Multi-class Semantic Video Segmentation with Exemplar-based Object Reasoning

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

Xuming He
NICTA/ANU
Xuming.He@anu.edu.au

Stephen Gould
ANU/NICTA
stephen.gould@anu.edu.au

Abstract

We tackle the problem of semantic segmentation of dynamic scene in video sequences. We propose to incorporate foreground object information into pixel labeling by jointly reasoning semantic labels of super-voxels, object instance tracks and geometric relations between objects. We take an exemplar approach to object modeling by using a small set of object annotations and exploring the temporal consistency of object motion. After generating a set of moving object hypotheses, we design a CRF framework that jointly models the supervoxel and object instances. The optimal semantic labeling is inferred by the MAP estimation of the model, which is solved by a single move-making based optimization procedure. We demonstrate the effectiveness of our method on three public datasets and show that our model can achieve superior or comparable results than the state-of-the-art with less object-level supervision.

1. Introduction

Semantic segmentation, which aims to jointly segment and detect object classes in images and videos, has become a core problem in scene understanding [8, 27], and has wide applications in automatic navigation, egocentric vision and surveillance. In order to achieve better semantic parsing of images, it is essential to explore object information [12, 28, 29] as well as multiple properties of the underlying scenes [6]. In particular, reasoning object instances and their relations with contexts has played an important role in the state-of-the-art methods for image segmentation [12, 24].

Semantic parsing of a single image requires relatively strong prior assumptions on the scene structure, and recent progress in object-augmented scene segmentation heavily relies on large datasets with object instance level annotation and/or pre-trained object detectors, which are expensive to obtain. For a complex scene with novel object classes, it remains a challenging task to reliably incorporate object instance knowledge into semantic image segmentation. To address these difficulties, we instead consider semantic parsing of image sequences in a video in this work.

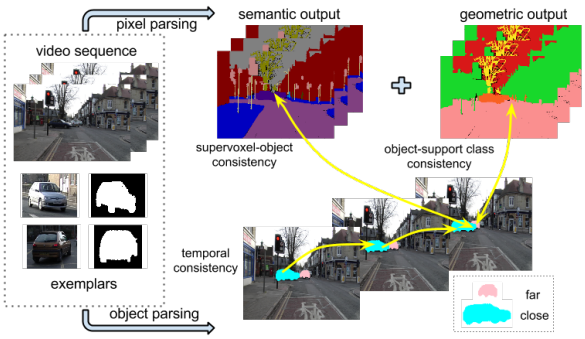

Figure 1: Overview of our approach. We jointly predict the semantic, geometric labels with respect to object detection, tracking and their relative ordering. The frontal-most object is in pink.

Video sequences of dynamic scene provide a natural setting to incorporate object level information in semantic segmentation. Motion cues can facilitate localizing object instances and inferring geometric relationships between objects. In addition, tracking of moving objects imposes long-range temporal consistency constraints to segmentation and allows weak supervision at object level. Most of previous video segmentation methods, however, either focus on capturing mid-level spatio-temporal consistency [3, 16], modeling static scenes [1], or using a single-class object detector trained with many additional object annotations [27].

In this work, we propose a joint framework for multi-class semantic video segmentation which integrates both region-level labeling and object-level reasoning. Given a video sequence taken from monocular camera, we formulate the segmentation as a supervoxel labeling problem. At region-level, we seek consistent semantic and geometric labeling of super-voxels that are smooth in spatio-temporal domain. At object-level, we infer foreground moving objects and their relative depth in a chunk of video frames, which imposes long-range spatial and temporal consistency of multi-class object segmentation. More importantly, we adopt a weak supervision strategy at object-level as in [7], in which only a small set of object exemplars is used for modeling each object class. An overview of our method is

shown in Figure 1.

Specifically, our approach consists of two stages. We first use object exemplars and dense point trajectories to generate a number of object segmentation hypotheses for each foreground class. Given these object hypotheses, we design a conditional random field (CRF) that jointly models semantic and geometric classes of supervoxels, as well as segmentation and relative depth of objects in videos. In particular, we propose a set of pairwise and higher-order potentials to impose the label consistency between objects and corresponding supervoxels, and to encode $occlusion$ and $support$ relations between object classes. To parse a video sequence, we compute the MAP estimation of the CRF model, which is formulated as minimizing a unified energy function and solved by an efficient move-making algorithm. We test our model on three public available datasets [1, 27, 26] and show that our method can achieve state-of-the-art or even better performance with much less training data.

