Multiclass Semantic Video Segmentation with Object-level Active Inference

Buyu Liu
ANU/NICTA
buyu.liu@anu.edu.au

Xuming He
NICTA/ANU
xuming.he@nicta.com.au

Abstract

We address the problem of integrating object reasoning with supervoxel labeling in multiclass semantic video segmentation. To this end, we first propose an object-augmented dense CRF in spatio-temporal domain, which captures long-range dependency between supervoxels, and imposes consistency between object and supervoxel labels. We develop an efficient mean field inference algorithm to jointly infer the supervoxel labels, object activations and their occlusion relations for a moderate number of object hypotheses. To scale up our method, we adopt an active inference strategy to improve the efficiency, which adaptively selects object subgraphs in the object-augmented dense CRF. We formulate the problem as a Markov Decision Process, which learns an approximate optimal policy based on a reward of accuracy improvement and a set of well-designed model and input features. We evaluate our method on three publicly available multiclass video semantic segmentation datasets and demonstrate superior efficiency and accuracy.

1. Introduction

Semantic scene parsing, which aims to segment images into conherent and semantically meaningful regions, has recently made much progress by incorporating high-level visual information, such as scene context and objects [8, 5, 14], and jointly solving multiple related vision tasks [32, 22, 26]. Such integration with object and scene-level reasoning not only provides more global cues and longer-range constraints for pixel-level labeling, but also enables us to generate a deeper parsing of images with multiple properties, ranging from object instance [26] to scene geometry [9].

However, such object-aware strategies require many hypotheses of object instances and their relations to accommodate uncertainty in object detection and localization [22, 25]. This leads to increasingly larger structure and/or higher complexity of the resulting models on pixels and objects, which has several drawbacks. First, with more object classes and their relations added, it becomes challenging to develop efficient inference algorithm in the joint models. Greedy strategies, such as alternating inference between pixels and objects, have been used to address such difficulty [26], which do not exploit the full potential of joint modeling. In addition, object hypotheses have to be pruned based on heuristic thresholds to provide a balance between precision and recall, which is tedious for adding new object classes. Furthermore, it is even more difficult to extend this to dynamic scene parsing in videos, as the number of object hypotheses may increase greatly due to object motion and longer image sequences.

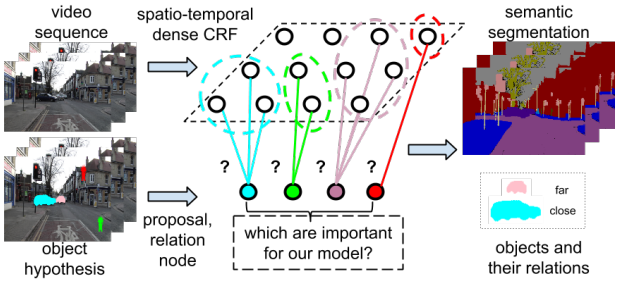

Figure 1: Overview of our approach. Example of the object-augmented dense CRF model. Our active inference adaptively selecting subgraphs thus improve the inference efficiency.

An alternative approach to addressing this difficulty is to adaptively select a subset of model components which is most informative for inference and within a budget constraint, or active inference [21, 29]. This allows users to achieve a balance between efficiency and performance in a principled way. Most of previous work, however, focuses on improving efficiency in feature computation except a few on model structure [28].

In this work, we aim to address the problem of joint pixel and object inference in semantic video segmentation. We take a hypothesize-and-verify approach [10, 17], in which we generate a pool of object hypotheses and formulate video segmentation as a joint labeling of pixels and object hypotheses. To handle a large number of object hypotheses, we adopt an active inference strategy at object level to select

an optimal subset of hypotheses for joint inference.

Specifically, given a video sequence, we first build an object-augmented dense CRF model that consists of a supervoxel layer and an object layer. The supervoxel layer, modeled by a dense CRF [12], captures long-range spatio-temporal dependency between supervoxels, while the object layer imposes consistency between semantic labeling of supervoxels and objects, as well as valid occlusion relation between overlapping objects. We develop an efficient mean field inference algorithm for the case with a moderate number of object hypotheses.

Inspired by [29], we propose to select an informative subset of objects and their relation nodes for scaling up inference with many object hypotheses. To this end, we build a set of subgraphs corresponding to the object hypotheses, which are selected in our inference procedure. We formulate the subgraph selection as a Markov Decision Process (MDP) and develop a learning approach to search the optimal policy for sequentially choosing the most informative subgraph. We define a reward function using the improvement on pixel-level per-class accuracy, and learn an approximate policy based on Q-learning [15]. Our policy takes long-range features generated by both current model uncertainty and video input, and predicts the most valuable subgraph to choose in next step. Furthermore, we also use an imitation learning scheme [1] to efficiently train a local classifier that approximates the optimal decision.

We evaluate our approach on three publicly available semantic video segmentation datasets. We demonstrate that our learned policy is capable of selecting informative object hypotheses and relations, leading to much simpler model structure and comparable or even higher segmentation accuracy. Our contribution has two folds. 1) We develop an object-augmented dense CRF semantic video segmentation with efficient mean field inference. 2) We explore the value of object hypotheses for semantic labeling and propose a MDP-based method to sequentially choose most informative objects during inference. Our method is capable of generating compact model structure, which is useful for achieving a balance between efficiency and accuracy.

