and FCP [51] are reported to cost 12 and 16 seconds, respectively.

Our implementation used for timing is not strictly optimized with most components performed in a single thread. To further improve the speed, proposal and feature extraction could be replaced with faster ones (*e.g.* [54]), while matching massive exemplars can benefit from parallelization platforms.

TABLE VI

PER-FRAME RUNNING TIME OF THE PROPOSED APPROACH AVERAGED ON THE YTO-JAIN-SUB DATASET

Stage	Time (secs/frame)
Proposals & optical flows	5.27
Features with Embedding	2.42
Exemplar matching	6.50
Proposal selection	0.17
Pixel-level refinement	1.76

VII. CONCLUSION

In this paper we explore weakly labeled images to address video object segmentation. To this end, a robust and efficient algorithm is firstly proposed for exemplar-driven object localization. We further observe that the second-order relationships among proposals are helpful for accurate localization, and propose a proposal selection algorithm that benefits from well-studied optimization theories. The proposed approach achieves impressive performance on two challenging benchmarks.

The results obtained so far indicate that high-quality results for video object segmentation could be achieved with scalable approaches that do not rely on massive accurate annotations. Our future work will address several limitations and further challenges. First, to simplify this task, the proposed approach extracts a single object in the video following many weakly supervised approaches [7], [59]. Thus, it would be interesting to enhance our approach to segment multiple classes/objects in a video. Second, we are also interested to extend the proposed image-to-video matching algorithm to address video-to-video matching, and investigate efficient iterative methodologies for joint object segmentation in a collection of videos.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their help in improving this paper.

REFERENCES

[1] *The PASCAL VOC 2012 Challenge*. Accessed: Jul. 20, 2017. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

[2] A. Adams, J. Baek, and M. Davis, "Fast high-dimensional filtering using the permutohedral lattice," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 753–762, 2010.

[3] E. Ahmed, S. Cohen, and B. Price, "Semantic object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3150–3157.

[4] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3762–3769.

[5] B. Drayer and T. Brox. (2016). "Object detection, tracking, and motion segmentation for object-level video segmentation." [Online]. Available: https://arxiv.org/abs/1608.03066

[6] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5320–5329.

[7] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1201–1210.

[8] Z. Dong, J. Omar, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 628–635.

[9] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[10] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen, "JumpCut: Non-successive mask transfer and interpolation for video cutout," *ACM Trans. Graph.*, vol. 34, no. 6, Nov. 2015, Art. no. 195.

[11] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1846–1853.

[12] S. Fujishige, *Submodular Functions and Optimization*, vol. 58. Amsterdam, The Netherlands: Elsevier, 2005.

[13] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 2578–2586.

[14] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 437–446.

[15] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 81–88.

[16] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 817–824.

[17] A. Gonzalez-Garcia, A. Vezhnevets, and V. Ferrari, "An active search strategy for efficient object class detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3022–3031.

[18] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[19] G. Hartmann *et al.*, "Weakly supervised learning of object segmentations from web-scale video," in *Proc. ECCV Workshop Web-Scale Vis. Social Media*, 2012, pp. 198–208.

[20] Z. Hayder, M. Salzmann, and X. He, "Object co-detection via efficient inference in a fully-connected CRF," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 330–345.

[21] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using Web-crawled videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2224–2232.

[22] A. Jain, S. Chatterjee, and R. Vidal, "Coarse-to-fine semantic video segmentation using supervoxel trees," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1865–1872.

[23] S. D. Jain, B. Xiong, and K. Grauman, "FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos," in *Proc. CVPR*, 2017, pp. 2117–2126.

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

[25] S.-D. Jain and K. Grauman, "Active image segmentation propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2864–2873.

[26] W.-D. Jang, C. Lee, and C.-S. Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 696–704.

[27] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3271–3279.

[28] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using Web-image priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2698–2705.

[29] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 109–117.

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

[31] A. Kundu, V. Vineet, and V. Koltun, "Feature space optimization for semantic video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3168–3175.

[32] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1995–2002.

[33] P. Lei and S. Todorovic, "Recurrent temporal deep field for semantic video labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 302–317.

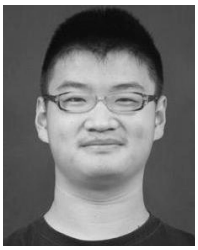
[34] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 247–256.
[35] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu, "Weakly supervised multiclass video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 57–64.
[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[37] J. Lu, R. Xu, and J. J. Corso, "Human action segmentation with hierarchical supervoxel consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3762–3771.
[38] N. Marki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 743–751.
[39] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized Prim's algorithm," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2536–2543.
[40] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Convolutional oriented boundaries," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 580–596.
[41] M. Muja and D. G. Lowe, "Fast matching of binary features," in *Proc. Comput. Robot Vis.*, May 2012, pp. 404–410.
[42] N. S. Nagaraja, F. R. Schmidt, and T. Brox, "Video segmentation with just a few strokes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3235–3243.
[43] M. Najafi, S. T. Namin, M. Salzmann, and L. Petersson, "Sample and filter: Nonparametric scene parsing via efficient filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 607–615.
[44] P. Ochs and T. Brox, "Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1583–1590.
[45] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.
[46] P. P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 328–335.
[47] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1777–1784.
[48] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3491–3500.
[49] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 724–732.
[50] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 724–732.
[51] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung, "Fully connected object proposals for video segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3227–3234.
[52] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3282–3289.
[53] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 6517–6525.
[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[55] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[56] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in Internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1939–1946.
[57] G. Seguin, P. Bojanowski, R. Lajugie, and I. Laptev, "Instance-level video segmentation from object tracks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3678–3687.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[59] K. K. Singh, F. Xiao, and Y. J. Lee, "Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3548–3556.
[60] W. Sultani and M. Shah, "What if we do not have multiple videos of the same action?—Video action localization using Web images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1077–1085.
[61] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe. (2014). "Scalable, high-quality object detection." [Online]. Available: https://arxiv.org/abs/1412.1441
[62] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2483–2490.
[63] B. Taylor, A. Ayvaci, A. Ravichandran, and S. Soatto, "Semantic video segmentation from occlusion relations within a convex optimization framework," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. 2013, pp. 195–208.
[64] B. Taylor, V. Karasev, and S. Soattoc, "Causal video object segmentation from persistence of occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4268–4276.
[65] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3899–3908.
[66] Y.-H. Tsai, G. Zhong, and M.-H. Yang, "Semantic co-segmentation in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 760–775.
[67] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun.2015, pp. 3395–3402.
[68] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2235–2244.
[69] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Unconstrained salient object detection via proposal subset optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5733–5742.
[70] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia, "Semantic object segmentation via detection in weakly labeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3641–3649.
[71] G. Zhong, Y.-H. Tsai, and M.-H. Yang, "Weakly-supervised video scene co-parsing," in *Proc. ACCV*, 2016, pp. 20–36.
[72] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 169–176.

Yu Zhang (S'17) received the B.E. degree in computer science from Beihang University in 2012, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering. His research interests include computer vision and video processing.

Xiaowu Chen (SM'15) received the Ph.D. degree in computer science from Beihang University in 2001. He is currently with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include virtual reality, augmented reality, computer graphics, and computer vision.

Jia Li (SM'15) received the B.E. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently an Associate Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image/video processing.

Wei Teng is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. Her research interests include computer vision and image processing.



Haokun Song is currently pursuing the M.Sc. degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and machine learning.