The main contributions of our work can be summarized as follows. First, we incorporate multi-class object reasoning to semantic video segmentation, which enables us to capture long-range dependency in spatial-temporal domain. We show that inferring object instances and their relationships is beneficial to video segmentation. Second, we propose a weak supervision approach to model the foreground object classes by exploiting temporal coherency, pixel-object label consistency and a few annotated object exemplars. Finally, our method produces a better understanding of dynamic scenes, which includes not only pixel-wise semantic segmentation of videos but also object-level parsing with object instance segmentation, tracking and their relative depth ordering.

2. Related Work

In recent years, semantic parsing of images has been extensively investigated in computer vision and a large number of techniques have been proposed to address the problem of pixel labeling with semantic class information. Our work builds on recent progress in semantic scene parsing and object segmentation of static images, in which object information is incorporated into (super-)pixel labeling [12, 28, 24]. Particularly, we take the holistic perspective of scene understanding [29] which jointly parses the scene at pixel, object and scene level.

Early approaches in semantic segmentation use pre-trained object detectors and impose the consistency between pixel labeling and object detection output by higher-order potentials [12]. Their performance critically depends on the accuracy of object detectors. Yang et al [28] also introduce a global relative depth ordering to improve modeling of occlusion relation between overlapped objects. More recently, Tighe et al [24] jointly infer scene labeling, object segmentation and relative depth ordering. Their method decomposes the problem into three coupled subtasks and solves them in an alternating way. Similarly, Kim et al [10] integrate object detection and semantic segmentation. Those methods reply on large amount of training data with instance annotations to build object models. Our exemplar-based object reasoning is inspired by [7], which uses a small set of annotated exemplars to segment multiple object instances of a single class. However, our work differs from [7] in two important aspects. First, we address multi-class object segmentation and model their relative depth. Also, we add temporal cues to generate object trajectory hypotheses in video sequences.

While much progress has been made in semantic image segmentation, dynamic scene segmentation in video attracts less attention. Most existing approaches focus on modeling temporal consistency at pixel or region level [3, 16, 23], which do not have object-level reasoning, or building sparse or dense 3D models of statics scenes based on structure from motion [1], which cannot handle multiple moving objects. Top-down object information is first introduced by Wojek et al [27, 26], and they combine object tracking and the pixel labeling with a pairwise CRF. However, their method depends on object detectors pre-trained on many examples and does not jointly infer pixel labeling, object and object relations.

Some works aim to model the relative depth and occlusion of objects in video. Wang et al [25] consider foreground object segmentation, tracking and occlusion reasoning with a unified MRF model. Similar to our work, [13] uses long-term trajectories to discover moving objects. However, they do not model multiple semantic object classes, nor do they capture contextual relations between objects and background classes. Taylor et al [21] jointly infer pixel semantic classes and occlusion relationship in video segmentation. Unlike our method, they do not incorporate object instance level reasoning.

Geometrically and semantically consistent labeling is first introduced in [6], and has been extended to video scene parsing in [23, 14]. Stixel world model [18] also exploits geometric cues in semantic segmentation. However, none of them jointly estimate object instance segmentation.

3. Our Approach

We address the multi-class semantic video segmentation problem from a holistic perspective, in which we jointly assign a category label to every pixel, and infer object instance segmentation and their geometric relations in a video. Our main focus is to explore temporal consistency of object motion and thus to integrate object-level information more effectively.

To this end, we propose an exemplar-based approach to incorporating object instance segmentation and object rela-

tions to semantic video segmentation. We first use a small number of objects and their masks as well as dense trajectories to generate a redundant set of dynamic object hypotheses. Given these object hypotheses, we design a spatio-temporal CRF to jointly label all supervoxels and infer activations and relations of object hypotheses.

3.1. Dynamic Object Hypothesis Generation

To handle multiple object instances, we first generate a set of object trajectory hypotheses from a video sequence, which are binary masks in spatio-temporal domain. This hypothesize-and-verify approach greatly reduces the search space of object instances.

Our hypothesis generation method consists of three steps. **The first step** detects object instances and generates their masks in a sparse set of key frames, which aims to improve the efficiency. We apply an exemplar SVM [15] detectors trained with a small number of examples (10–20) to obtain object proposals in the key frames as in [7].

Given these static proposals, we then propagate them to the entire sequence in **the second step**. We perform both forward and backward propagation based on pixel-level trajectories, which are obtained by employing the method in [19]. We compute an affine transformation of each object mask using the pixel trajectories passing through the mask. This enables us to partially accommodate the variation of shapes due to motion. Compare with the per-frame detection method [27], we can generate static proposals more effectively and efficiently by exploring the temporal information. Some examples of generated object hypotheses are shown in Figure 2.