2. Related work

Semantic scene understanding has attracted much attention in recent years and a large number of methods have been proposed [6]. High-level visual information has played an important role in improving the state-of-the-art of semantic segmentation. Early approaches integrate object information into semantic labeling task by imposing consistency between pixel labels and object detection outputs [14]. More recently, object instances and pixel labels are jointly inferred through structural models defined on both entities [22, 26]. In particular, multiple aspects of scene property, ranging from pixel labels to scene categories, have been combined to provide more effective constraints in understanding the scene as a whole [32].

Despite the progress in static setting, much less work has been done in semantic video segmentation, partially due to lack of training data and greater complexity of the problem. The main focus of many existing methods is to model temporal coherence of pixel labeling [4, 19, 25], which lacks object-level reasoning. Other methods construct 3D models of statics scenes based on structure from motion [2, 13] for pixel labeling, but are limited in parsing moving objects. Object detection and tracking is first integrated with semantic video segmentation by Wojek et al [31, 30]. However, they deal with each object instance separately and no occlusion is modeled.

Most video object segmentation approaches focus on single class foreground objects. Wang et al [27] consider foreground object segmentation, tracking and occlusion reasoning with a unified MRF model. Lezama et al [16] use optical flow based long-term trajectories to discover moving objects. Nevertheless, they do not include multiple object classes, and thus they cannot capture semantic context of moving objects. Taylor et al [24] infer multiple semantic classes and occlusion relationship in video segmentation. Unlike our method, they do not represent individual object instance and reason their relations.

Similar to our work, Liu et al [17] also consider joint inference of supervoxel labels and object instances. The main difference of our approach is introducing active inference for handling a large number of object hypotheses. In addition, our model is based on the fully-connected CRF [12] and is augmented with object potentials.

Active or budgeted inference has recently been introduced to improve efficiency of inference in structured models [29, 28, 21]. Roig et al [21], makes use of perturb-and-MAP inference model to compute and select informative unary potentials. Weiss et al [29] develop a reinforcement learning framework for feature selection in structured models. Our method adopts the same framework as [29], but we focus on selecting object hypotheses instead of input-related features. In [20], a local classifier is learned to select views for multi-view semantic labeling. We build our selection policy based on a principled MDP framework. In addition, anytime structural learning has been proposed in[7] for model learning. In this work, we mainly focus on improving inference procedure.

3. Object-augmented spatio-temporal CRF

We first introduce our spatio-temporal CRF model for multi-class semantic video segmentation. Our model consists of mainly two parts: at supervoxel-level, we build a pairwise CRF with dense connections [12], which captures the long-range dependency between frames; we also apply object detectors and a tracking method to propose object

hypotheses, which augment the supervoxel CRF with object and object relation potentials [22, 17]. The joint CRF model has a mixed structure with both dense pairwise and sparse ternary potentials. We develop an efficient mean field inference for estimating node marginal distributions, which is also critical for the active inference in Sec 4.

3.1. Model setup and notations

Given a video sequence $\mathcal{T}$, we first compute its supervoxel representation based on [3]. We denote the semantic class of the i-th supervoxel as l_i with $i \in \{1, ..., N\}$ indexing all the supervoxels. The semantic labeling of the full sequence is denoted by $\mathbf{L} = (l_1, ..., l_N)$.

We then generate a set of object trajectory hypotheses from object detection and tracking efficiently as in [17]. We denote the hypothesis pool as $\mathcal{O}$ and $m \in \{1, ..., M\}$ indexing all the hypotheses. For the m-th hypothesis, we introduce a binary variable d_m to indicate whether it is a true positive detection or background, and o_m to represent the spatio-temporal regions it occupies. The object states of the full sequence is denoted by $\mathbf{D} = (d_1, ..., d_M)$.

We model the object relations by considering the relative depth ordering between them. To this end, we divide the hypotheses into two groups, one of which represents the singleton objects with no overlap with others (on image plane), and the other consists of overlapping object instances. We denote the singleton set as $\mathcal{S}$ and the set of all pairs of overlapping hypotheses as $\mathcal{P}$.

We assume the relative depth ordering of an object pair keeps unchanged in a short period, and for each pair $(m, n) \in \mathcal{P}$, introduce a variable $h_{mn} \in \{-1, 0, 1\}$ to describe their occlusion relations. The value -1 and 1 denotes m-th hypothesis is occluded by or occludes n-th hypothesis respectively and 0 denotes there exists no occlusion relations between current pair, or equivalently, at least one of the hypotheses is background. To handle longer sequences, we divide them into short video chunks and introduce a set of $\{h_{mn}\}$ for each chunk. We use $\mathbf{H} = \{h_{mn}\}_{(m,n)\in\mathcal{P}}$ to denote the states of ordering for the entire sequence.

We formulate the semantic video segmentation as a joint labeling problem of supervoxels, object hypotheses and object relations, and construct an object-augmented spatio-temporal CRF model on those entities. The overview of our model is in Figure 2. We define the overall energy function of our CRF, $E(\mathbf{L}, \mathbf{D}, \mathbf{H}|\mathcal{T})$, as follows:

$$E(\mathbf{L}, \mathbf{D}, \mathbf{H}|\mathcal{T}) = E_v(\mathbf{L}|\mathcal{T}) + \sum_{m\in\mathcal{S}} E_s(\mathbf{L}, d_m|\mathcal{T}) + \sum_{p\in\mathcal{P}, p=(m,n)} E_r(\mathbf{L}, d_m, d_n, h_{mn}|\mathcal{T}) \quad (1)$$

where $E_v(\mathbf{L}|\mathcal{T})$ denotes the potentials at supervoxel level, $E_s(\mathbf{L}, d_m|\mathcal{T})$ are potentials for the singleton object hypothesis m, and $E_r(\mathbf{L}, d_m, d_n, h_{mn}|\mathcal{T})$ are the potentials for the object pair relations in $\mathcal{P}$. We will describe the details of those potentials in the following, and omit $\mathcal{T}$ for clarity.

