

# Multiple-Instance Learning with Structured Bag Models

Jonathan Warrell      Philip H. S. Torr

jwarrell@brookes.ac.uk      ptorr@brookes.ac.uk  
Oxford Brookes University  
Oxford, UK

<http://cms.brookes.ac.uk/research/visiongroup>

**Abstract.** Traditional approaches to Multiple-Instance Learning (MIL) operate under the assumption that the instances of a bag are generated independently, and therefore typically learn an instance-level classifier which does not take into account possible dependencies between instances. This assumption is particularly inappropriate in visual data, where spatial dependencies are the norm. We introduce here techniques for incorporating MIL constraints into Conditional Random Field models, thus providing a set of tools for constructing *structured bag* models, in which spatial (or other) dependencies are represented. Further, we show how Deterministic Annealing, which has proved a successful method for training non-structured MIL models, can also form the basis of training models with structured bags. Results are given on various segmentation tasks.

## 1 Introduction

Multiple instance learning (MIL) has received intense interest recently in the field of computer vision, with algorithms being developed for a number of different learning scenarios and specific applications. Methods have been proposed based on the frameworks of boosting [3, 21], SVMs [6, 25], and random forests [12, 16], and special algorithms/techniques proposed for online scenarios [16], scenarios where the proportion of positives in bag is known [6], and multi-label generalizations [24, 25]. In several of these methods, the technique of *deterministic annealing* (DA) has proved an effective method for learning [6, 12, 16], which provides a general framework for learning in weakly supervised settings [13, 14]. Particular vision applications where MIL has had notable success include tracking [3], detection [21], and image classification [12, 25].

Other vision problems have also been targeted for MIL solutions. The framework which MIL provides, with labels at two distinct levels, is in principle ideally suited for tasks such as semantic scene segmentation [17], where it is expensive to collect detailed segmentations of whole scenes at the pixel level, but much easier to acquire annotations at the image level of the objects an image contains. Attempts to apply MIL techniques directly to this problem though have met with limited success. Vezhnevets *et al.* [18] for instance found it necessary to include a multitask learning subproblem to achieve reasonable results with MIL, as the former appeared to be needed as a regularizer over the unlabeled data. This may be explained by appealing to the underlying assumptions: while MIL assumes all instances in a bag to be independent samples, labeling tasks in vision such as scene segmentation typically exhibit strong spatial dependencies between neighboring labels. A recent approach which considers modeling dependencies between

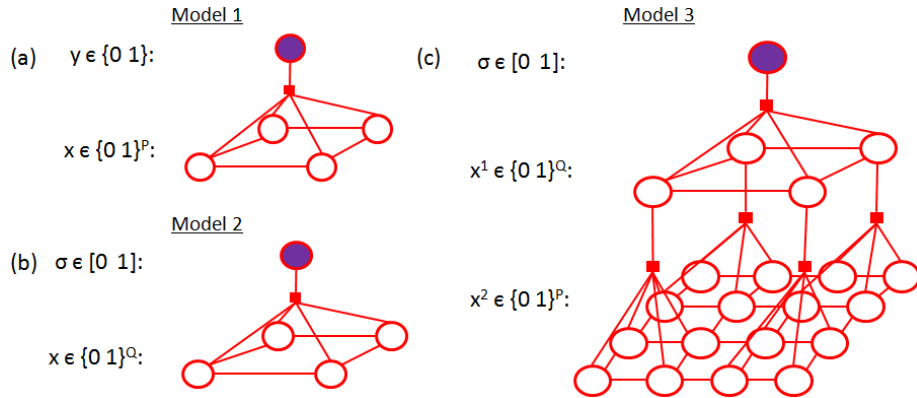
the instances of a bag was proposed in [26]. This approach though builds a graph-kernel classifier directly at the bag level, and so is not directly applicable if we are interested primarily in learning instance classifiers (as in semantic segmentation). A further method uses Hidden Conditional Random Fields and probabilistic EM learning for MIL in a combined scene segmentation/classification setting [24]. In this form, the MIL approach is highly related to other probabilistic models which integrate labels at multiple levels for scene segmentation (e.g. [7]). The disadvantage of such models from an MIL perspective is that they propose structures in which probabilistic sampling is necessary, the efficiency of which determines the feasibility of the learning.