Finally, we build longer-range object *trajectory* hypotheses in **the third step**, which extends the trajectories generated from the propagation in the second step. To this end, we first construct a directed graph of the static object proposals. Each node of the graph is a static proposal and we add an edge between two nodes if they are: (1) from consecutive frames (2) of the same category (3) of similar mask size and heavily overlapped after propagation. We refer the reader to Section 4.2 for details. We define edge direction as the direction of time evolution. Given the directed graph, we use depth-first search to generate possible paths starting from all the earliest static object proposals which correspond to the nodes without parents in the graph. In the end, we collect all the generated paths as the object trajectory hypotheses.

3.2. Spatio-temporal CRF with Object Reasoning

We represent a video sequence as a set of supervoxels, which are computed based on [2]. For long video sequences, we take a sliding window approach and consider a video chunk with length T each time. We then augment the supervoxel representation with a set of object trajectory hypotheses, and introduce reasoning of the activation of objects and their occlusion and support relations.

Figure 2: Examples of proposals from detectors (top) and from propagation (bottom) in CamVid. Occluded objects are failed to be detected by detectors but can be successfully proposed from propagation.

Formally, given a video chunk $\mathcal{T}$, let $i \in \{1, ..., N\}$ index the supervoxels in $\mathcal{T}$. We denote the label of superpixel i as $l_i = \{l_i^g, l_i^s\}$, which is a random variable in the joint space of semantic class $\mathcal{L}^s$ and geometric classes $\mathcal{L}^g$. The variable $\mathbf{L} = (l_1, ..., l_N)$ is the label configuration of the whole video chunk.

For the set of object trajectory hypotheses, we want to infer the true active objects jointly with the supervoxel labeling and remove the false ones. To that end, we introduce a binary variable d_m indicating whether the m-th hypothesis is activated or not, and let $m \in \{1, ..., M\}$ indexing from hypothesis pool $\mathcal{O}$ as described in Section 3.1. For hypothesis m, we denote its trajectory as $\mathbf{m} = \{m_1, ..., m_{t_m}\}$ and m_t is the static object proposal in t-th frame. Note that once the m-th hypothesis is activated, all static proposals on its trajectory are activated. The variable $\mathbf{D} = (d_1, ..., d_M)$ is the configuration of all the object hypotheses. Also, we denote the object class of the m-th hypothesis as o_m and the set of supervoxels it occupies as S_m.

In addition, we capture the relative depth ordering of object instances by introducing an occlusion variable $h_{mn} \in \{-1, 0, 1\}$ for each pair of overlapped proposals $\{m, n\}$. The value -1 and 1 denotes m-th proposal is occluded by or occludes n-th proposal respectively and 0 denotes there exists no occlusion relation between them, which means at least one of the proposals is inactive. We denote the set of all pairs of overlapped proposals as $\mathcal{P}$, and represent the configuration of all occlusion variables by $\mathbf{H} = \{h_{mn}\}_{(m,n)\in\mathcal{P}}$. Details of the overlapping proposal pairs will be explained in 4.2.

We formulate the semantic video parsing problem as a joint labeling of supervoxels, object hypotheses and object relations, and build a joint Conditional Random Field (CRF) on the label variables $\mathbf{L}$, $\mathbf{D}$, and $\mathbf{H}$. An overview of our graphical model is shown in Figure 3. The overall energy function of our CRF model consists of three main components (we omit the input $\mathcal{T}$ for clarity):

$$E(\mathbf{L}, \mathbf{D}, \mathbf{H}) = E_s(\mathbf{L}) + E_o(\mathbf{D}, \mathbf{H}) + E_c(\mathbf{L}, \mathbf{D}, \mathbf{H}), \quad (1)$$

where $E_s(\mathbf{L})$ represents supervoxel-level potentials,

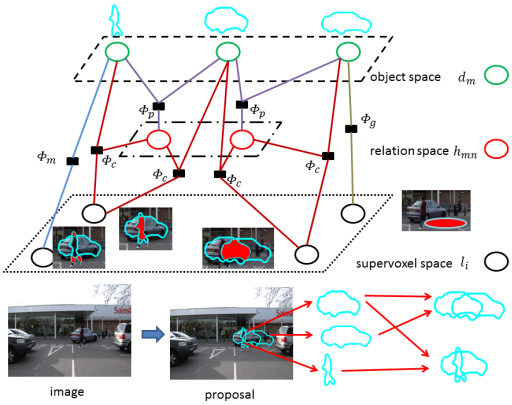

Figure 3: Graphical representation of our CRF. Note that all unary terms and supervoxel pairwise terms are not shown for clarity. Two proposals are occluded pairs and one is exclusive pair (two $Cars$). Supervoxels are shown in red.