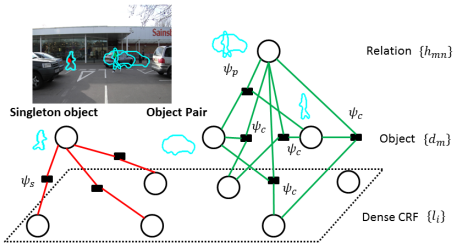

Figure 2: Example of subgraphs for singleton and object pair potentials. Different colors denote different subgraphs.

3.2. Supervoxel pairwise CRF E_v

At supervoxel level, we build a dense pairwise CRF [12] to capture long-range dependency of supervoxel labels in spatio-temporal domain. Specifically, the potential functions of $E_v(\mathbf{L})$ consists of two terms,

$$E_v(\mathbf{L}) = \sum_{i=1}^{N} \phi_v(l_i) + \sum_{i\neq j} \psi_v(l_i, l_j) \quad (2)$$

where $\phi_v(l_i)$, $\psi_v(l_i, l_j)$ are the unary and pairwise term, respectively.

Supervoxel unary. The unary term specifies the cost of assigning l_i to supervoxel i and is defined as $\phi_v(l_i) = -\log P_l(l_i)$, where $P_l(l_i)$ is the output of a probabilistic classifier. Here we train a random forest classifier based on color, texture, HoG and geometric features [17].

Pairwise Potential We introduce the dense pairwise term $\psi_v(l_i, l_j)$ to enforce spatio-temporal smoothness of labeling. We define a contrast-sensitive two-kernel potential as in [12]. The appearance kernel uses the CIE-Lab color space, and both the appearance and smoothness kernel take spatial location and time in $\mathcal{T}$ as position features.

3.3. Singleton object potentials E_s

For each singleton object hypothesis $m \in \mathcal{S}$, we define a unary potential ϕ_o to encode the likelihood of being true positive, and a pairwise potential ψ_s to impose the consistency between object activation and supervoxel labeling,

$$E_s(\mathbf{L}, d_m) = \phi_o(d_m) + \sum_{i\in o_m} \psi_s(d_m, l_i) \quad (3)$$

where $\{i \in o_m\}$ include all the supervoxels occupied by the hypothesis m.

Object unary The object unary term $\phi_o(d_m)$ models the cost of activating m-th hypothesis with the following form,

$$\phi_o(d_m) = \alpha_{mdl} d_m - \alpha_o \log \frac{p_c(d_m)}{1 - p_c(d_m)} d_m \quad (4)$$

where $p_c(d_m)$ is the probability of activating m-th hypothesis, which is generated by a classifier. We extract detector output, position, object mask size, appearance and edge distance as features and train a multiclass logistic regressor to predict $p_c(d_m)$. α_o is the weight for the classifier score and α_{mdl} is to enforce the sparsity of detection.

Object and supervoxel consistency The pairwise potential $\psi_s(d_m, l_i)$ penalizes the inconsistency between the class of activated hypothesis m and supervoxel labeling l_i,

$$\psi_s(d_m, l_i) = \alpha_s \frac{V_i}{V_{o_m}} d_m [\![l_i \neq c_m]\!] \tag{5}$$

where c_m denotes the m-th hypothesis's object class. V_i denotes the volume of supervoxel i and V_{o_m} is the volume of object hypothesis m. We use $[\![f]\!]$ as the indicator function, which equals to 1 if f is true and 0 otherwise.

3.4. Object-pair relation potentials E_r

We model the relative depth ordering relationship between every overlapping object pair using the object-pair relation potentials. To this end, we design a potential that encodes the likelihood of valid object pair relation and the consistency between object pair label configuration and supervoxel labeling. Specifically,

$$E_r(\mathbf{L}, d_m, d_n, h_{mn}) = \phi_h(h_{mn}) + \psi_p(h_{mn}, d_m, d_n) + \sum_{k \in \{m,n\}} \left(\phi_o(d_k) + \sum_{i \in o_k} \psi_c(h_{mn}, l_i, d_k) \right) \tag{6}$$

where ϕ_h is the unary potential of relative ordering, ψ_p encodes the depth ordering consistency between objects within a pair, and ψ_c enforces consistency between supervoxel labeling and object pair configuration.

Ordering unary The ordering unary term $\phi_h(h_{mn})$ models the likelihood of h_{mn} taking one of the three states, $\{-1, 0, 1\}$. We define $\phi_h(h_{mn}) = -\alpha_h \log P_h(h_{mn})$, where P_h is the probabilistic score from a classifier. We generate P_h by taking the object unaries, positions, sizes, appearance (color histogram), number of terminated optical flows as features for each object pair and training a random forest classifier to predict the ordering probability.

Ordering consistency We define ψ_p to enforce that object state d_m, d_n and relation variable h_{mn} should be consistent. h_{mn} is nonzero iff both object are true positives, and two hypotheses cannot be heavily overlapped:

$$\psi_p(h_{mn}, d_m, d_n) = \alpha_{inf} \Big([\![\neg(d_m d_n = 1 \wedge h_{mn} \neq 0)]\!] + [\![\neg(d_m d_n = 0 \wedge h_{mn} = 0)]\!] + [\![\tfrac{o_m \cap o_n}{o_m \cup o_n} > \tau]\!] d_m d_n \Big) \tag{7}$$

where α_{inf} is a large penalty and τ is a threshold for non-maximum suppression of overlapped objects.

