Cross-Granularity Graph Inference for Semantic Video Object Segmentation

Huiling Wang[1], **Tinghuai Wang**[2], **Ke Chen**[1], **Joni-Kristian Kämäräinen**[1]
[1]Department of Signal Processing, Tampere University of Technology, Finland
[2]Nokia Technologies, Finland
{huiling.wang, ke.chen, joni.kamarainen}@tut.fi, tinghuai.wang@nokia.com

Abstract

We address semantic video object segmentation via a novel cross-granularity hierarchical graphical model to integrate tracklet and object proposal reasoning with superpixel labeling. Tracklet characterizes varying spatial-temporal relations of video object which, however, quite often suffers from sporadic local outliers. In order to acquire high-quality tracklets, we propose a transductive inference model which is capable of calibrating short-range noisy object tracklets with respect to long-range dependencies and high-level context cues. In the center of this work lies a new paradigm of semantic video object segmentation beyond modeling appearance and motion of objects locally, where the semantic label is inferred by jointly exploiting multi-scale contextual information and spatial-temporal relations of video object. We evaluate our method on two popular semantic video object segmentation benchmarks and demonstrate that it advances the state-of-the-art by achieving superior accuracy performance than other leading methods.

1 Introduction

Semantic video object segmentation aims at grouping pixels in video frames into spatio-temporal regions belonging to a unique semantic class label. Notable progress has been made toward this problem by incorporating middle- and high-level visual information, such as object detection [Zhang *et al.*, 2015; Wang *et al.*, 2016; Drayer and Brox, 2016], to build an explicit semantic notion of video objects. Such an integration with object recognition and segmentation not only facilitates a holistic object model, but also provide a middle-level geometric representations for delineating semantic objects.

Existing detection-segmentation based approaches usually fail to capture long-range and high-level contexts due to the lack of joint modeling and inference of contexts and segmentation. Those methods either directly employ detected object proposals, *i.e.* local context, from independent frames associated in temporal domain as constraints to enforce labelling consistence [Zhang *et al.*, 2015; Drayer and Brox, 2016] or build holistic object models using adapted image-based raw detections [Wang *et al.*, 2016]. However, object detections followed by temporal association may contain errors due to inconsistent object appearance across frames and occlusions. Using independent object proposals as constraint for labelling without incorporating various context cues leads to mis-segmentations. To discover and utilize both local and global contexts, we propose a novel cross-granularity graphical model for jointly modeling and inference of different granularities, *i.e.* tracklets, frame-wise object proposals and superpixels.

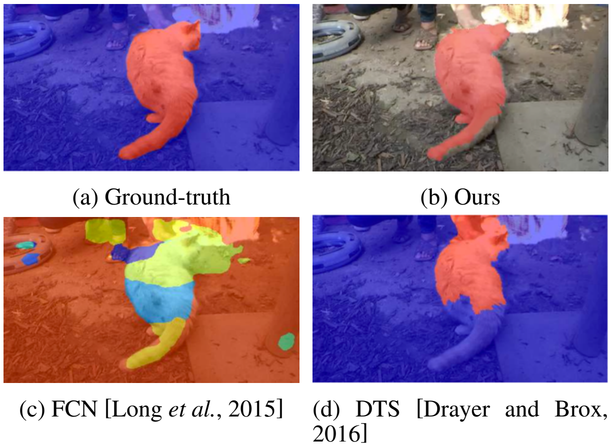
(a) Ground-truth
(b) Ours
(c) FCN [Long *et al.*, 2015]
(d) DTS [Drayer and Brox, 2016]

Figure 1: Semantic segmentation examples from the proposed approach and the state-of-the-art image and video semantic segmentation methods.

Tracklet characterizes the spatial-temporal evolution of various object features and are commonly tied to higher-level contexts such as object interactions and behaviours [Choi and Savarese, 2012]. High-quality tracklets play a crucial role in building our cross-granularity graphical model by introducing global contexts to semantic labeling. To this end, given a video sequence, we first build a graph which consists of object proposal tracklets each of which captures short-range spatio-temporal contexts. We then propose a novel graph-based transductive inference model which operates on the constructed object graph of tracklets. Our transductive inference model enables the calibration of the noisy object proposals by exploring the long-range spatio-temporal contexts. The rationale is that graph-based transductive inference captures the synergy of objects belonging to the same category in deep feature space and thus propagates the confidence along the graph with respect to long-range spatio-temporal dependencies and high-level semantics.

We then construct a hierarchical graphical model which consists of calibrated tracklets, object proposals and superpixels, where each of the nodes captures spatio-temporal contexts and semantics from coarse to fine granularities. The tracklet level nodes encode long-range contexts and higher-level semantic synergies, which resolve local motion and appearance ambiguities. The object proposal nodes enforce short-range dependencies and local geometric representation, while the superpixel nodes play the role as perceptually meaningful atomic regions to guarantee efficient and accurate segmentation despite that they are much less informative. The information flow across nodes of various granularities enables effective inference which accounts for both bottom-up and top-down semantic cues.