$E_o(\mathbf{D},\mathbf{H})$ is the object-level potentials, and $E_c(\mathbf{L},\mathbf{D},\mathbf{H})$ are the potentials imposing the consistency between the object and supervoxel labeling. We will describe the details of these three terms in the following subsections.

3.2.1 Supervoxel-level potentials

The supervoxel-level potentials $E_s(\mathbf{L})$ include a data term potential for every supervoxel and a pairwise potential that encodes spatio-temporal smoothness of the supervoxel labeling:

$$E_s(\mathbf{L}) = \sum_i \phi_s(l_i) + \sum_{(i,j)\in\mathcal{N}_l} \Phi_s(l_i,l_j) \tag{2}$$

where $\mathcal{N}_l$ denotes the set of spatio-temporal adjacent supervoxels.

The **supervoxel unary term** $\phi_s(l_i)$ is the cost of assigning l_i to supervoxel i. We define $\phi_s(l_i) = -\log P_l(l_i)$, where $P_l(l_i)$ is the output of a unary classifier (See Sec 4.2 for details.)

We use an edge sensitive Potts model as the **supervoxel pairwise term** $\Phi_s(l_i,l_j)$, which to enforce adjacent supervoxels to take the same label unless there is an intensity edge in between. Specifically, $\Phi_s(l_i,l_j) = \alpha_l \exp(-\frac{\|\mathbf{f}(i)-\mathbf{f}(j)\|^2}{\beta})[\![l_i \neq l_j]\!]$, where $\mathbf{f}(i)$ is the averaged color feature of supervoxel i in CIELab space, and $[\![\cdot]\!]$ is the indicator function.

3.2.2 Object-level potentials

The object-level potentials $E_o(\mathbf{D},\mathbf{H})$ describe the data term for object proposals, occlusion variables and the relationship between object $\mathbf{D}$ and the occlusion $\mathbf{H}$:

$$E_o(\mathbf{D},\mathbf{H}) = \sum_m \phi_o(d_m) + \sum_{(m,n)\in\mathcal{P}} \Big(\phi_h(h_{mn}) + \Phi_p(h_{mn},d_m,d_n)\Big) \tag{3}$$

where $\phi_o(d_m)$ and $\phi_h(h_{mn})$ is the object and occlusion unary respectively, and $\Phi_p(h_{mn},d_m,d_n)$ is the relation term.

The **object unary term** $\phi_o(d_m)$ models the cost of activating m-th proposal and has the following form,

$$\phi_o(d_m) = \left(\alpha_c - \alpha_o \log \tfrac{P_m}{1-P_m}\right) d_m \tag{4}$$

where P_m is the probability of activating m-th proposal, which is obtained from a trained classifier's output (See Sec 4.2 for details). α_o and α_c are two weight parameters and α_c is introduced to encourage sparse detections.

The **occlusion unary term** $\phi_h(h_{mn})$ models the cost of h_{mn} taking one of the three states and is defined as $\phi_h(h_{mn}) = -\alpha_h \log P_h(h_{mn})$. Similarly, P_h is the probabilistic score from a trained classifier (See Sec 4.2).

The **occlusion relation term** Φ_p encodes that the occlusion variable h_{mn} should be consistent with the states of d_m and d_n, i.e., $h_{mn} \neq 0$ iff d_m and d_n are active, and an exclusive relation:

$$\Phi_p(h_{mn},d_m,d_n) = \alpha_{\texttt{inf}}\Big([\![d_m d_n = 0 \wedge h_{mn} \neq 0]\!] + [\![d_m d_n = 1 \wedge h_{mn} = 0]\!] + [\![d_m d_n = 1 \wedge \{m,n\} \in \mathcal{E}_o]\!]\Big) \tag{5}$$

where the exclusive set $\mathcal{E}_o$ consists of pairs that share the same category and are significantly overlapped (See Sec 4.2), and $\alpha_{\texttt{inf}}$ is a large cost.

3.2.3 Supervoxel-object label consistency potentials

The supervoxel-object label consistency potentials E_c enforce the consistency of an object activation and the labels of supervoxels related to the object. It consists of three terms, encoding overlap, support and occlusion consistency respectively,

$$E_c(\mathbf{L},\mathbf{D},\mathbf{H}) = \sum_{m\in\mathcal{S}} \Big(\sum_{i\in S_m} \Phi_v(d_m,l_i) + \sum_{i\in B_m} \Phi_g(d_m,l_i)\Big) + \sum_{(m,n)\in\mathcal{P}} \sum_{k\in\{m,n\}} \sum_{i\in S_k} \Phi_c(h_{mn},l_i,d_k) \tag{6}$$

where $\mathcal{S}$ denotes the set of isolated object proposals, and B_m is the neighboring supervoxels located below the m-th object hypothesis.