Supervoxel occlusion consistency The occlusion term penalizes the inconsistency between supervoxel labeling and object pair configuration w.r.t the occlusion relations. We enforce the supervoxels in the overlapping regions to be consistent with the foremost object that is activated,

$$\psi_c(h_{mn}, l_i, d_m) = \alpha_c \frac{V_i}{V_{o_m}} \Big([\![l_i \neq c_m \wedge h_{mn} = 1]\!] + [\![l_i \neq c_m \wedge h_{mn} = 0 \wedge d_m = 1]\!] + [\![l_i \neq c_m \wedge i \in \{o_m \setminus o_n\} \wedge h_{mn} = -1]\!] \Big) \tag{8}$$

where α_c is the weighting coefficient, c_m, V_i and V_{o_m} are defined in Eq (5). $\{o_m \setminus o_n\}$ is the set of supervoxels occupied by the object m but not by n.

3.5. Model Inference and learning

Model inference The object-augmented spatial-temporal CRF in Eq (1) has a dense pairwise potential on supervoxels and a sparse ternary potential on object and relation nodes. The inference is challenging due to this mixed structure. We develop a mean field approximate inference algorithm to jointly infer supervoxel, object and object relation labels. Specifically, we approximate the joint model distribution by a fully factorized model $q(\mathbf{L}, \mathbf{D}, \mathbf{H}) = \prod_i q_v(l_i) \prod_m q_o(d_m) \prod_p q_p(h_p)$ where $p = (m, n)$ indexes pairs. The mean field updating equations can be derived by minimizing the KL divergence between model distribution and q. For clarity, we only introduce the updating equation for $q_v(l_i)$ in the following and leave the rest to the supplementary material:

$$\hat{q}_v(l_i) \propto \sum_{i \neq j} \langle \psi_v(l_i, l_j) \rangle_{q_v(l_j)} + \sum_{m \in \mathcal{S}} \sum_{i \in o^m} \langle \psi_s(d_m, l_i) \rangle_{q_o(d_m)} + \sum_{p \in \mathcal{P}} \sum_{\substack{k \in \{m,n\} \\ i \in o_k}} \langle \psi_c(h_p, l_i, d_k) \rangle_{q_o(d_k) q_p(h_p)} + \phi_v(l_i) \tag{9}$$

We note that the first summation has the same form as in [12], and can be computed efficiently. The next two terms are summed over sparse connections, which can also be easily computed. The overall mean field algorithm shares similar efficiency as the original dense CRF as long as the number of object hypotheses and relations is moderate.

Given the approximate marginals, $q_v(l_i)$, $q_o(d_m)$, and $q_p(h_p)$, we take the modes of marginals to obtain the supervoxel and object label predictions [12]. In particular, for supervoxel i, we compute $l_i^* = \arg\max_{l_i} q_v(l_i)$. Empirically, this also provides us consistent labeling results over object activation and object relations.

Parameter learning In this work, we assume the model parameters are pre-learned and focus on the inference problem. For completeness, we briefly discuss the model learning. We use the piece-wise learning [23] to incrementally

learn parameters while any other suitable learning method can also be applied. We first learn the weights and kernel parameters in the supervoxel pairwise CRF. Then we estimate the parameters of the singleton object potentials. Here we incrementally learn parameters for *Car*, *Pedestrian* and *Bicyclist* class and treat all hypotheses as singletons. Finally, we tune the parameters of the object relation potentials while keeping the supervoxel and object potentials fixed. All parameters are learned on validation set by grid search and we choose the set of parameters that maximizes the per-class accuracy. The constant α_{inf} is set to be 10^{20}.

4. Learning to infer object potentials

To scale up the inference algorithm in Sec 3.5, we adopt an active inference approach to adaptively select a subset of informative object hypotheses. As our test budget is unknown, we follow [29] and formulate the selection task as a sequential Markov Decision Process. Our goal is to learn an optimal policy to sequentially add object hypotheses, and at each step, the most informative object hypothesis with respect to the current model is selected. We will first introduce the formulation of our selection process and then discuss the policy learning.

4.1. Active inference with object subgraphs

Given the object-augmented dense CRF in Sec 3, we aim to select a subset of object-related potentials to improve efficiency in the inference. To this end, we decompose the model graph into a set of smaller subgraphs, which correspond to the object hypotheses or object relations.

Specifically, we make use of the model structure defined in Eq (1) and build the set of subgraphs based on the potential functions E_v, E_s and E_r. As the inference on basic dense CRF $E_v(\mathbf{L})$ is efficient, we always start from the subgraph of E_v and select a subset of subgraphs corresponding to the set of $\{E_s(\mathbf{L}, d_m)\}_{m\in\mathcal{S}} \cup \{E_r(\mathbf{L}, d_m, d_n, h_p)\}_{p\in\mathcal{P}}$. We note that our subgraphs are overlapped in the model graph and our de-selection refers to deactivating the corresponding object nodes instead of removing both the nodes and edges.

More formally, we introduce a subgraph selection state vector $\mathbf{z} = [\mathbf{z}^s, \mathbf{z}^r]$, where $\mathbf{z}^s$ for the singleton object set $\mathcal{S}$ and $\mathbf{z}^r$ for the object pair set $\mathcal{P}$. Each subgraph $k \in \mathcal{S} \cup \mathcal{P}$ is associated with an binary indicator z_k and $z_k = 1$ means k is selected. The full model for active inference can be written as

$$E(\mathbf{L},\mathbf{D},\mathbf{H},\mathbf{z}|\mathcal{T}) = E_v + \sum_{m\in\mathcal{S}} z_m^s E_s^m + \sum_{p\in\mathcal{P}} z_p^r E_r^p \quad (10)$$

where E_s^m is the m-th singleton object potential and E_r^p is the p-th object pair potential. Given a sparse $\mathbf{z}$, we can efficiently compute the node marginals of the model in Eq (10).