The contribution of this paper is two-fold. First, our paper is the first attempt to jointly model and infer complex dynamics and multi-scale contextual cues via a cross-granularity hierarchical graph for semantic video object segmentation. Second, we develop a novel transductive inference model which is capable of calibrating the noisy semantic confidence of short-range object tracklets with respect to long-range and high-level contexts. Experiments on two popular benchmarks verify superior performance of the proposed method to the state-of-the-arts.

2 Related Work

This work falls in the field of unsupervised or weakly supervised video object segmentation, where the current literature can be generally grouped into motion segmentation, generic object segmentation and semantic object segmentation in videos.

Methods from the first category normally cluster pixels using appearance and optical flow based motion information across all frames [Wang *et al.*, 2009; Sundberg *et al.*, 2011; Ayvaci and Soatto, 2012] or take a bottom-up approach based on spatio-temporal appearance and motion constraints [Papazoglou and Ferrari, 2013; Giordano *et al.*, 2015], which work to some extent if the objects show some independent motion in the video. However, motion segmentation methods typically fail on static objects or almost all motion is due to camera motion, or objects move as a unit.

The second category consists of methods generally utilizing two middle-level representations, i.e., object proposals and salient region detection, to form an explicit notion of generic objects. Methods [Lee *et al.*, 2011; Zhao and Fu, 2015; Wang and Wang, 2016; Xiao and Lee, 2016] have been proposed to explore recurring object-like regions from still images by measuring generic object appearance [Endres and Hoiem, 2010]. Saliency measure has been used to detect generic objects in methods [Banica *et al.*, 2013; Wang *et al.*, 2015]. These approaches typically aim to segment the primary object or all foreground objects regardless of semantic labels.

Methods from the third category, which is the closest to our approach, segment the video object with assigned semantic labels. These methods either train weakly supervised classifiers from collections of positive and negative videos [Hartmann *et al.*, 2012; Tang *et al.*, 2013; Liu *et al.*, 2014] or employ off-the-shelf object detection models [Taylor *et al.*, 2013; Zhang *et al.*, 2015; Wang *et al.*, 2016; Drayer and Brox, 2016]. [Hartmann *et al.*, 2012] formulated the problem of semantic video object segmentation as learning weakly supervised classifiers for a set of independent spatio-temporal segments. A discriminative model based on distance matrix is proposed by [Tang *et al.*, 2013] who leverage labelled positive videos and a large collection of negative examples. [Liu *et al.*, 2014] proposed nearest neighbour-based label transfer algorithm which encourages smoothness between regions that are spatio-temporally adjacent and similar in appearance. In another line of work, [Taylor *et al.*, 2013] develop an approach to incorporate scene topology and semantics from trained Textonboost classifier. [Zhang *et al.*, 2015] adopt pre-trained Deformable Part Model based object detector to generate a set of raw detections which are used to enforce spatio-temporal labeling consistence. [Wang *et al.*, 2016] employ a pre-trained image recognition model and build semi-supervised graphical model for domain adaptation to generate smooth semantic confidence map. Drayer and Brox [Drayer and Brox, 2016] enhance the motion segmentation approach [Papazoglou and Ferrari, 2013] with detected and tracked object regions to improve segmentation accuracy on videos with no motion, dominant camera motion, and objects that move as a unit.

Our approach differs from [Taylor *et al.*, 2013; Zhang *et al.*, 2015; Wang *et al.*, 2016; Drayer and Brox, 2016] mainly in two aspects. First, our approach performs object tracklets level inference incorporating long-range dependencies to alleviate sporadic local outliers instead of using temporally associated raw detections with only short-range cues as labeling constraints. Second, we construct a cross-granularity hierarchical graphical model consisting of calibrated tracklets, object proposals and superpixels to account for contextual cues at multiple scales, whilst the existing approaches use single-layer graph with local contexts for inference.

3 Transductive Inference of Tracklets

We first introduce our graph-based transductive inference model which takes as input the noisy object tracklets with short-range contexts, and enables the calibration of the confidence of tracklets by exploring the long-range contextual information. The calibrated object tracklets augments the hierarchical graphical model with high quality object proposals and tracklets i.e., longer-term spatio-temporal object relations and contexts, which are critical for the segmentation model in Sec. 4.

3.1 Object Tracklet Generation

Given one video sequence with unknown object category, our goal is to firstly generate sets of spatio-temporally associated regions corresponding to the recurrence of the same objects in consecutive frames, i.e., object tracklets. One of the major challenges to generate tracklets is to associate thousands of object proposals from different objects while maintaining spatio-temporal consistence against complex motion, occlusion relationships and appearance variation. To this end, we combine the still-image object detection and generic object tracking together to exploit the discriminative capability of object detector and the ability of handling complex motion from object tracker.