The **overlap consistency term** $\Phi_v(d_m,l_i)$ penalizes the inconsistency between the class of an active object and the semantic label of the supervoxels it contains. The more inconsistency exists between local supervoxel prediction and the object class, the higher cost will be assigned if the object is active. We define the cost as:

$$\Phi_v(d_m,l_i) = \alpha_v \frac{vol(i)}{vol(S_m)} d_m [\![l_i \neq o_m]\!] \tag{7}$$

where S_m denotes the set of supervoxels contained by object proposal m and o_m is the object class. $vol(\cdot)$ computes the volume of a region, and α_v is the weight coefficient.

The **support consistency term** $\Phi_g(d_m, l_i)$, is introduced to encode the supporting relation between the foreground objects and its supporting surface. We enforce that active objects should be supported from below by supervoxels with geometric label $Horizontal$. That is,

$$\Phi_g(d_m, l_i) = \alpha_g \frac{vol(i)}{vol(B_m)} d_m [\![l_i^g \neq Horizontal]\!] \quad (8)$$

where B_m is define in Eq (6) and α_g is the weight coefficient.

The **occlusion consistency term** penalizes the inconsistency between the class of object proposals and the labels of the supervoxel they contain with respect to the occlusion relations. Specifically, we enforce the supervoxels in the overlapped regions should be explained by the frontal-most object if both proposals are activated. Otherwise, their labeling should be consistent with the active one:

$$\Phi_c(h_{mn}, l_i, d_m) = \alpha_p \tfrac{vol(i)}{vol(S_m)} [\![l_i \neq o_m]\!] \Big([\![h_{mn} = 1]\!] + [\![h_{mn} = 0 \wedge d_m = 1]\!] + [\![i \in \{S_m \setminus S_n\} \wedge h_{mn} = -1]\!] \Big) \quad (9)$$

where $\{S_m \setminus S_n\}$ donates the set of supervoxels occupied by m-th object proposal but not by n-th. α_p is the weight coefficient.

3.3. Model Inference and Learning

Due to the complexity of our model, we adopt the piecewise learning [20] approach to incrementally estimate parameters. We firstly learn the parameters of pairwise CRF. It consists only ϕ_s and Φ_s. Then we fix learnt parameter α_l and learn the coefficients α_c, α_o, α_v, α_g for the object potentials in the CRF, including ϕ_o, Φ_v and Φ_g. We refer to this partial model as the incremental CRF. Finally, we learn the rest parameters of our full model (α_h, α_p). Note that all parameters are learnt on validation set by grid search and $\alpha_{\texttt{inf}}$ is set to be a large number (10^{20}). We automatically choose the set of parameters that maximize the per-class accuracy during the piecewise learning.

Given the parameters and test sequences, we compute the maximum a posterior (MAP) estimate by minimizing the energy $E(\mathbf{L}, \mathbf{D}, \mathbf{H})$ according to the method in [17]. Specifically, we apply the improved version of QPBO (QPBOI) introduced in [17] and prefer the unassigned nodes keep their original labels during the expansion move.

4. Experiments

4.1. Datasets

We test the efficacy of the proposed framework on three multi-class semantic video segmentation datasets. We focus on the CamVid dataset here as they provide multiple foreground classes. To demonstrate the generalisability of our model, we also evaluate on the MPIScene and DynamicScene datasets.

Color
C1: mean and variance in CIE-Lab color space
Texture
T1: mean and covariance of 17-dimensional filterbank response
HoG
H1: mean and variance of HOG feature on 8x8 patches
Optical Flow
O1: weighted dense optical flow histogram and mean O2: flow differential : Histogram of differential of dense optical flow in x and y, across 3 kernel size of differential(3, 5 and 7)
Semantic and Geometric Label Feature
S1: average, max and variance of semantic probability for each class G1: average, max and variance of geometric probability for each class

Table 1: Statistics computed to represent supervoxels.

CamVid [1] consists of 5 video sequences captured during the daytime and dusk. These sequences are sparsely labelled at 1Hz with 11 semantic classes. We follow the data split in [1] and annotate 20 exemplars for Car, $Bicyclists$ and $Pedestrian$ respectively. To obtain the ground truth geometric label, we apply a simple mapping from 11 semantic classes to 5 geometric classes.

MPIScene [26] consists of 156 annotated frames with 5 semantic classes. We follow the set-up in [16] and annotate 10 exemplars for $Vehicle$.