4.2. Subgraph selection as MDP

We formulate the subgraph selection as a Markov Decision Process. The state of the MDP s is $(\mathcal{T}, \mathbf{z})$ for an input video $\mathcal{T}$ and subgraph selection state $\mathbf{z}$. The initial state is $s_0 = (\mathcal{T}, \mathbf{0})$, which means we only select the basic dense CRF in E_v. The action space $\mathcal{A}(s) = \{i|z_i = 0\} \cup \{0\}$ means we can either choose a subgraph that is not selected before or terminate the process. If we are in state s_t and will take action r , the next state is represented as $s_{t+1} = (\mathcal{T}, \mathbf{z}^t + \mathbf{e}_r)$ where $\mathbf{e}_r$ is an indicator vector with the r-th column is 1 and all others are 0.

Then we define the expected reward of action r, or activating a subgraph r in state s_t as follows:

$$R(s_t, r, s_{t+1}) = \begin{cases} \eta(s_{t+1}) - \eta(s_t) & \text{if not terminated} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $\eta(s_t)$ is the expected per-class label accuracy of prediction given the state s_t. Our target is to learn a (deterministic) policy $\pi(s) \rightarrow r \in \mathcal{A}(s)$ that maximizes the expected reward.

Approximate policy learning with model features We parametrize the policy by defining a priority function $\gamma(f(s, r))$, where $f(s, r)$ is a set of object-based features and model uncertainty features. The policy $\pi(s)$ is defined as $\pi(s) = \arg\max_r \gamma(f(s, r))$. We learn the function $\gamma(\cdot)$ based on Q-learning [15, 29]. In Q-learning, we evaluate both linear regressor and random forest regressor as $\gamma(\cdot)$. We refer the reader to Sec 5.1 for the details of the feature design of $f(s, r)$.

Imitation learning with local classifier Another faster approximate scheme is the imitation learning [1], which learns a classifier based on local cues to predict the optimal action trajectory demonstrated by an expert. To this end, we first generate the optimal trajectories on training set based on dynamic programming. We then take the same set of features as in Q-learning and train a classifier to score the quality of a selection. More concretely, we build a training dataset in which the states in the optimal trajectories are treated as positive samples and the other states are negative. Then a random forest classifier is trained to predict the probability score of an action.

5. Experiments

We evaluate our object augmented dense CRF and the proposed active inference on three publicly available multi-class semantic video segmentation datasets. We focus on the CamVid [2] dataset here as it provides multiple foreground classes. To demonstrate the generalisability of our model, we also evaluate on the MPIScene [30] and DynamicScene [31] datasets. The details of these datasets are summarized in the supplementary material.

5.1. Implementation details

Object hypotheses We follow the exemplar-driven approach [25, 17], in which we manually annotate 20 exemplars for each of foreground object classes in CamVid. We apply the detectors every 10 frames and propagate detected objects to the whole sequence based on long-range trajectories [16]. We refer the reader to [17] for further details.

Policy learning features We design two sets of features: the detection related features and contextual features. The detection related features consist of object unary potentials, object position, object size, and detection score. The contextual features have two parts: supervoxel-level and object-level features. The supervoxel-level features include average supervoxel entropy for all classes, entropy on averaged supervoxel distribution, average entropy of foreground object class based on current supervoxel marginals and average supervoxel marginals. We compute these features in each stage of the incremental hypothesis selection based on the statistics of the model output at that stage. We also add the difference of those features between the current and previous stage to the supervoxel-level features. For object-level features, we compute the averaged entropy on object terms, the entropy on averaged objects based on their marginals, the average and maximum object marginals, as well as their stage-wise difference.

5.2. Baseline Methods

We compare to three baselines: simple entropy-based approach, Expected Labeling Change (ELC) and the Greedy graph induction (GreedyC) method. The entropy baseline computes the probabilistic marginals and selects the subgraph that has the highest average entropy in the marginals.

The ELC method activates the subgraph that generates the largest change in final prediction, or the pixel-level labeling in our case. Among all the baselines, the ELC is the most time-consuming one as it enumerates all the potential actions and performs inference for each of them.

The Greedy graph induction learns a classifier to imitate a policy which is generated by sequentially selecting most rewarding subgraph locally. This is a myopia version of imitation learning. We extract features as described in 5.1 for each sample and train a random forest classifier.

5.3. Experiment results

We conduct three sets of experiments on our selective inference CRF model: 1) Detailed comparison with baselines and the state-of-the-art methods on CamVid. 2) Scalability evaluation in terms of the number of object classes, object hypotheses and frames in video. 3) Generalization to other video segmentation datasets.

Active Inference on CamVid We first show the quantitative results of our method and compare with state-of-the-art methods in Table 1. We compute the accuracy and Intersection-Over-Union (IOU) score of semantic segmentation on CamVid[1].

We can see that our method achieves better performance than the dense CRF, which is a strong baseline. In addition, the overall pixel-average and class-average accuracy of our method are comparable to the state-of-the-art methods. Specifically, we outperform GeoF[11] significantly in overall IOU and Liu [17] in most (8 out of 11) of classes. We also compute the F1 score and our method achieves 59.8%, which is better than 59.1% in [17] and 58.2% in [24]. More importantly, we achieve the competitive performance using only one-third of object hypotheses. We visualize our subgraph selection procedure in Figure 4.

