Multiple Structured-Instance Learning for Semantic Segmentation with Uncertain Training Data

Feng-Ju Chang[1,2] Yen-Yu Lin[1] Kuang-Jui Hsu[1]
[1]Academia Sinica, Taiwan [2]University of Southern California, USA

Abstract

We present an approach MSIL-CRF that incorporates multiple instance learning (MIL) into conditional random fields (CRFs). It can generalize CRFs to work on training data with uncertain labels by the principle of MIL. In this work, it is applied to saving manual efforts on annotating training data for semantic segmentation. Specifically, we consider the setting in which the training dataset for semantic segmentation is a mixture of a few object segments and an abundant set of objects' bounding boxes. Our goal is to infer the unknown object segments enclosed by the bounding boxes so that they can serve as training data for semantic segmentation. To this end, we generate multiple segment hypotheses for each bounding box with the assumption that at least one hypothesis is close to the ground truth. By treating a bounding box as a bag with its segment hypotheses as structured instances, MSIL-CRF selects the most likely segment hypotheses by leveraging the knowledge derived from both the labeled and uncertain training data. The experimental results on the Pascal VOC segmentation task demonstrate that MSIL-CRF can provide effective alternatives to manually labeled segments for semantic segmentation.

1. Introduction

Semantic segmentation [1, 3, 12, 15, 18, 20, 24, 27, 29] aims to assign one of predefined object classes or background to each pixel in an image. Distinct from the conventional image segmentation task, *e.g.*, [2, 7, 22, 23], semantic segmentation not only determines the shapes of objects of interest but also identifies their categories. As a key component of image analysis, semantic segmentation is essential to widespread applications, such as scene understanding, object recognition, and image/video editing.

Recent research efforts have advanced semantic segmentation in many aspects, such as more powerful features [24], combination of information derived from different levels of image quantization [12, 15], and exploration of contextual relations among object classes [1, 15]. These approaches are often built on graphical models such as *conditional random fields* (CRFs) [25] for their merits in fusing diverse evidences and ensuring spatial consistency. However, learning graphical models for complex semantic segmentation tasks, *e.g.*, *Pascal VOC* [11], usually requires sufficient training data in form of object segments, *i.e.*, pixel-wise annotation. The heavy annotation cost hinders the advances in semantic segmentation.

In this work, we aim at reducing annotation cost in semantic segmentation, and consider that a few *object segments* and a set of *objects' bounding boxes* are given. We focus on inferring unknown object segments in the bounding boxes, and use the inferred segments as the training data for semantic segmentation. Since labeling a bounding box of an object takes only four clicks, replacing object segments with bounding boxes reveals the potential of reducing the annotation cost. Figure 1 gives an overview of the proposed framework.

Motivated by the capacity of multiple image segmentations [4, 10] for inferring object segments, we first generate multiple object segment hypotheses for each bounding box. By assuming that at least one segment hypothesis is close to the ground truth, the inference of the object segment in a bounding box can be achieved by *picking* the best hypothesis. It can be observed that the bounding boxes and their segment hypotheses here match the two-layer structure, *bags and instances*, in *multiple instance learning* (MIL) [9]. Namely, each bounding box can be regarded as a *positive bag* with its segment hypotheses as *instances*, and the hypothesis closest to the ground truth corresponds to the *positive instance*. Beyond MIL, each segment hypothesis is composed of superpixels. Exploring the structural information among superpixels generally facilitates ranking the hypotheses. Therefore, we cast the inference of the object segments in the bounding boxes as an instance of the *multiple structured-instance learning* (MSIL) problem.

We develop an algorithm, called *MSIL-CRF*, which solves the task of MSIL upon CRFs. It uses the principle of MIL to deal with our uncertainty of the segment ground truth, and leverages the formulation of CRFs to model the structural information. Moreover, it can jointly consider

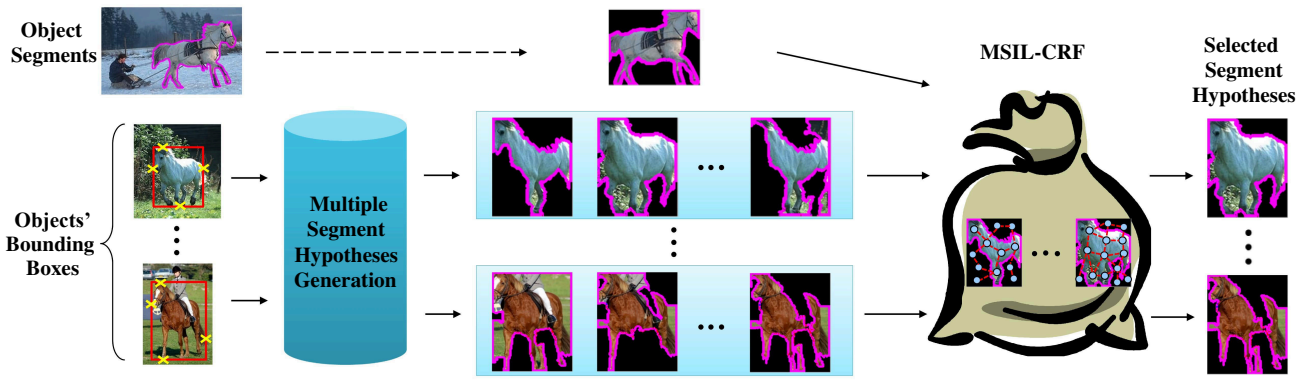

Figure 1. Given a few object segments and a set of bounding boxes of objects, our approach is designed to infer the unknown object segments in the bounding boxes by taking both the labeled data (object segments) and uncertain data (bounding boxes) into account jointly.

both the labeled data (manually labeled object segments) and the uncertain data (objects' bounding boxes), and establish more accurate models for segment inference. After completing the optimization, the most plausible object segment of each bounding box is determined. Any off-the-shelf approach to semantic segmentation can then be adopted and trained with the inferred object segments.