We generate object proposals using [Endres and Hoiem, 2010] which produces bottom-up grouped object-like regions. As the majority object proposals are negative samples and may not correspond to any objects belonging to the target PASCAL VOC 20 classes, we use fast R-CNN [Girshick, 2015] to remove easy negative object proposals whose detection score of the 20 classes are below a certain low threshold

(0.1). The retained object proposals constitute a pool of object candidates Ω, which are assigned with a semantic label with respect to the highest scoring class from R-CNN. We define a subset of high-confidence object candidates $\Omega^+ \subseteq \Omega$ with detection score exceeding a higher threshold (0.5). We generate a negative proposal set Ω^- by randomly sampling negative instances in each frame whose Intersection-over-Union (IoU) with any proposal from Ω^+ are less than 0.3.

For each object class, we generate tracklets by tracking high-confidence proposals from Ω^+ in the video sequence using SR-DCF tracker [Danelljan *et al.*, 2015]. In each iteration, the unassigned proposal in Ω^+ with the highest detection confidence initializes trackings to both ends of the video sequence simultaneously, and any object proposals from Ω whose boxes have a sufficient IoU with the tracker box are considered as candidates to constitute a tracklet. In frames where multiple overlapping candidates are detected, the one with the highest detection confidence is finally chosen to be added to the tracklet. Forward and backward tracklets are subsequently concatenated to form one complete tracklet. This process is performed iteratively until all proposals from Ω^+ are assigned to at least one tracklet. Finally, we extract a set of noisy tracklets denoted as $\mathcal{T}$ which comprises both high- and low-confidence detections. Note that, we do not assume perfect tracking against large motions or heavy occlusions, as short-range spatio-temporal coherence within only a few frames is expected from object tracklets at this stage.

3.2 Transductive Inference Model for Tracklets

The generated object tracklets contain sporadic spurious detections whereas they preserve high-confidence short-range contextual information. We propose a transductive inference model to calibrate the semantics of tracklets with respect to the long-range contexts and global tracklet relationships. We define a weighted space-time graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ on positive and negative object proposals from $\mathcal{T}$ and Ω^- respectively. An undirected acyclic subgraph is naturally formed by temporally connecting all the object proposals from the same tracklet; the induced subgraphs are subsequently connected via the k-nearest neighbors among all the constituent nodes of each pair of tracklets from $\mathcal{T}$; the graph is complemented by adding nodes from Ω^-, connecting to k-nearest other negative nodes or tracklets where the nearest nodes of each tracklet subgraph is connected. This graph accounts for both the short- and longer-range object relationships, with negative examples being sparsely connected to calibrate the sporadic spurious positive object detections. The sparsity of the graph is preserved to facilitate effective and efficient information flowing within structural properties during inference. An exemplar graph is illustrated in Fig. 2.

We solve transductive inference by minimizing an energy function $E(\mathbf{Z})$ with respect to all nodes confidence $\mathbf{Z}$ ($\mathbf{Z} \in [-1, 1]$):

$$\min_Z E(\mathbf{Z}) = \min_Z \mu \sum_{i=1}^{N} ||\mathbf{z}_i - \mathbf{y}_i||^2 + \sum_{i,j=1}^{N} A_{ij} ||\mathbf{z}_i d_i^{-\frac{1}{2}} - \mathbf{z}_j d_j^{-\frac{1}{2}}||^2, \tag{1}$$

where μ is a parameter, and $\mathbf{z}_i$ are the desirable confidence of node i which are regulated by prior confidence $\mathbf{y}_i$. The first term in (1) encourages the inferred confidence to agree with the prior knowledge, whilst the second term promotes the coherence of inferred confidence among adjacent nodes lying in a close vicinity in the feature space. Let the node degree matrix $D = \text{diag}([d_1, \ldots, d_N])$ be defined as $d_i = \sum_{j=1}^{N} A_{ij}$, where $N = |\mathcal{V}|$.

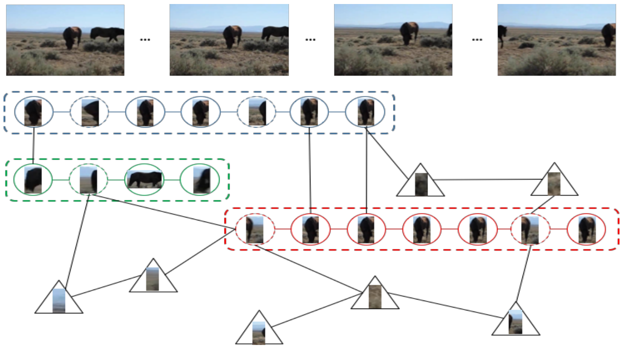

Figure 2: An example of tracklet graph, where rectangles indicate tracklets, circles represent object proposals and triangles stand for the negative boxes. Solid circles are high-confidence proposals whilst dashed circles are weakly detected proposals.