In addition, we demonstrate the efficiency of our proposed methods by showing how the accuracy improves with increasing number of subgraphs selected. We compare Local Imitation Learning (LocalC) and Q-learning with three baseline methods. We evaluate both linear and non-linear regressor for Q-learning. Figure 3 shows the prediction curve for average class accuracy, average foreground object accuracy and average pixel accuracy. We can see our proposed methods can always have earlier stop with superior performance. Overall, our method achieves a better trade-off in accuracy and efficiency than baseline methods. We will use the LocalC and Q-learning (linear) in the rest of our experiments.

Scalability Evaluation on CamVid We evaluate the scaling up property of our method in three natural scenarios. In the first setting, we gradually increase the number of object classes in the hypothesis pool. We start from the basic dense CRF and incrementally add hypotheses from Car, $Pedestrian$ and $Bicyclists$ detection. The results are shown in the left panel of Figure 5. We note that both Q-learning and LocalC improve their accuracies and when all three classes are added, the final performance is the same as the full model. In addition, despite the number of hypotheses increases significantly, the size of selected subsets grows slowly when more classes are added.

The second setting evaluates our method on video sequences with different average lengths. We test our model on the video lengths of 61, 121 and 241 frames. Results in the right panel of Figure 5 show that longer sequences benefit from better temporal smoothness effect and the number of selected hypotheses also increases slowly.

In the third setting, we generate more hypotheses of a single object class to improve the recall of object detection and evaluate our model with more hypotheses. We generate three different hypothesis sets by changing the threshold of detector. In our experiment, we choose -0.85,-0.9

[1]We do not include the methods that use strong object detectors and work solely on static images (e.g.,[14]).

Accuracy	Road	Building	Sky	Tree	Sidewalk	**Car**	Pole	Fence	**Pedestrian**	**Bicylist**	Sign	Pixel	Class
DenseCRF	89.3	74.4	96.0	**81.5**	**87.6**	68.8	10.7	**61.9**	40.5	6.3	**54.7**	82.5	61.1
Ours	88.9	73.8	96.0	81.2	87.4	81.3	10.5	60.3	43.8	9.3	**54.7**	82.8	62.4
Liu [17]	**92.4**	73.8	95.5	79.2	73.6	**81.7**	9.7	29	**60.9**	**42.1**	50.3	82.5	**62.5**
Tighe [25]	95.9	**87.0**	**96.9**	67.1	70.0	62.7	1.7	17.9	14.7	19.4	30.1	83.3	51.2
IOU score													
DenseCRF	85.1	67.2	**90.3**	**66.7**	63.4	56.1	8.4	17.6	26.9	5.8	21.3	-	46.1
Ours	**85.8**	66.8	90.1	**66.6**	**63.5**	**62.9**	**8.3**	**17.8**	**28.0**	8.5	21.4	-	**47.2**
Liu [17]	85.5	**67.3**	89.8	65.7	61.4	55.6	7.3	11.8	22.4	**14.9**	**22.2**	-	45.8
GeoF[11]	-	-	-	-	-	-	-	-	-	-	-	-	38.3

Table 1: Averaged semantic accuracy in CamVid. Our method can achieve the state of the art segmentation performance. Foreground object classes with hypotheses are in bold. We outperform the state-of-the-art in IOU significantly.

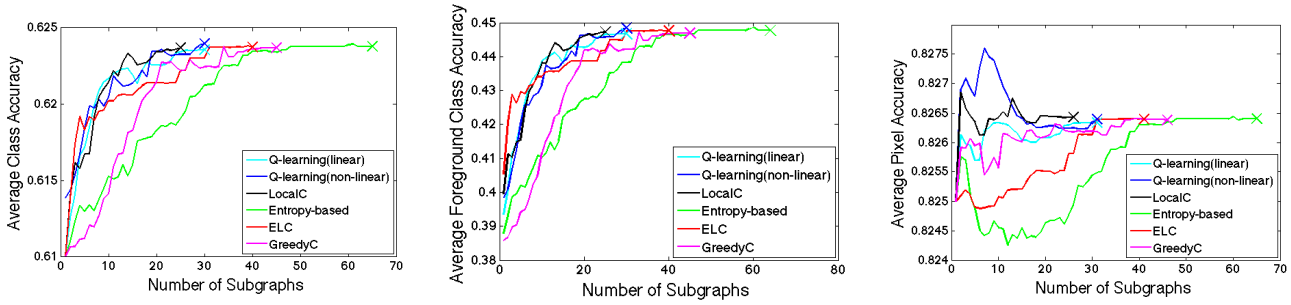

Figure 3: Traded-off performance on the CamVid dataset. The curve shows the increase in accuracy over the selective inference model as a function of subgraph number. The cross shows the termination point for inference. Best viewed in color.

and -0.95 as thresholds for $Half$, $Same$ and $Double$ settings respectively. We have tested the results in three object classes and all of them shows the significant improvement in terms of segmentation accuracy. Here we only show $Bicyclists$ since it performs the worse in three object classes. We can see from Figure 6 that lowering the threshold brings in more noise in detection results and makes the inference inefficient. Again, our method only infers a small subset of all the hypotheses and gains much improvement in single class segmentation performance. And the overall performance also increases a little in all three settings. More noticeably, the Q-learning with $Double$ setting in $Bicyclists$ can even achieve similar performance compared with the full model, which means we do not sacrifice other classes in this setting. More details about the overall performance in $Bicyclists$ and other classes can be viewed in supplementary materials.