The main contribution of this work is the development of MSIL-CRF. On the one hand, it generalizes CRFs to work on training data with uncertain labels. On the other hand, it provides a way of dealing with structured instances in MIL. Technically, it adopts the *smooth maximum function* [26] to express our belief over the instances in each bag. The resulting objective function is differentiable with respect to the variables to be optimized. It follows that MSIL-CRF, like CRFs, can be optimized by efficient solvers, such as L-BFGS [17]. With the application to reducing the annotation cost in semantic segmentation, we demonstrate that the estimated object segments by MSIL-CRF are of high quality, and can replace the manually labeled segments even in challenging segmentation tasks such as Pascal VOC [11].

2. Related Work

We review some relevant research topics in this section.

Semantic Segmentation. Methods of this category, *e.g.*, [1, 3, 12, 15, 18, 20, 24, 27, 29], aim to identify objects of interest and segment them out at the same time. Owing to the high flexibility in modeling the dependencies among variables and observations, CRFs have been widely adopted in the task. For instance, Shotton *et al*. [24] presented an abundant set of features for predicting object classes in the level of pixels, and used CRFs to combine these features. Various high order potential functions for CRFs have been introduced in [1, 15] for expressing the contextual information among object classes. Despite the effectiveness, training CRFs for semantic segmentation usually requires a vast amount of manual efforts on labeling training images in form of object segments.

Multiple Image Segmentations. Although algorithms for image driven segmentation [2, 7, 23] or figure-ground separation [5, 14, 22] are developed with theoretic merits, no universal algorithm or parameter setting can segment all objects with adequate results. This phenomenon has been pointed out in [28]. To alleviate this problem, the strategy of multiple segmentations, *e.g.*, [4, 6, 10], attempts to produce a set of segmentation hypotheses by distinct segmentation algorithms, parameters, or seeding methods. Our approach is relevant to [4, 6, 10], which assume the unknown object contour can be discovered by at least one segment hypothesis.

Segmentation with Low Labeling Cost. Recent research efforts have been made on reducing the labeling cost for object segmentation. *Weakly supervised methods* or *co-segmentation*, *e.g.*, [13, 18, 29], support training data labeled in the levels of images or bounding boxes, instead of object maps. As information regarding object classes has been annotated, the class-specific clues were extracted in these methods to enhance object segmentation. However, weakly supervised labeling is susceptible to large intra-class variations, which obstruct the discovery of the latent object contours. Another type of methods, *e.g.*, [5, 14, 16, 22], for saving manual labeling is interactive segmentation, in which the segmentation process is guided by user input. Distinct from these approaches, our approach adopts MIL to handle the lack of ground truth, and infers the object segments in the bounding boxes by leveraging knowledge transferred from a few manually labeled object contours.

It is worth noting that the *MI-CRF* (conditional random field for multiple instance learning) [8] and the proposed MSIL-CRF are similar in their abbreviations, but they address different problems. MI-CRF deals with an MIL task over the formulation of CRFs. It models bags as nodes in CRFs with instances as their states. The *mi-Graph* [31] and the *MILSD* [30] are two MIL algorithms that further ex-

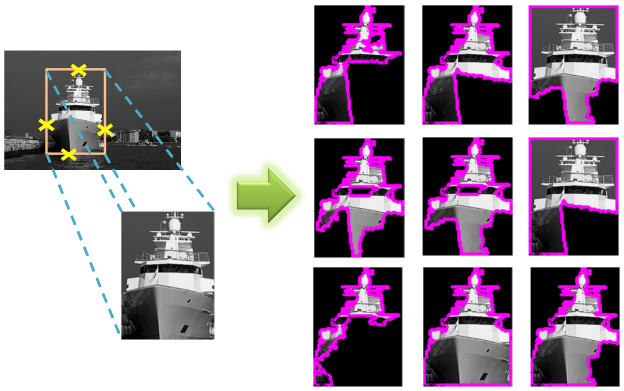

Figure 2. A bounding box and some generated tight segments.

plore the relationships among instances in the same bag. All the aforementioned approaches to MIL work on instances in form of feature vectors. Besides, the *HCRFs* (hidden-state conditional random fields) [21] augment intermediate hidden variables to model the latent structure of the observation. Unlike the foregoing approaches, our approach introduces MIL for addressing the problem caused by the lack of ground truth in learning CRFs. From another perspective, it works on *structured instances* in MIL. Specifically, an instance in our case corresponds to the graph structure over the superpixels in an object's bounding box. To the best of our knowledge, such a generalization of multiple instance learning is novel.