Denoting $\mathbf{S} = D^{-1/2} A D^{-1/2}$, this energy function can be minimized iteratively [Zhou *et al.*, 2004] as

$$\mathbf{Z}^{k+1} = \alpha \mathbf{S} \mathbf{Z}^k + (1 - \alpha) \mathbf{Y}$$

until convergence, where α controls the relative amount of information from its neighboring nodes and its prior knowledge. In each iteration, each node adapts its confidence by receiving the information propagated from its neighboring nodes while preserving its initial confidence. The confidence is adapted symmetrically since S is symmetric. The symmetrically normalized affinity matrix A of $\mathcal{G}_t$ enables the convergence of the following iteration.

Each node in the graph is characterized by the L2-normalized VGG-16 Net [Simonyan and Zisserman, 2014] *fc6* features F_i of its box, and the affinity matrix A of $\mathcal{G}_t$ is computed as the inner-product between the feature vectors of neighboring nodes, i.e., $A_{i,j} =< F_i, F_j >$.

We alternatively solve the optimization problem as a linear system of equations which is more efficient. Differentiating $E(\mathbf{Z})$ with respect to $\mathbf{Z}$ we have

$$\nabla E(\mathbf{Z})|_{\mathbf{Z}=\mathbf{Z}^*} = \mathbf{Z}^* - S\mathbf{Z}^* + \mu(\mathbf{Z}^* - \mathbf{Y}) = 0, \tag{2}$$

which can be transformed as

$$\mathbf{Z}^* - \frac{1}{1+\mu} \mathbf{S} \mathbf{Z}^* - \frac{\mu}{1+\mu} \mathbf{Y} = 0 \tag{3}$$

Denoting $\gamma = \frac{\mu}{1+\mu}$, we have $(I - (1 - \gamma)\mathbf{S})\mathbf{Z}^* = \gamma \mathbf{Y}$. The optimal solution for $\mathbf{Z}$ can be obtained using the preconditioned (Incomplete Cholesky factorisation) conjugate gradient method with very fast convergence.

The initial confidence $\mathbf{Y}$ is initialized based on the detection confidences of R-CNN. Specifically, $\mathbf{Y}$ of positive nodes whose detection confidences are higher than a threshold η ($\eta = 0.1$) is assigned with the detection confidence. The positive nodes with detection confidences below η are deemed as

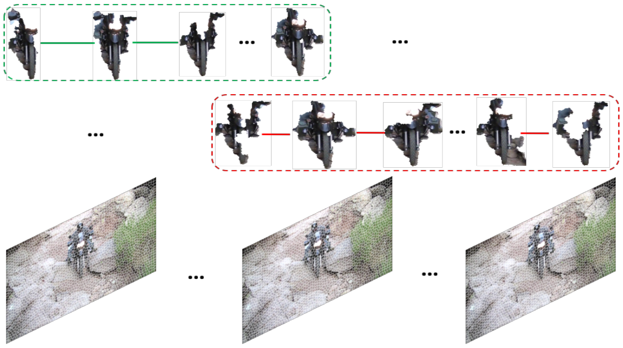

Figure 3: Exemplar tracklets comprising object proposals and underlying superpixel maps.

unlabeled, and their values $\mathbf{Y}$ are initially assigned as 0. $\mathbf{Y}$ of all negative nodes are initially assigned as -1.

The inference process normally involves two separable confidence propagation from labeled (positive or negative) nodes to unlabeled nodes respectively, with initial labels $\mathbf{Y}$ in (1) substituted as $\mathbf{Y}_+$ and $\mathbf{Y}_-$ respectively:

$$\mathbf{Y}_+ = \begin{cases} \mathbf{Y} & \text{if } \mathbf{Y} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and

$$\mathbf{Y}_- = \begin{cases} -\mathbf{Y} & \text{if } \mathbf{Y} < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We propose to combine both propagation processes to produce more efficient and coherent labelling, harnessing the complementary properties of positive and negative nodes. We perform the optimization for two propagation processes simultaneously as follows:

$$\mathbf{Z}^* = \gamma(I - (1-\gamma)\mathbf{S})^{-1}(\mathbf{Y}_+ - \mathbf{Y}_-). \tag{6}$$

Combined inference process enables more efficient and stable optimization while yielding equivalent results to the individual label inferences. The confidences of all object proposals $\mathcal{O}$ are thus calibrated by incorporating local and global tracklet relationship; object proposals with inferred confidence $\mathbf{Z} < 0$ are deemed as false positives and are consequently removed from the constituting tracklets.

4 Hierarchical Graphical Model for Segmentation

Given the calibrated tracklets, we formulate the semantic object segmentation as a superpixel labeling problem. We propose a novel hierarchical graphical model to combine bottom-up motion and appearance cues with top-down recognition, long-term object relations and spatio-temporal contexts under one framework, to assign each node with labels $\mathbf{x} \in \{0, K\}$.