In addition, we compute the inference time v.s. number of hypotheses, and compare with [17] and the baseline mean field in Table 2. We can see that our algorithm is much faster than the other two methods, and its inference time increases only sub-linearly with respect to the number of hypotheses.

Overall, all three sets of experiments show that our method can scale up well in terms of number of object classes, video length and number of object hypotheses from single class.

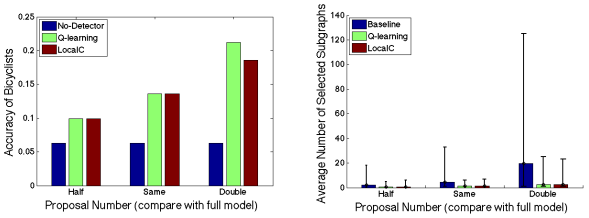

Figure 6: Scalibility with more hypotheses from a single class ($Bicyclists$). Left: Class accuracy; Right: Model complexity.

# of Subgraphs	Time(s)		
	[17] (GraphCut)	DCRF (MeanField)	Our Method
21.6	4.3	2.6	**1.5**
41.1	5.8	3.3	**1.6**
81.8	8.3	5.3	**1.8**
165	14.4	10.9	**2.4**

Table 2: Inference efficiency v.s. number of hypotheses of different methods on CamVid. See text for details.

Extension to Other Datasets We also apply our learned policy to two more video datasets, MPIScene and DynamicScene. Because of fewer foreground semantic classes, we can only generate the results from the Car or $Vehicle$ class. In these two datasets, we adjust our chunk length to the length of the full sequence and apply the parameters trained on CamVid directly.

Overall, we can achieve comparable or better perfor-

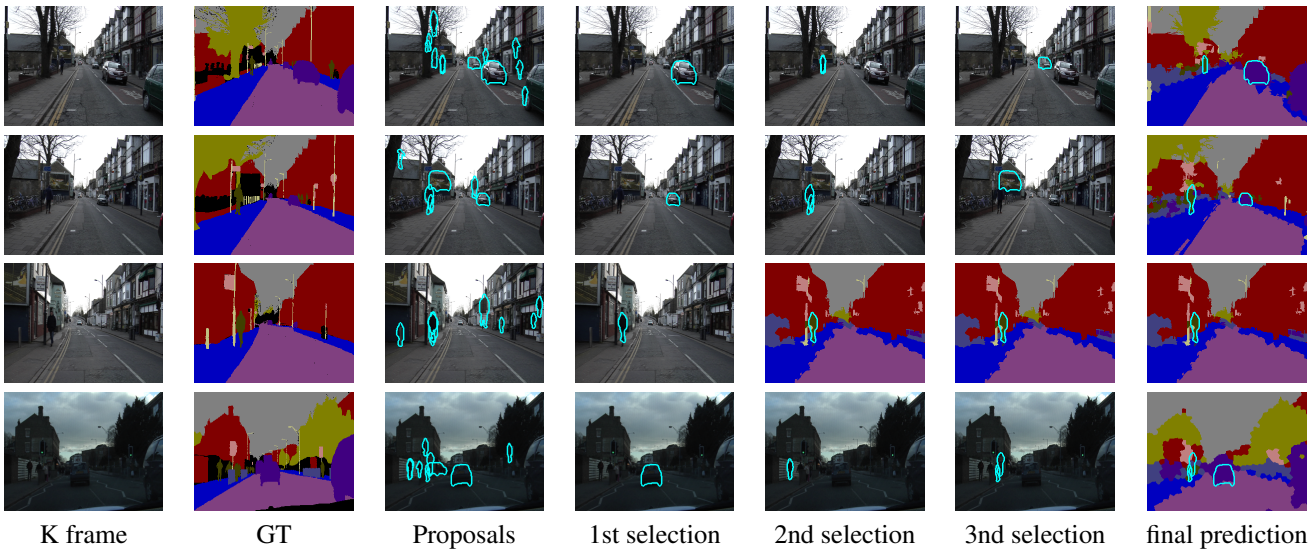

Figure 4: Examples of our selective inference on CamVid. We can always select the most informative subgraphs with high priority. The third row shows the early stop in subgraph selection that only choose one subgraph.

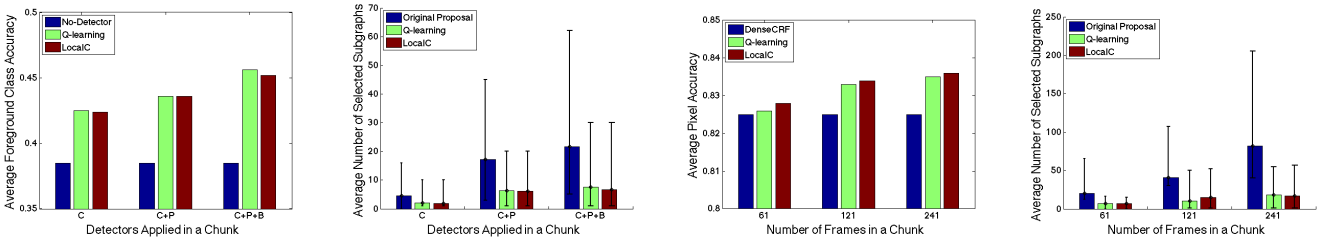

Figure 5: Left panel: Scalability with number of object classes (C: 'Car'; P:'Pedestrian'; B:'Bicyclist'); Right panel: Scalability with different length of videos. Both the average accuracy and number of selected subgraphs are shown. See text for details.