3. Multiple Tight Segment Generation

A set of object segment hypotheses is required for each bounding box to serve as the input to MSIL-CRF. In this work, we adopt the algorithm in [6] to compile multiple *tight segments* for each bounding box. The tight segments are used as the segment hypotheses. A segment is tight with respect to a bounding box if it touches all the four sides of the bounding box. The reason of using tight segments is that the bounding box is the smallest rectangle covering the real object segment, which must be tight. The approach in [6] integrates *bounding box prior* [16] to yield multiple tight segments for a bounding box. More importantly, the approach in [6] makes it more likely that at least one tight segment is close to the real object segment. This property supports the use of MIL in the framework. An example of the bounding box of an object and some of the yielded tight segments is shown in Figure 2.

4. Our Approach

We define the notations, give a brief review of CRFs, and introduce the proposed MSIL-CRF in this section.

4.1. Notation

Suppose a few object segments as well as objects' bounding boxes are annotated in images of an object class, say `horse` in Figure 1. We crop the ROIs in the images, and partition each ROI into *superpixels* by *mean-shift* algorithm [7]. For ROIs of the annotated object segments, we denote them by $L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\ell}$, where ℓ is the number of the segments and $\mathbf{x}_i$ is the feature representation or observation. Vector $\mathbf{y}_i = [\mathbf{y}_i(p)] \in \{0,1\}^{n_i}$ is the label, where n_i is the number of superpixels. $\mathbf{y}_i(p)$ takes value 1 if superpixel p belongs to foreground, and 0 otherwise.

As for ROIs of the annotated bounding boxes, we have $U = \{(\mathbf{x}_i, \{\mathbf{y}_{ij}\}_{j=1}^{T_i})\}_{i=\ell+1}^{\ell+u}$, where u is the number of the bounding boxes, and T_i is the number of the generated tight segments in bounding box i. Note that the segment ground truth in the bounding boxes is unknown. Here, we use the generated tight segments as the candidates. Namely, each bounding box i consists of multiple label vectors, and $\mathbf{y}_{ij} = [\mathbf{y}_{ij}(p)] \in \{0,1\}^{n_i}$ is the segment hypothesis induced by its jth tight segment. Our goal is to leverage information available in $L \cup U$, and select the most plausible tight segment for each bounding box. Then, the selected tight segments are used as training data for any of off-the-shelf approaches to semantic segmentation.

4.2. Conditional random fields

For a given ROI $\mathbf{x}$ of n superpixels, each superpixel is associated with a variable node with two states, *i.e.*, foreground and background, while an edge is added between two nodes if their corresponding superpixels are adjacent. The conditional random fields (CRFs) [25] model the conditional distribution of the figure-ground configurations by $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, where $\mathbf{y} \in \mathcal{Y} = \{0,1\}^n$ and $\boldsymbol{\theta}$ is the set of the model parameters. The posterior distribution $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ of CRFs is a Gibbs distribution, and is written as

$$P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{Z_{\mathbf{x}}} \exp\left(-E(\mathbf{y}, \mathbf{x})\right), \tag{1}$$

where *energy function* $E(\mathbf{y}, \mathbf{x})$ and *partition function* $Z_{\mathbf{x}}$ for normalization are defined as

$$E(\mathbf{y}, \mathbf{x}) = \sum_m \lambda_m f_m(\mathbf{y}, \mathbf{x}), \text{ and} \tag{2}$$

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp\left(-E(\mathbf{y}', \mathbf{x})\right). \tag{3}$$

The energy function is composed of *feature functions* $\{f_m\}$ as well as feature weights $\boldsymbol{\theta} = \{\lambda_m\}$. For the sake of clearness, the adopted feature functions are described later.

With training data $L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\ell}$, parameters $\boldsymbol{\theta}$ in CRFs can be estimated by *maximizing likelihood*, *i.e.*,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{\ell} P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) \tag{4}$$

For numerical consideration, $\boldsymbol{\theta}$ is typically derived by maximizing the log likelihood function

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{\ell} -E(\mathbf{y}_i, \mathbf{x}_i) - \log Z_{\mathbf{x}_i}. \quad (5)$$

$J(\boldsymbol{\theta})$ in general cannot be optimized in closed form. Thus, methods based on gradient ascent are often used to solve (5) with partial derivative

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \lambda_m} = \sum_{i=1}^{\ell} \{-f_m(\mathbf{y}_i, \mathbf{x}_i) + \sum_{\mathbf{y}'_i \in \mathcal{Y}_i} P(\mathbf{y}'_i|\mathbf{x}_i, \boldsymbol{\theta}) f_m(\mathbf{y}'_i, \mathbf{x}_i)\}. \quad (6)$$

The partial derivative in (6) has an intuitive meaning. Its first term is the empirical value of function f_m, while the second term is the expectation of f_m under the current model $\boldsymbol{\theta}$. λ_m is optimized by minimizing their difference.

4.3. The proposed MSIL-CRF

Parameter set $\boldsymbol{\theta}$ in CRFs is learned with few labeled training data $L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\ell}$ in our case. Estimating parameters is hence at a high risk of *overfitting*. Besides, the learned $\boldsymbol{\theta}$ may not well predict $U = \{(\mathbf{x}_i, \{\mathbf{y}_{ij}\}_{j=1}^{T_i})\}_{i=\ell+1}^{\ell+u}$ owing to the *large intra-class variations*. We address the two problems by including U in training. In this way, abundant information in U can regularize the estimation of parameters, and all the data to be predicted are covered in training. Thus, the two problems can be alleviated.