We define a hierarchical graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$ for constructing the segmentation model. Firstly the tracklets are modeled as top layer graph nodes, connecting with each other. The second layer is comprised of object proposals which are modeled as undirected acyclic sub-graph within each tracklet. Object proposal nodes are connected to the associated tracklet and their constituent superpixels. The bottom layer consists of superpixels connected to represent the local spatial and temporal relationships. Exemplar tracklets which comprises object proposals and the underlying superpixel maps are shown in Fig. 3

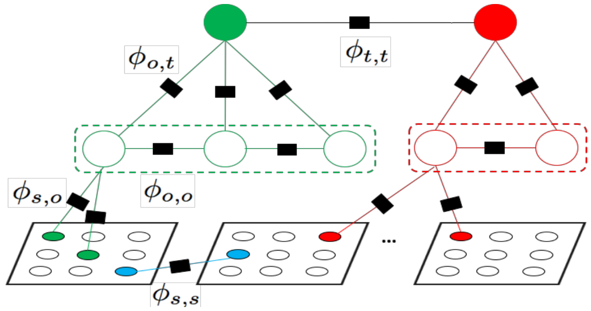

Figure 4: CRF of hierarchical graph for various pairwise potentials.

This hierarchical graph can be formulated as a Conditional Random Field (CRF) for nodes with label $\mathbf{x} \in \{0, K\}$ where an energy function can be defined as:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}\, E_U(\mathbf{x}) + E_P(\mathbf{x}). \tag{7}$$

where E_U and E_P denote the unary and pairwise potentials respectively. An illustration of various pairwise potentials of the hierarchical graphical model is shown in Fig. 3. We adopt alpha expansion [Boykov *et al.*, 2001] to minimise (7) and the resulting label assignment gives the semantic object segmentation of the present categories.

4.1 Unary Potentials

We define unary potentials for nodes to measure the compatibility of the observed feature and their labels,

$$E_U(\mathbf{x}) = \psi_s(\mathbf{x}) + \psi_o(\mathbf{x}) + \psi_t(\mathbf{x}) \tag{8}$$

where ψ_s, ψ_o and ψ_t correspond to the sets of superpixels, object proposals and tracklets respectively.

The unary potentials of superpixel nodes combine both the deep feature and color-based appearance models:

$$\psi_s(\mathbf{x}) = \sum_{i \in \mathcal{S}} \Phi_d(x_i) + \Phi_c(x_i).$$

where $\Phi_d(\cdot)$ is the deep feature model term and $\Phi_c(\cdot)$ is the color-based appearance model term. We learn a SVM model based on hierarchical CNN features [Ma *et al.*, 2015] and GMM model based on CIE Lab colors, by sampling from the calibrated tracklets.

The unary potentials of object proposals ψ_o are computed by averaging the unary potentials of their constituent superpixels; similarly, the unary potentials of tracklet nodes ψ_t are computed by averaging over their constituent object proposals.

4.2 Pairwise Potentials

We define the pairwise potentials resembling Potts model to encourage both local and global contexts of labelling while preserving discontinuity in the data,

$$\begin{aligned} E_P(\mathbf{x}) =& \phi_{s,s}(\mathbf{x}) + \phi_{s,o}(\mathbf{x}) + \\ & \phi_{o,o}(\mathbf{x}) + \phi_{o,t}(\mathbf{x}) + \phi_{t,t}(\mathbf{x}) \end{aligned} \tag{9}$$

where the paired subscripts from $\{s, o, t\}$ indicate the pairwise potentials between different types of nodes.

The spatial and temporal pairwise potentials of superpixels $\phi_{s,s} = \{\phi^s_{s,s}, \phi^t_{s,s}\}$ are defined as follows. The spatial pairwise potential $\phi^s_{s,s}$ which penalizes different labels assigned to spatially adjacent superpixels is defined as,

$$\phi^s_{s,s} = \sum_{i,j \in \mathcal{S}} \delta(x_i, x_j) \frac{e^{-d^c(i,j)}}{d^s(i,j)}$$

where $\delta(\cdot)$ is the Kronecker delta, the functions $d^s(i,j)$ and $d^c(i,j)$ computes the spatial and color distance respectively between spatially neighboring superpixel nodes i and j as $d^c(i,j) = \frac{||c_i - c_j||^2}{2 < ||c_i - c_j||^2 >}$, where $||c_i - c_j||^2$ is the squared Euclidean distance between two adjacent superpixel nodes in CIE Lab colorspace.

The temporal pairwise potential is defined over edges where superpixels are temporally connected by at least one optical flow motion vector on consecutive frames. The temporal pairwise potential is defined as

$$\phi^t_{s,s} = \sum_{i,j \in \mathcal{S}} \delta(x_i, x_j) \frac{e^{-d^c(i,j)}}{d^t(i,j)},$$

where $d^t(i,j)$ computes the temporal distance between i and j which is measured by the the ratio of pixels within the two superpixel nodes that are connected by motion vectors over the union of two superpixels.