DynamicScene [27] consists of 176 sequences with 11 successive images each, and the last frame of each sequence is labelled with 8 classes. Half of the sequences are used for training and the remaining ones are for testing. Among all the training data they provide, we use only 46 images for unary term learning and 10 samples for detector. And we test our model on the same test set as [27].

4.2. Implementation Details

The **supervoxel** generation and supervoxel unary term are different in three datasets. In **CamVid** dataset, we use a sliding window approach, in which each window consists of 61 frames and shares one image with previous one to maintain temporal label consistency. Then we employ the method in [2] to generate spatio-temporal supervoxels. We extract a set of image and motion features at each pixel, such as color, texture and HoG features. We also apply methods proposed in [5] and [9] in each frame to obtain per-pixel semantic and geometric probability independently. Then we train the random forest classifier [4] for P_l on these features (See Table 1 for a summary).

In **MPIScene** and **DynamicScene** datasets, we set $\mathcal{T}$ as 156 and 11 respectively with no overlapping frames with other chunks. We follow the same supervoxel generation procedure as in the CamVid and set P_l as the averaged semantic probability for each class.

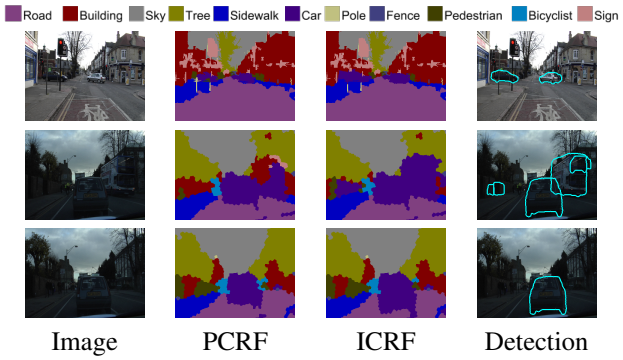

Figure 4: Semantic segmentation results from PCRF and ICRF and the activated detections in ICRF in the CamVid. Active detections impose strong labeling consistency in our model.

The proposal and relation parts are the same for three datasets. We obtain detectors by the exemplar SVM [15]. As no **proposal** ground truth is available for probability regression on object instances, we automatically generate "ground-truth" from segmentation ground truth by defining a proposal as"true positive" if more than half of its occupied pixels are consistent with its category. Then we extract $S1, C1$ as in Table 1 and the Chamfer distance between object mask and edge as features and train logistic regressors to obtain P_o.

We define two **relations** for proposed chains. In particular, proposal chains **M** and **N** are exclusive if : (1) they belong to the same category. (2) some values of $f_{overlap}(m_t, n_t)$ are larger than λ_1, where λ_1 is empirically set as 0.75. Besides, we define **M** and **N** are overlapped if some values of $f_{overlap}(m_t, n_t)$ are larger than λ_2. λ_2 is selected as 0.15 in experiment. We manually label the relations of 20 pairs as ground truth and fit a multiclass logistic regressor for P_h. The features for regressor are $C1$, the Chamfer distance between mask and edge, and number of terminated trajectories for each proposal chain.

4.3. Results on CamVid Dataset

4.3.1 Segmentation Results

There are three settings of our model, the baseline—pairwise CRF (PCRF), incremental CRF (ICRF) and our full model. PCRF consists only supervoxel unary and pairwise terms while ICRF combines multiple detections without reasoning object occlusion relation. Quantitative semantic segmentation results in Table 2 show that by exploring occlusion relation, we can achieve (1) better performance in terms of overall measurement and three foreground classes with respect to PRCF and ICRF. (2) comparable results in overall measurement and better performance in interested classes compare with state-of-the-art under much less supervision.

Figure 4 compares segmentation results between PCRF and ICRF. The second and last row show that the strong labeling consistency is enforced by activated Car detections. As can be seen from Table 2, performance in three interested classes is improved but the overall performance is roughly the same. There are two possible reasons. Firstly, supervision for detection is quite weak and we may introduce too much noise during regression. Secondly, spatially overlapped detections introduce conflicts in ICRF and thus deteriorate the performance.

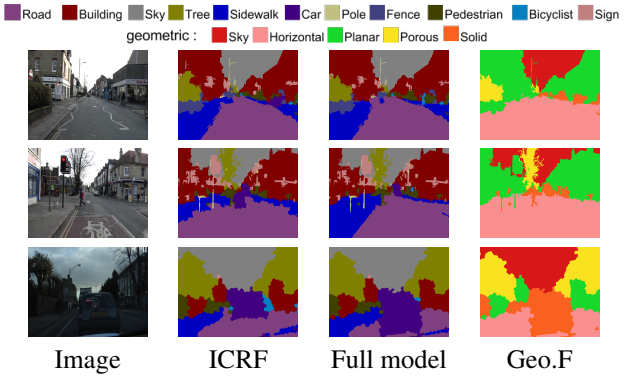

Figure 5: Semantic segmentation results of ICRF and full model in CamVid. Geo.F represents the geometric labeling results of the full model.