As mentioned in Section 3, one important property about U is that at least one tight segment in each bounding box is close to the ground truth. The maximum likelihood solution of the proposed MSIL-CRF can be accordingly defined as

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{\ell} P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) \prod_{i=\ell+1}^{\ell+u} \max_j P(\mathbf{y}_{ij}|\mathbf{x}_i, \boldsymbol{\theta}). \quad (7)$$

Implied by (7), the mode of seeking the positive instances in MIL is included to deal with the uncertainty in data labeling. The corresponding log likelihood function is

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{\ell} -E(\mathbf{y}_i, \mathbf{x}_i) - \log Z_{\mathbf{x}_i} + \sum_{i=\ell+1}^{\ell+u} \max_j(-E(\mathbf{y}_{ij}, \mathbf{x}_i)) - \log Z_{\mathbf{x}_i}. \quad (8)$$

The max operation in (8) has made $J(\boldsymbol{\theta})$ no longer differentiable. We introduce the *smooth maximum function* [26] (or the log-sum-exp trick) to overcome this problem. It gives the differentiable approximation of max operation by

$$\max_j(-E(\mathbf{y}_{ij}, \mathbf{x}_i)) \simeq \frac{1}{\gamma} \log(\sum_{j=1}^{T_i} \exp(-\gamma E(\mathbf{y}_{ij}, \mathbf{x}_i))), \quad (9)$$

where γ is a positive constant, and is used to control the degree of precision in approximation. We empirically set $\gamma = 2^4$ in this work. It gives a good surrogate for the max function.

By substituting (9) into (8), it can be verified that the partial derivative of the log likelihood function with respect to each feature weight λ_m is

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \lambda_m} = \sum_{i=1}^{\ell} \{-f_m(\mathbf{y}_i, \mathbf{x}_i) + \sum_{\mathbf{y}'_i \in \mathcal{Y}_i} P(\mathbf{y}'_i|\mathbf{x}_i, \boldsymbol{\theta}) f_m(\mathbf{y}'_i, \mathbf{x}_i)\} + \sum_{i=\ell+1}^{\ell+u} \{\sum_{j=1}^{T_i} -\kappa_{ij} f_m(\mathbf{y}_{ij}, \mathbf{x}_i) + \sum_{\mathbf{y}'_i \in \mathcal{Y}_i} P(\mathbf{y}'_i|\mathbf{x}_i, \boldsymbol{\theta}) f_m(\mathbf{y}'_i, \mathbf{x}_i)\}, \quad (10)$$

$$\text{where } \kappa_{ij} = \frac{\exp(-\gamma E(\mathbf{y}_{ij}, \mathbf{x}_i))}{\sum_{j'=1}^{T_i} \exp(-\gamma E(\mathbf{y}_{ij'}, \mathbf{x}_i))}. \quad (11)$$

The partial derivative in (10) also has intuitive justification. By comparing (10) with (6), the part of derivative contributed by data in L is exactly the same. As for the part by data in U, it can be checked that κ_{ij} in (11) is non-negative, and $\sum_{j=1}^{T_i} \kappa_{ij} = 1$. The distribution of $\{\kappa_{ij}\}_{j=1}^{T_i}$ represents our *belief* over all the tight segments of bounding box i. The less the energy, the larger the belief. Therefore, the empirical value of function f_m in this part is a weighted combination of those induced by the guessed labels.

The log likelihood function of MSIL-CRF, $J(\boldsymbol{\theta})$ in (8), is differentiable. The efficient method L-BFGS is adopted to solve the optimization problem in our implementation.

4.4. Tight segment selection

After the optimized parameters θ^* of MSIL-CRF in (8) are obtained, we infer the most plausible tight segment for each bounding box. Specifically, for each bounding box i in $U = \{(\mathbf{x}_i, \{\mathbf{y}_{ij}\}_{j=1}^{T_i})\}_{i=\ell+1}^{\ell+u}$, we pick its π_ith tight segment with

$$\pi_i = \arg\max_j P(\mathbf{y}_{ij}|\mathbf{x}_i, \boldsymbol{\theta}^*). \quad (12)$$

We collect the set of training data $D = L \cup \tilde{U}$ in form of object segments, where $\tilde{U} = \{(\mathbf{x}_i, \mathbf{y}_{i\pi_i})\}_{i=\ell+1}^{\ell+u}$. D can then be used as the input to any of the off-the-shelf semantic segmentation methods, *e.g.*, [12, 27] in our experiments.

4.5. Implementation details

As implied in (8), a training instance induced by an object segment is treated as importantly as that by a bounding box. For better performance, one tunable parameter is introduced for reweighting instances induced by bounding boxes. In addition, we use a regularization term to penalize parameter set whose norm, $\|\boldsymbol{\theta}\|$, is too large. It makes the learned model more stable and effective. The values of the two parameters for reweighting and regularization are determined by cross validation, in which the performance is measured on the labeled training set L. Note that $J(\boldsymbol{\theta})$ in (8) is nonconcave in general, and a local optimum is reached

in practice. We set all the optimization variables as zero in initialization in the experiments.