The pairwise potential between superpixel and associated object proposal is defined to encourage the superpixels which constitute the same object proposal to be assigned with the same label while still allowing some of them to have different labels,

$$\phi_{s,o} = \sum_{i \in \mathcal{S}, j \in \mathcal{O}} \delta(x_i, x_j) \frac{e^{-\sigma_j}(1 - |p_i - p_j|) N^p_i}{N^p_j} \qquad (10)$$

where p_i denotes the likelihood of node i to be labeled as x_i based on trained hierarchical CNN feature and GMM color model, σ_j is the standard deviation of superpixel unary potentials constituent object proposal node j, N^p_i and N^p_j are the cardinalities of node i and j respectively in terms of pixel counts. This potential computes the penalty of assigning different labels to superpixel node i and object proposal node j. This penalty is lower if either node i is small relative to j, or node i and j appear different, or object proposal node j contains large uncertainties.

Pairwise potential between object proposals $\phi_{o,o}$ is following the inner-product of paired CNN features defined in Sec. 3.2, to exploit higher level semantics in deep feature space

$$\phi_{o,o} = \sum_{i,j \in \mathcal{O}} \delta(x_i, x_j) e^{-(1 - <F_i, F_j>)} \qquad (11)$$

Potential function $\phi_{o,t}$ follows a similar definition as in (10),

$$\phi_{o,t} = \sum_{i \in \mathcal{O}, j \in \mathcal{T}} \delta(x_i, x_j) \frac{e^{-\sigma_j}(1 - |p_i - p_j|)}{N^o_j} \qquad (12)$$

where N^o_j stands for the cardinality of tracklet node j in terms of object proposal count. This function encourages the object proposal in each tracklet to be labeled the same, except they are of divergent confidence or tracklet node j consists of proposals with significant variance of uncertainties. $\phi_{t,t}$ is computed similar to $\phi_{o,o}$, where the tracklet node feature is computed by averaging all the CNN feature vectors of constituent object proposals.

Table 1: Intersection-over-union overlap accuracies on YouTube-Objects Dataset

	LDW	DSW	SSW	DAW	DTS	SLM	Ours
Plane	0.517	0.178	0.758	**0.760**	0.744	0.749	0.757
Bird	0.175	0.198	0.608	0.747	0.721	0.754	**0.766**
Boat	0.344	0.225	0.437	0.588	0.585	0.623	**0.666**
Car	0.347	0.383	0.711	0.659	0.600	0.676	**0.758**
Cat	0.223	0.236	0.465	0.557	0.457	0.548	**0.624**
Cow	0.179	0.268	0.546	0.675	0.612	0.679	**0.720**
Dog	0.135	0.237	0.555	0.574	0.552	0.582	**0.671**
Horse	0.267	0.140	0.549	**0.575**	0.566	0.503	0.526
Mbike	0.412	0.125	0.424	**0.569**	0.421	0.516	0.547
Train	0.250	0.404	0.358	**0.430**	0.367	0.358	0.392
Avg.	0.285	0.239	0.541	0.613	0.562	0.599	**0.643**

5 Experiments

In order to evaluate the performance of semantic video object segmentation, many motion segmentation or figure-ground segmentation datasets are not suitable due to either the ambiguous object annotation in ground-truth (one label for all foreground moving objects or no annotation for static objects) or the insufficient number of videos/frames per object class. We evaluate on two large-scale video object segmentation datasets, YouTube-Objects [Prest *et al.*, 2012], egoMotion [Shankar Nagaraja *et al.*, 2015], which are totally over 30,000 frames. The categories of these two datasets are subsets of the pretrained 20 classes of PASCAL VOC 2012 in R-CNN.

The YouTube-Objects dataset has become the dataset on which state-of-the-art methods report their results. YouTube-Objects consists of videos from 10 object classes with pixel-level ground truth for totally more than $20,000$ frames. These videos are very challenging and completely unconstrained, with objects of similar colour to the background, fast motion, non-rigid deformations, and fast camera motion. The egoMotion dataset consists of 11882 frames from 4 object classes (*cars, cats, chairs, dogs*), where the main challenge is the dominant camera motion. We measure the segmentation performance using the standard the average IoU overlap as accuracy metric, $IoU = \frac{S \cap GT}{S \cup GT}$.

5.1 YouTube-Objects

We compare our approach with 5 state-of-the-art semantic video object segmentation approaches on this dataset, i.e., [Prest *et al.*, 2012] (LDW), [Tang *et al.*, 2013] (DSW), [Zhang *et al.*, 2015] (SSW), [Wang *et al.*, 2016] (DAW) and [Drayer and Brox, 2016] (DTS).

As shown in Table 1, our approach surpasses the competing methods in 6 out of 10 classes, with gains up to 3% in average over the best competing method DAW. It is worth noting that DAW can not separate interacting objects with similar colors (e.g., motorbike and rider in Fig. 5) due to the lack of long-range object interactions and behaviors information, whereas our decreased performance in *Mbike* is caused by the inaccurate ground-truth which labels interact-