IOU	Sky	Horizon	Planar	Porous	Solid	Class
Stage [14]	88.3	88.7	69.2	52.7	40.3	67.8
Full model	**90.0**	**93.3**	**76.1**	**65.7**	**48.7**	**74.8**

Table 3: CamVid dataset geometric labeling results.

We then incorporate object relations to address the existing problems in ICRF. We explicitly model the labeling consistency with respect to occlusions in full model. The quantitative results in Table 2 show that incorporating relations do boost the performance. Compared with state-of-the-art, our model can achieve better or comparable results with much weaker supervision. Figure 5 shows the quantitative semantic results of both ICRF and full model as well as the geometric output of full model. The fourth and last row show that object relation reasoning inactivates the false positive and activates the true positive respectively. The significant improvement in both three object classes and overall measurement can be seen in Table 2.

We also compare our geometric predictions with the state-of-the-art geometric labeling method [14] in Table 3. We can see that our method outperforms the state-of-the-art significantly on the CamVid dataset.

4.3.2 Object Segmentation and Proposal Efficiency

Table 4 shows the detector performance under weak supervision. Although detectors fail to detect more than half of the objects and propose a large number of false positives,

Accuracy	Road	Building	Sky	Tree	Sidewalk	Car	Column-Pole	Fence	Pedestrian	Bicylist	Sign-symbol	Pixel	Class
Static	95.6	84	94.9	76.7	53.1	66.9	4.9	9.2	28.7	17.6	5.9	78.2	48.9
PCRF	90.7	74.8	95.6	**80.3**	70.3	76.4	10.2	30.6	60.3	36.9	51.4	81.3	61.6
ICRF	90.4	73.9	95.6	80.0	69.9	79.3	10	29.9	60.6	38.3	51.2	81	61.7
Full model	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 [23]	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
Tighe [22]	**96**	83.1	94.6	73.5	71.2	78.1	5.3	32.8	58.6	**45.9**	**71.2**	**83.9**	**62.5**
Ladicky [12]	93.9	81.5	96.2	76.6	**81.5**	78.7	**14.3**	**47.6**	43	33.9	40.2	83.8	**62.5**
IOU score													
ICRF	82.1	66.7	90.0	66	53.7	54.1	7.3	12.3	22.2	15.0	21.4	-	44.6
Full model	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
F1 score													
ICRF	90.3	80.1	**94.7**	**79.5**	69.9	70.2	13.7	21.9	-	**26.1**	35.2	-	58.1
Full model	**92.2**	**80.4**	94.6	**79.3**	**76.1**	**71.5**	13.6	21.0	-	25.9	36.3	-	**59.1**
Occlusion[21]	87.0	71.6	92.9	57.0	**76.1**	60.9	**19.8**	**37.5**	-	24.3	**54.4**	-	58.2

Table 2: Averaged semantic recall, intersection-over-union (IOU) and F1 score of existing methods in CamVid. Note that supervision in our model is much weaker than the baselines. F1 score in $Pedestrian$ is not provided to have a fair comparison with [21].

Category	Precision	Recall	IOU
Car	64.7/74.8	30/30	26.3/27.2
Pedes.	2.9/32.3	3.2/10	2/8.2
Bicy.	0/6.8	0/4.7	0/2.9

Table 4: Precision, recall and IOU score of detection performance at equal FPPI in CamVid. In each cell, the first one is the object segmentation result from detectors only method and the second is that of full model.

our model can still learn from weak detections and benefit from object relation reasoning. The detector performance is improved significantly in terms of precision, recall and IOU criterion.

In addition to the improvement in object segmentation, our model also boosts the efficiency compare with per-frame detection method. In terms of time consuming for proposal generation in each chunk, per-frame method takes 858s while that of our method is less than 248s. Overall, we improve the detection efficiency by 3.5 times.

4.3.3 Object Reasoning

As can be seen in the top three rows in Figure 6, the full model inactivates false positives that fail to be identified in ICRF by reasoning object relations. The last two rows of Figure 6 show the relative ordering of overlapped proposals. We can see that the full model provides a better representation of scenes by inferring the invisible parts of occluded objects.