5. The Energy Function

We describe the adopted energy function (2) in this work. It is composed of three types of energy. For an ROI $\mathbf{x}$ with n superpixels, let $\mathbf{y} = [\mathbf{y}(p)] \in \{0,1\}^n$, $\mathcal{V}$, and $\mathcal{E}$ respectively denote the labeling, the set of superpixels, and the set of edges that connect adjacent superpixels. The energy function is designed as

$$E(\mathbf{y},\mathbf{x}) = \sum_{p\in\mathcal{V}} \phi(\mathbf{y}(p),\mathbf{x}) + \sum_{p\in\mathcal{V}} \psi(\mathbf{y}(p),\mathbf{x}) + \sum_{(p,q)\in\mathcal{E}} \varphi(\mathbf{y}(p),\mathbf{y}(q),\mathbf{x}), \quad (13)$$

where ϕ, ψ, and φ are the *instance-specific unary energy*, the *class-consistent unary energy*, and the *pairwise energy*, respectively.

On Designing ϕ. Energy ϕ encodes the negative log probabilities of each superpixel belonging to foreground and background by considering the features extracted from this ROI. Following [16], we fit a GMM (*Gaussian mixture model*) to the RGB color vectors in the strip of 10 pixels around the ROI, and sort pixels inside the ROI according to their probabilities measured by the GMM. Two GMMs $\mathbf{f}$ and $\mathbf{b}$ are learned with the last 33% and the first 33% of the sorted pixels respectively. The probability of each pixel belonging to foreground (background) can be estimated by $\mathbf{f}$ ($\mathbf{b}$). Since ϕ is applied to superpixels, we simply average the probabilities of pixels falling into each superpixel. Besides, we generalize this energy by replacing the RGB feature with *SIFT* [19] and *texton* [24], respectively. Thus, energy ϕ is composed of three feature functions.

On Designing ψ. Energy ψ encodes the negative log probabilities of each superpixel according to features extracted from ROIs of the same class. Specifically, we generate tight segments for each ROI in L. Following [4], each segment is characterized by the *mid-level features*, and its accuracy, in $[0,1]$, w.r.t. the ground truth. With the features and accuracy rates (target values) of these segments, a regressor is learned to rank tight segments. For ROI $\mathbf{x}$, suppose its first tight segment is predicted by the learned regressor and with regression value r. The superpixels inside the tight segment are accumulated with r, while the rest are with $1-r$. The procedure is repeated for each tight segment of $\mathbf{x}$. Then, the probability of each superpixel belonging to foreground can be estimated. As suggested in [4], we divide the *mid-level features* into three groups, with each of which a regressor is learned. Hence, ψ is composed of three feature functions.

On Designing φ. We adopt the Potts model here.

6. Experimental Results

In this section, we assess the performance of our approach by conducting two sets of experiments in Pascal VOC segmentation task. First, the quality of the tight segments picked by MSIL-CRF is measured. It tests whether MSIL-CRF works well on training data with uncertain labels. Second, the efficacy of using the picked tight segments for semantic segmentation is evaluated. It checks if the picked tight segments by MSIL-CRF can replace the manually labeled segments in semantic segmentation, the underlying goal of this work.

6.1. Pascal VOC 2007 dataset

The Pascal VOC 2007 segmentation dataset [11] contains 20 object classes and one additional category of background. Each object class consists of $30 \sim 100$ annotated object segments except for class `person`, which has 345 ones. The dataset consists of highly deformable objects, and results in substantial annotation costs for manually labeling object segments. Nevertheless, it serves as an appropriate test bed to verify the effectiveness of our approach.

For each annotated object segment, we set its bounding box as the ROI, and crop it from the image. In addition, the resolutions of objects in the dataset are different. Most segmentation algorithms are sensitive to resolutions. Thus, we resize each ROI to around $80,000$ pixels, without changing their aspect ratios.

6.2. Baselines

MSIL-CRF infers the object segments enclosed by a bounding box. For performance comparison, we implemented seven baselines of the following two categories.

Single image figure-ground segmentation. Methods in this category perform figure-ground segmentation by considering a bounding box at a time. Specifically, we adopted the following four approaches, each of which is denoted below in bold and in abbreviation:

- **GrabCut** [22]: It works with initial foreground and background models. The foreground model is initialized with the whole bounding box, while the background model is fitted to the outside region.
- **TS** (tight segment) [16]: Bounding box prior is integrated into figure-ground segmentation. It further ensures that the resulting foreground segment is tight.
- **OP** (object proposal) [10]: It produces a set of object proposals. A pretrained regressor is used to rank these proposals and pick the best one.
- **FG** (F-G classification) [5]: It compiles various foreground priors and one common background prior. Multiple segment hypotheses are generated with different foreground priors. The one that maximizes the score of segmentation quality is selected.

method	**GrabCut** [22]	**TS** [16]	**OP** [10]	**FG** [5]	**DCCoSeg** [13]	**CRFs**	**SSSVR** [6]	**Ours**
median	74.03	68.61	64.97	72.67	62.93	78.19	77.78	**81.63**
mean	71.90	67.34	61.58	70.60	65.90	74.76	74.33	**77.78**

Table 1. The accuracies (%) of various figure-ground segmentation methods in Train+Val set of Pascal VOC 2007.