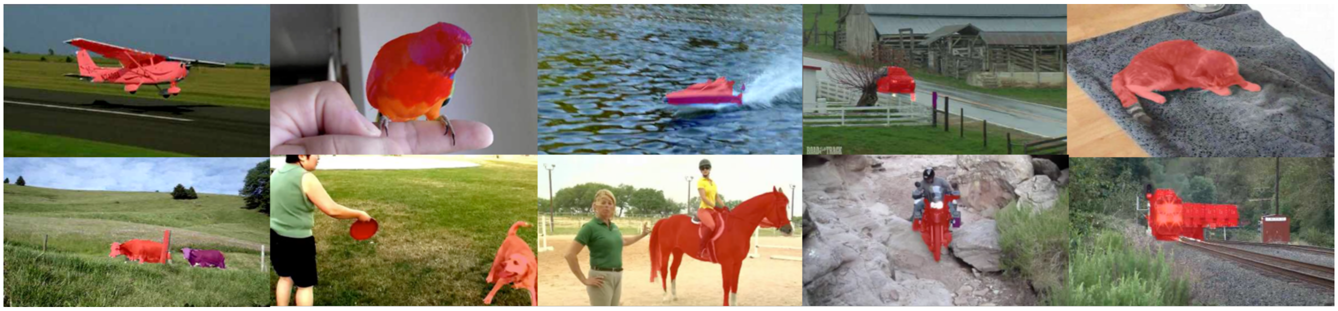

Figure 5: Qualitative results for YouTube-Objects Dataset.

ing (motorbike and rider) objects as one object. Other underperformances in *Plane*, *Horse* and *Train* compared with DAW are mainly owing to the noisy boundaries of object proposals, where DAW uses multiple overlapping proposals to compensate for the mis-segmentations of proposals.

Our approach beats DTS and SSW in all categories with large margins of 8.1% and 10.2% respectively by exploiting higher level contextual information and high-quality tracklets. Our method doubles or triples the accuracy of DSW and LDW in most categories, with an exception in *Train* category where DSW slightly surpasses our approach. This is probably owing to that DSW uses a large number of similar training videos which may capture objects in rare view. We also compare with a baseline scheme of our proposed approach by replacing the hierarchical model with a single-layer graphical model (SLM) consisting of only superpixels and the rest remains the same. Comparing with the baseline, a gain of 4.4% is benefited by adopting the proposed hierarchical model. It is remarkable that baseline scheme SLM outperforms DTS which incorporates tracklets locally in a single layer model, indicating the effectiveness of our transductive inference model in generating high-quality tracklets.

Fig. 5 shows some qualitative results of the proposed algorithm on YouTube-Objects dataset, where it detects and delineates the stationary objects (horse), rapid moving objects (boat), rigid (car, train) or non-rigid objects (cat, dog), multiple objects (cow), as well as interacting objects exhibiting similar color or motion (cat, horse, motorbike).

5.2 egoMotion

We compare with four automatic object segmentation methods [Papazoglou and Ferrari, 2013] (FOS), [Ochs *et al.*, 2014] (MLT), [Keuper *et al.*, 2015] (MTS), [Drayer and Brox, 2016] (DTS), and the state-of-the-art CNN based semantic image segmentation method [Long *et al.*, 2015] (FCN). As shown in Table 2 , our method surpasses the competing methods on 2 out of 4 classes, with a large average precision improvement of 5.1% over the best competing method DTS. DTS, which heavily relies on local motion cues, outperforms our approach on rigid object class *Chair* but exhibits poorer performance on *Car* and non-rigid object classes *Cat* (see Fig. 1) and *Dog* where our approach exploits long-range contexts and object relations to better deal with dynamic object appearance variations and agile motion. Per-frame segmentation using FCN shows similar behavior as DTS, and also fails to segment object with unusual viewpoint as shown in Fig. 1. Fig. 6 shows the qualitative results of the proposed algorithm on egoMotion dataset, which confirms its superior robustness in handling rigid (*Car*, *Chair*) or non-rigid objects (*Cat*, *Dog*).

Table 2: Intersection-over-union overlap accuracies on egoMotion Dataset

	MLT	MTS	FOS	FCN	DTS	Ours
Car	0.336	0.379	0.476	**0.861**	0.780	0.786
Cat	0.135	0.453	0.566	0.166	0.657	**0.843**
Chair	0.162	0.198	0.595	0.390	**0.735**	0.708
Dog	0.417	0.534	0.642	0.471	0.752	**0.792**
Average	0.263	0.391	0.570	0.472	0.731	**0.782**

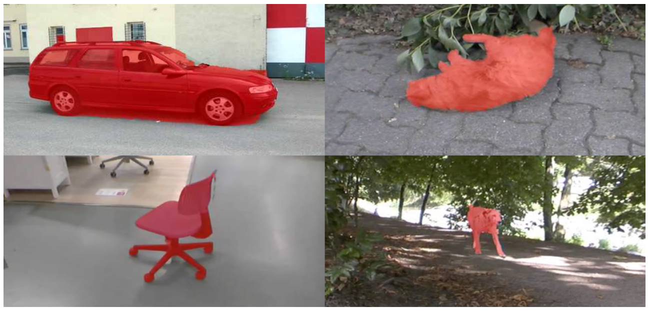

Figure 6: Qualitative results for egoMotion Dataset.