4.4. Extension to Other Datasets

To test the generalization of our model, we also evaluate our framework on MPIScene and DynamicScene datasets. Except the supervoxel unary and exemplars, we do not re-

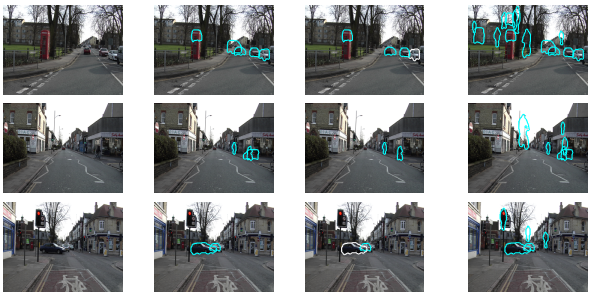

Figure 6: Examples of all detector proposals, active detections from ICRF and full model in CamVid. Full model can successfully infer the relative depth of overlapped objects and suppress the false positives. The frontal-most proposal is colored in red.

F1 Score	Background	Road	Lane	Vehicle	Sky	Class
PCRF	89.8	**91.7**	11.2	66.8	**95.2**	71
Full model	**90.2**	**91.7**	11.2	**72.5**	**95.2**	**72.2**
Ondrej [16]	73	34	**33**	28	56	53.7
IOU score						
PCRF	81.6	**84.7**	**6**	50.1	**90.8**	62.6
Full model	**82.2**	84.6	**6**	**56.9**	**90.8**	**64.1**
Recall						
PCRF	**83.6**	**98.7**	6	74.7	99.3	72.4
Full model	83.2	98.1	6	**90.9**	99.3	**75.5**

Table 5: Semantic segmentation performance in MPIScene. Our results outperform the state-of-the-art significantly.

train the CRF parameters specifically for these two datasets but apply those we learned in CamVid directly.

Table 5 shows the semantic segmentation results on MPIScene dataset. Our model outperforms the state-of-the-art

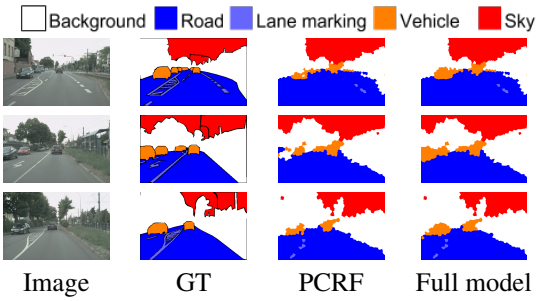

Figure 7: Comparison of the semantic segmentation results of PCRF and full model on MPIScene dataset.

method [16] significantly. Qualitative results can be viewed in Figure 7.

Our model also outperforms the state-of-the-art method [27] on DynamicScene dataset. The per-class and per-pixel accuracy on DynamicScene are 91.6% and 69.8% respectively from our model. According to [27], their results are 88.3% and 68.4% respectively. It is worth mentioning that our results are obtained under much weaker supervision and tested on the same test set as [27]. Particularly, we utilize one third of their labeling training data and less than 1% of object annotations during training.

5. Conclusion

In this paper, we propose a multiclass semantic video segmentation method that joint infers semantic, geometric labeling and relative ordering of proposed objects in dynamic scenes. We build a unified CRF model with a variety of potential functions that encode object relations with local semantic and geometric labeling. We also solve the MAP inference problem efficiently in an one-shot optimization procedure. Moreover, we exploit the temporal information in videos and propose an efficient way to generate object proposals with less training samples. We show that we can achieve comparable or better performance than state-of-the-art methods in three popular multi-class segmentation datasets.

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

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

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

[4] P. Dollár. Piotr's Image and Video Matlab Toolbox (PMT). `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`.

[5] S. Gould. DARWIN: A framework for machine learning and computer vision research and development. *Journal of Machine Learning Research*, 2012.

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

[7] X. He and S. Gould. An exemplar-based crf for multi-instance object segmentation. In *CVPR*, Columbus, USA, 2014.

[8] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[9] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[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.

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

[12] 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.

[13] 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.

[14] B. Liu, X. He, and S. Gould. Joint semantic and geometric segmentation of videos with a stage model. In *WACV*, 2014.

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

[16] 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.

[17] C. Rother, V. Kolmogorov, V. S. Lempitsky, and M. Szummer. Optimizing binary mrfs via extended roof duality. In *CVPR*, 2007.

[18] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. Stixmantics: A medium-level model for real-time semantic scene understanding. In *ECCV*, 2014.

[19] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

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

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

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

[23] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*, 2013.

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

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

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

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

[28] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object models for image segmentation. *IEEE TPAMI*, 2011.

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