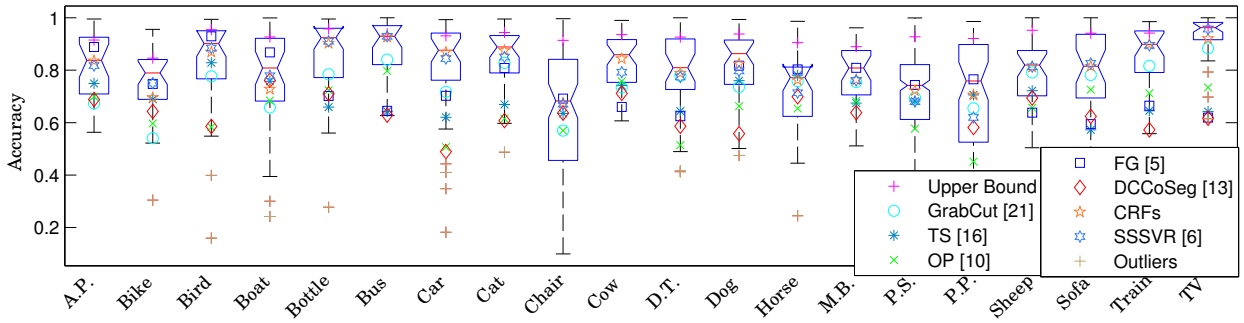

Figure 3. The accuracy distributions of our approach on the 20 object classes. The edges of each blue box are the 25th and 75th percentiles, while the red line indicates the median. The accuracy rates of the seven baselines and the upper bound of our approach are also given.

Class-based figure-ground segmentation. Methods of this category work by considering all the bounding boxes of an object class jointly. Thus, the class-specific knowledge can be derived to benefit figure-ground segmentation. We adopted the following three approaches:

- **DCCoSeg** (discriminative clustering for co-segmentation) [13]: It utilizes a discriminative clustering algorithm to jointly segment out the objects enclosed by the bounding boxes.
- **CRFs**: We learn the CRFs model by using the same setting as that of MSIL-CRF. The only difference is that the CRFs model is learned in (4) with labeled object segments, whereas MSIL-CRF in (7) considers both the labeled segments and the bounding boxes. Comparing MSIL-CRF to the baseline shows the advantage of including bounding boxes in training.
- **SSSVR** (semi-supervised support vector regression) [6]: It learns a semi-supervised SVR regressor by considering both the object segments and the bounding boxes at the same time. It represents each tight segment as a feature vector, while MSIL-CRF treats a tight segment as a structured instance. This baseline is useful to measure the benefit of respecting structured information in segmentation.

6.3. On figure-ground segmentation

We assess the qualities of inferred object segments by comparing them with the ground truth. For a segment $\boldsymbol{\ell}$, its accuracy is computed by $1 - \frac{XOR(R(\boldsymbol{\ell}),GT)}{\#pixel}$, where GT is the ground truth, XOR is the function of *exclusive or*, and $R(\boldsymbol{\ell})$ is a binary vector that indicates each pixel in $\boldsymbol{\ell}$ assigned to either foreground or background.

The four baselines [22, 16, 10, 5] in the first category infer the object contour in a bounding box at a time. Baseline DCCoSeg jointly segments out the common objects in the bounding boxes of the same class. CRFs, SSSVR, and MSIL-CRF require a few labeled object segments to learn their models. Thus, we randomly select 10% of bounding boxes coming with the ground truth, *i.e.*, L, while the rest are treated as bounding boxes, U, and their object segments are assumed to be unknown. Actually, we analyzed the qualities of the inferred object segments of U with different sizes of L, including $\{5\%, 10\%, 15\%, ..., 35\%\}$, and found that the average performance converges when L contains about 10% of bounding boxes. MSIL-CRF picks one tight segment for each bounding box, so the accuracy of the best tight segment is its performance upper bound. The performances of our approach and the baselines are evaluated on U. We first compute the median and mean accuracies over data of each class. The accuracies in median of all the approaches are shown in Figure 3. The average accuracies across the 20 classes are reported in Table 1.

It can be observed in Table 1 that GrabCut, TS, OP, and FG work on a single bounding box where only restricted information is accessible, they often result in suboptimal performance. DCCoSeg seeks object segments with common appearance. However, this assumption may not hold, since there exist large intra-class variations in Pascal VOC. CRFs and SSSVR achieve similar average accuracies, but their class-wise accuracies, shown in Figure 3, are differently distributed. It reveals that the structural clues used in CRFs and the unlabeled data used in SSSVR are complementary. The proposed MSIL-CRF is consistently superior to CRFs and SSSVR. It indicates that MSIL-CRF can effectively make the most of both types of information, leading to promising accuracies. We regard that MSIL-CRF remarkably outperforms all the baselines, since the space for accuracy improvement is limited in Pascal VOC. As shown in Figure 3, each object class consists of several *outliers*, *i.e.*, bounding boxes with low segmentation accuracies. The