6 Conclusion

We have proposed a novel tracklet-object-aware hierarchical model for semantic video object segmentation, which jointly models and segments object incorporating both local and global context features and longer-term object interactions and behaviors. To deal with a large number of noisy object hypotheses, we further proposed a transductive inference model which is capable of calibrating short-range noisy object tracklets with respect to long-range object relations and high-level context cues. We have demonstrated that our approach advanced the state-of-the-art performance on two large scale video datasets.

References

[Ayvaci and Soatto, 2012] Alper Ayvaci and Stefano Soatto. Detachable object detection: Segmentation and depth ordering from short-baseline video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):1942–1951, 2012.

[Banica *et al.*, 2013] Dan Banica, Alexandru Agape, Adrian Ion, and Cristian Sminchisescu. Video object segmentation by salient segment chain composition. In *ICCV Workshops*, pages 283–290, 2013.

[Boykov *et al.*, 2001] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[Choi and Savarese, 2012] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, pages 215–230. Springer, 2012.

[Danelljan *et al.*, 2015] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, pages 4310–4318, 2015.

[Drayer and Brox, 2016] Benjamin Drayer and Thomas Brox. Object detection, tracking, and motion segmentation for object-level video segmentation. *arXiv preprint arXiv:1608.03066*, 2016.

[Endres and Hoiem, 2010] Ian Endres and Derek Hoiem. Category independent object proposals. In *ECCV*, pages 575–588, 2010.

[Giordano *et al.*, 2015] Daniela Giordano, Francesca Murabito, Simone Palazzo, and Concetto Spampinato. Superpixel-based video object segmentation using perceptual organization and location prior. In *CVPR*, pages 4814–4822, 2015.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[Hartmann *et al.*, 2012] Glenn Hartmann, Matthias Grundmann, Judy Hoffman, David Tsai, Vivek Kwatra, Omid Madani, Sudheendra Vijayanarasimhan, Irfan A. Essa, James M. Rehg, and Rahul Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV Workshop*, pages 198–208, 2012.

[Keuper *et al.*, 2015] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *ICCV*, pages 3271–3279, 2015.

[Lee *et al.*, 2011] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *ICCV*, pages 1995–2002, 2011.

[Liu *et al.*, 2014] Xiao Liu, Dacheng Tao, Mingli Song, Ying Ruan, Chun Chen, and Jiajun Bu. Weakly supervised multiclass video segmentation. In *CVPR*, pages 57–64, 2014.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Ma *et al.*, 2015] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, pages 3074–3082, 2015.

[Ochs *et al.*, 2014] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1187–1200, 2014.

[Papazoglou and Ferrari, 2013] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, pages 1777–1784, 2013.

[Prest *et al.*, 2012] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, pages 3282–3289, 2012.

[Shankar Nagaraja *et al.*, 2015] Naveen Shankar Nagaraja, Frank R Schmidt, and Thomas Brox. Video segmentation with just a few strokes. In *ICCV*, pages 3235–3243, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sundberg *et al.*, 2011] Patrik Sundberg, Thomas Brox, Michael Maire, Pablo Arbelaez, and Jitendra Malik. Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR*, pages 2233–2240, 2011.

[Tang *et al.*, 2013] Kevin D. Tang, Rahul Sukthankar, Jay Yagnik, and Fei-Fei Li. Discriminative segment annotation in weakly labeled video. In *CVPR*, pages 2483–2490, 2013.

[Taylor *et al.*, 2013] Brian Taylor, Alper Ayvaci, Avinash Ravichandran, and Stefano Soatto. Semantic video segmentation from occlusion relations within a convex optimization framework. In *EMMCVPR*, pages 195–208. Springer, 2013.

[Wang and Wang, 2016] Huiling Wang and Tinghuai Wang. Primary object discovery and segmentation in videos via graph-based transductive inference. *Comput. Vis. Image Underst.*, 143(2):159–172, 2016.

[Wang *et al.*, 2009] Chaohui Wang, Martin de La Gorce, and Nikos Paragios. Segmentation, ordering and multi-object tracking using graphical models. In *ICCV*, pages 747–754, 2009.

[Wang *et al.*, 2015] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, pages 3395–3402, 2015.

[Wang *et al.*, 2016] Huiling Wang, Tapani Raiko, Lasse Lensu, Tinghuai Wang, and Juha Karhunen. Semi-supervised domain adaptation for weakly labeled semantic video object segmentation. In *ACCV*, 2016.

[Xiao and Lee, 2016] Fanyi Xiao and Yong Jae Lee. Track and segment: An iterative unsupervised approach for video object proposals. In *CVPR*, pages 551–566, 2016.

[Zhang *et al.*, 2015] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, and Changqun Xia. Semantic object segmentation via detection in weakly labeled video. In *CVPR*, pages 3641–3649, 2015.

[Zhao and Fu, 2015] Handong Zhao and Yun Fu. Semantic single video segmentation with robust graph representation. In *IJCAI*, pages 2219–2226, 2015.

[Zhou *et al.*, 2004] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Sch. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004.