F1 Score	Bkgd	Road	Lane	**Vehicle**	Sky	Class
Ours	88.4	91.5	13.3	63.2	94.6	70.2
Liu [17]	**90.2**	**91.7**	11.2	**72.5**	**95.2**	**72.2**
Ondrej [19]	73	34	**33**	28	56	53.7

Table 3: Semantic segmentation performance (average per-class accuracy) in MPIScene dataset. See text for details.

Accuracy	Class	Pixel	# hypothesis
Ours	69.1	91	4
Liu [17]	**69.8**	**91.6**	11
Wojek [31]	68.4	91	-

Table 4: Semantic segmentation performance (average per-class and per-pixel accuracy) and the number of selected subgraphs in DynamicScene dataset.

mance with active inference, and so our method has higher efficiency and the learned policy can be easily generalized to new datasets. Because we have only one sequence in MPIScene dataset, we do not include the statistics of selected hypotheses in Table 3. We also show the overall performance in DynamicScene dataset compared with state-of-the-art methods. We can see that although our method does not outperform [17] but we require much fewer subgraphs to generate comparable results (See Table 4).

6. Conclusion

We have proposed an object-aware dense CRF model for multiclass semantic video segmentation, which jointly infers supervoxel labels, object activation and their occlusion relationship. We derive an efficient mean field inference algorithm for such joint model with moderate number of object hypotheses. To deal with a large number of object hypotheses in video, we propose to take an active inference approach, which chooses informative hypotheses to activate in the dense CRF. We investigate two approaches within the MDP framework, one of which is based on Q-learning and the other is a classifier trained by imitation learning. We have demonstrated that both approaches achieve the state-of-the-art performance on three video datasets with fewer object hypotheses used in the final model inference and are able to scale up for video parsing with more object classes and/or longer sequences.

Acknowledgments NICTA is funded by the Australian Government as represented by the Dept. of Communications and the ARC through the ICT Centre of Excellence program.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004. 2, 5

[2] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008. 2, 5

[3] J. Chang, D. Wei, and J. W. F. III. A Video Representation Using Temporal Superpixels. In *CVPR*, 2013. 3

[4] A. Y. C. Chen and J. J. Corso. Temporally consistent multi-class video-object segmentation with the video graph-shifts algorithm. In *WMVC*, 2011. 2

[5] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 1

[6] S. Gould and X. He. Scene understanding by labeling pixels. *Commun. ACM*, 57(11):68–77, 2014. 2

[7] A. Grubb, D. Munoz, J. A. Bagnell, and M. Hebert. Speed-machines: Anytime structured prediction. *arXiv preprint arXiv:1312.0579*, 2013. 2

[8] X. He, R. S. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *Computer Vision–ECCV 2006*, pages 338–351. Springer, 2006. 1

[9] P. Isola and C. Liu. Scene collaging: analysis and synthesis of natural images with semantic layers. In *(ICCV), 2013 IEEE International Conference on*, 2013. 1

[10] B. Kim, M. Sun, P. Kohli, and S. Savarese. Relating things and stuff by high-order potential modeling. In *in ECCV'12 Workshop on Higher-Order Models and Global Constraints in Computer Vision*, 2012. 1

[11] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. Geof: Geodesic forests for learning coupled predictors. In *CVPR*, 2013. 6, 7

[12] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 2, 3, 4

[13] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014. 2

[14] L. Ladick, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010. 1, 2, 6

[15] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003. 2, 5

[16] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011. 2, 6

[17] B. Liu, X. He, and S. Gould. Multi-class semantic video segmentation with exemplar-based object reasoning. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, 2015. 1, 2, 3, 6, 7, 8

[18] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[19] O. Miksik, D. Munoz, J. A. Bagnell, and M. Hebert. Efficient temporal consistency for streaming video scene analysis. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2013. 2, 8

[20] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool. Learning where to classify in multi-view semantic segmentation. In *Computer Vision–ECCV 2014*, pages 516–532. Springer, 2014. 2

[21] G. Roig, X. Boix, R. D. Nijs, S. Ramos, K. Kuhnlenz, and L. V. Gool. Active MAP Inference in CRFs for Efficient Semantic Segmentation. In *2013 IEEE International Conference on Computer Vision*, 2013. 1, 2

[22] M. Sun, B.-s. Kim, P. Kohli, and S. Savarese. Relating things and stuff via object property interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7), July 2014. 1, 2, 3

[23] C. A. Sutton and A. McCallum. Piecewise training for undirected models. *CoRR*, 2012. 4

[24] B. Taylor, A. Ayvaci, A. Ravichandran, and S. Soatto. Semantic video segmentation from occlusion relations within a convex optimization framework. In *EMMCVPR*, 2013. 2, 6

[25] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*, 2013. 1, 2, 6, 7

[26] J. Tighe and S. L. Marc Niethammer. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014. 1, 2

[27] C. Wang, M. de La Gorce, and N. Paragios. Segmentation, ordering and multi-object tracking using graphical models. In *ICCV*, 2009. 2

[28] D. Weiss, B. Sapp, and B. Taskar. Dynamic structured model selection. In *Computer Vision (ICCV)*, 2013. 1, 2

[29] D. Weiss and B. Taskar. Learning Adaptive Value of Information for Structured Prediction. In *Advances in Neural Information Processing (NIPS)*, 2013. 1, 2, 5

[30] C. Wojek, S. Roth, K. Schindler, and B. Schiele. Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes. In *ECCV*, 2010. 2, 5

[31] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, 2008. 2, 5, 8

[32] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 1, 2