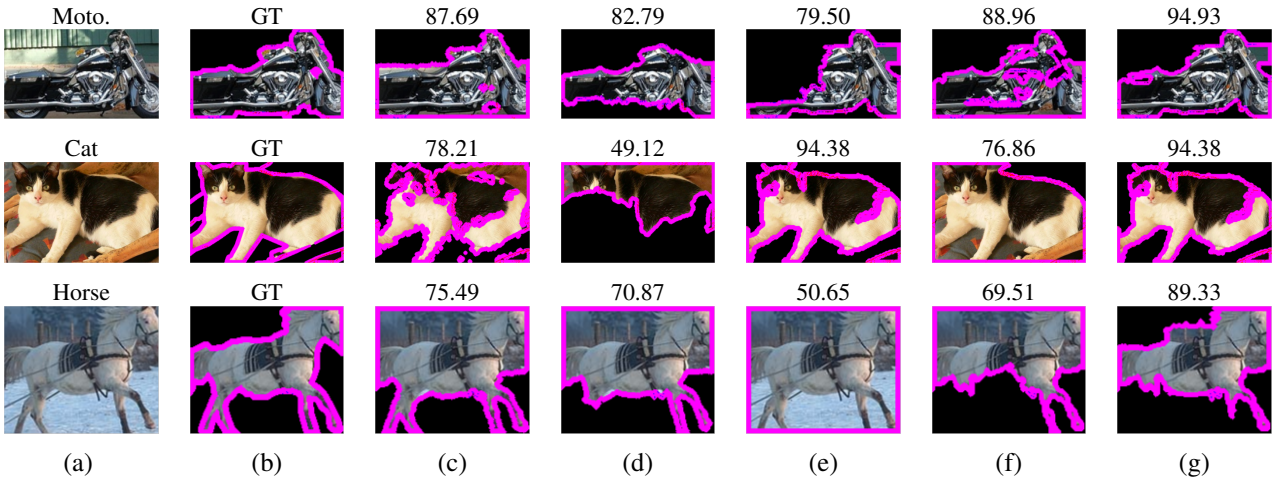

Figure 4. Inferred object segments by various approaches, together with the accuracies (%) shown above. (a) Bounding box. (b) Ground truth. (c) GrabCut [22]. (d) OP [10]. (e) CRFs. (f) SSSVR [6]. (g) Ours.

appearances of these outliers are often far different from those of the rest data. It is difficult to well segment these outliers by existing image descriptors.

In Figure 3, we see that MSIL-CRF works best in most classes. Aside from dealing with the simpler objects with convex shapes like `Bus` and `TV`, MSIL-CRF can also conquer the highly deformable objects, such as `Bird` and `Cat`. The performance of our approach is bounded by the quality of tight segments generated by the method [6]. Our approach does not perform well when none of the tight segments are close to the ground truth. This phenomenon is found in class `Bike`. The worst performance of our approach occurs in class `Chair`. It results from the large intra-class variations presented in this class. The training and inference time of our approach is within two minutes for each of the twenty classes, except for class `person`. Like CRFs, calculating marginal probabilities is the most time-consuming step in our approach.

To gain insight into the quantitative results, several inferred object segments by different approaches are shown in Figure 4. In the first example `Moto.`, MSIL-CRF successfully discovers object segments despite the complex color or texture distributions within the objects, and is superior to other baselines. In the highly deformable objects, *e.g.*, `Cat`, both MSIL-CRF and CRFs surpass SSSVR. The two examples reveal the advantages of using structured information to preserve local consistency. MSIL-CRF significantly outperforms CRFs in the last example `horse` owing to including U in training, since the background in this bounding box is much different from those in L.

6.4. On semantic segmentation

The second experiment aims to corroborate the effectiveness of MSIL-CRF in annotating training data for semantic segmentation. To this end, the ground truth (GT) and the object segments inferred by MSIL-CRF and the seven baselines are respectively used as training data for two state-of-the-art semantic segmentation algorithms [12, 27]. That is, we replace the manually labeled GT by automatically inferred object segments. The semantic segmentation methods [12, 27] learned with the eight (our approach as well as the seven baselines) distinct annotated training data are evaluated on the testing dataset, `test`, in the Pascal VOC 2007 segmentation task.

The quantitative results of the semantic segmentation algorithm [12] w.r.t. the eight copies of training data and the ground truth are reported in Table 2. Table 3 shows the results for another semantic segmentation method [27]. It can be observed that compared with the eight baselines, training with MSIL-CRF's results gives the best performance in semantic segmentation. Our approach also achieves similar performance to training with GT, *i.e.*, 26.01 vs. 25.51 using [12] or 18.99 vs. 19.01 using [27]. It shows that MSIL-CRF can automatically infer object segments in bounding boxes with sufficient quality, and be an effective alternate for the manually drawn GT in semantic segmentation.

It is worth mentioning that the accuracy by MSIL-CRF is slightly higher than that by ground truth in Table 2. It may be because the method [12] tends to overfit the difficult data provided by precise annotations of manual drawings. Our approach annotates training data by leveraging class-specific knowledge, including object contours and bounding boxes. Since the difficult data are relatively sparse, our approach tends to ignore these data due to their inconsistency with the whole class. Vague annotations resulted from our method may instead lower down the importance of this kinds of difficult training data, and lead to a better performance.

	avg.	A.P.	Bike.	Bird	Boat	Bottle.	Bus	Car	Cat	Chair	Cow	D.T.	Dog	Horse	M.B.	P.S.	P.P.	Sheep	Sofa	Train	Tv
GT	25.51	4.01	11.27	0.41	0.03	5.70	25.81	37.79	52.15	12.73	3.30	12.80	33.67	19.35	62.09	59.31	11.89	**22.02**	7.71	44.88	36.00
GrabCut [22]	24.12	5.82	5.44	10.82	4.08	0.73	23.50	**39.91**	52.42	13.21	5.41	6.75	24.70	**28.24**	**65.84**	59.42	10.40	13.64	7.73	41.02	28.81
TS [16]	24.95	12.08	13.93	17.66	3.60	1.41	29.41	36.81	49.98	10.61	2.28	7.56	**33.94**	16.95	54.73	63.86	**12.34**	13.78	7.84	35.10	34.55
OP [10]	21.85	13.92	1.93	13.75	4.38	4.01	18.11	37.02	37.43	15.91	3.76	10.79	32.84	12.06	36.69	59.33	7.20	13.02	7.52	38.00	23.47
FG [5]	24.18	2.90	**24.25**	1.64	1.59	0.36	19.57	39.47	**62.24**	15.09	0.67	7.26	31.48	19.92	59.90	61.45	6.90	15.77	7.76	34.13	27.48
DCCoSeg [13]	23.14	**18.27**	12.59	8.74	**15.04**	**7.14**	23.12	33.79	45.88	**20.07**	**7.90**	31.36	30.89	27.20	45.86	20.32	11.26	16.26	**9.61**	31.43	31.75
CRFs	24.36	6.15	9.29	18.53	2.13	1.02	21.37	39.55	54.94	13.39	2.99	10.61	21.81	22.77	63.16	**61.81**	8.86	16.06	7.18	39.56	26.68
SSSVR [6]	25.03	11.69	1.10	6.08	1.87	2.00	**39.75**	39.44	62.13	6.92	7.51	**37.56**	20.92	11.31	64.04	50.55	9.41	20.67	8.30	39.43	22.80
Ours	**26.01**	2.90	16.12	**23.43**	2.12	1.62	20.66	37.54	56.62	15.45	3.51	14.90	29.60	24.16	60.27	54.67	12.03	15.10	7.52	**46.68**	**37.44**

Table 2. Performance of [12] on Pascal VOC segmentation task w.r.t. various annotated training data by different approaches.

method	**GT**	[22]	[16]	[10]	[5]	[13]	**CRFs**	[6]	**Ours**
avg.	**19.01**	18.33	17.60	17.12	17.70	17.74	18.21	18.11	18.99

Table 3. Average accuracies of [27] on Pascal VOC segmentation task with training data generated by various approaches.

7. Conclusions

We have presented an approach, called MSIL-CRF, which adopts the principle of MIL to learn CRFs with training data with uncertainty in labels. On the other hand, it utilizes the expressive power of CRFs to enhance the performance of MIL by taking structured information into account. In the paper, our approach is applied to inferring object segments enclosed by bounding boxes, and is evaluated on Pascal VOC segmentation task. The promising results demonstrate that the inferred segments are good enough to replace the manually labeled training data in semantic segmentation. In addition, it is developed with theoretic merits: With the differentiable objective function, it can be optimized efficiently by gradient ascent methods, such as L-BFGS. For future work, we aim to generalize MSIL-CRF to handle negative bags, and test it with the applications in which structured information is appreciated, such as image parsing and contextual object recognition.

Acknowledgments. This work was supported in part by grant NSC 102-2221-E-001-025.

References

[1] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Serrat, and J. González. Harmony potentials - fusing global and local scale for semantic image segmentation. *IJCV*, 2012.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 2001.

[3] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *IJCV*, 2012.

[4] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012.

[5] Y. Chen, A. Chan, and G. Wang. Adaptive figure-ground classification. In *CVPR*, 2012.

[6] J.-Z. Cheng, F.-J. Chang, K.-J. Hsu, and Y.-Y. Lin. Knowledge leverage from contours to bounding boxes: A concise approach to annotation. In *ACCV*, 2012.

[7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 2002.

[8] T. Deselaers and V. Ferrari. A conditional random field for multiple-instance learning. In *ICML*, 2010.

[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *AI*, 1997.

[10] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[12] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[13] A. Joulin, F. R. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[14] Z. Kuang, D. Schnieders, H. Zhou, K.-Y. K. Wong, Y. Yu, and B. Peng. Learning image-specific parameters for interactive segmentation. In *CVPR*, 2012.

[15] L. Ladický, C. Russell, P. Kohli, and P. Torr. Associative hierarchical random fields. *TPAMI*, 2013.

[16] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[17] D. Liu and J. Nocedal. On the limited memory BFGS method for large-scale optimization. *Math. Program.*, 1989.

[18] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, 2013.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[20] N. Payet and S. Todorovic. Hough forest random field for object recognition and segmentation. *TPAMI*, 2013.

[21] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *TPAMI*, 2007.

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

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

[24] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[25] C. Sutton and A. McCallum. *An Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2007.

[26] G. Takács. *Convex polyhedron learning and its applications*. PhD thesis, Budapest University of Technology and Economics, Hungary, 2010.

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

[28] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *TPAMI*, 2007.

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

[30] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence. Multiple instance learning on structured data. In *NIPS*, 2011.

[31] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-I.I.D. samples. In *ICML*, 2009.