Recent advances in optimization techniques for Conditional Random Field models (CRFs) have made it possible to perform inference in models which encode a variety of structural constraints or assumptions concerning the form of the output. A technique which has been applied widely in this regard is that of *dual decomposition* (DD) [4], which arises from the integer programming formulation of maximum posterior (MAP) inference in CRF models. This approach has been applied for instance to general minimization of non-submodular energies [9], joint optimization of foreground/background segmentation and color models in the context of interactive segmentation [19], penalizing disagreement between foreground histograms in the context of cosegmentation [20], and enforcing marginal statistics on the solutions of labeling problems in binary and multilabel settings [23]. An advantage of dual decomposition is that it allows complex varieties of constraints to be modeled, while still permitting efficient energy minimization techniques such as graph-cuts to be used.

In this paper we propose to use such techniques to embed MIL constraints directly in CRF models. This will provide us a way of building *structured bag* models, which can effectively model the kinds of dependencies between instance variables that we expect in for instance semantic segmentation applications. We consider constraints of two kinds: first the traditional ‘hard’ MIL constraint, where we have a 0-1 bag label indicating the presence or absence of positives at the instance level, and second a ‘soft’ constraint, where a continuous bag label in the interval  $[0, 1]$  indicates the expected proportion of positives (similar to [6], where this information is added to the positive bags to help training). We also consider hierarchical bag models which incorporate both kinds of constraint at different levels. We then provide a method for extending the deterministic annealing framework discussed above to perform learning in such structured models, using boosted models as our instance level classifiers, and requiring only that approximate MAP inference can be performed at the bag level.

We demonstrate the potential of our approach on two vision tasks which both permit MIL formulations, and exhibit structural dependencies appropriate to our models. The first is a binary semantic labeling task involving road-crack detection using the LabelCracks dataset of [22], which we extend with various weaker MIL annotations. The second is the task of interactive cosegmentation, using the iCoseg dataset of [2], in which we consider the extended scenario where the user is free to add soft ‘bag-level’ annotations. We stress though the broad applicability of the techniques introduced, particularly for scene segmentation problems with weak annotations as discussed above.

Sec. 2 introduces our basic structured bag models, Secs. 3 and 4 give details on inference and training, 5 and 6 give experimental evaluations and 7 offers a discussion.



**Fig. 1.** Summary of structured bag models considered in the paper. Model 1 (a) enforces a hard MIL constraint between a binary bag label  $y$  and instance labels  $x$ , with internal pairwise dependencies, while model 2 (b) enforces a soft constraint between a continuous bag label  $\sigma$  and the instances. In model 3 (c) a two level bag model is formed, combining both hard and soft constraints. The bag labels  $y$  and  $\sigma$  may be observed or unobserved during inference (see Sec. 3).

## 2 Structured Bag Models

We begin by outlining the basic bag models we will consider in Sections 2.1 and 2.2, involving hard and soft MIL constraints respectively. Section 2.3 then outlines a more complex 2-level bag model, which is motivated by the experimental data of section 5.

### 2.1 Model 1 (hard constraints)

The first model we consider is shown in Figure 1a. Here we have a simple bag model, consisting of a binary *bag label*  $y \in \{0, 1\}$ , and *instances*  $x_p$ ,  $p = 1 \dots P$ , also taking binary labels. We shall denote observations associated with each instance by  $\mathbf{z}_p$ . Given this notation, we define an energy for the bag as:

$$E^1(\mathbf{x}, y|\mathbf{z}) = E_{\text{base}}(\mathbf{x}|z) + \phi^{\text{hard-MIL}}(\mathbf{x}, y) \quad (1)$$

where:

$$E_{\text{base}}(\mathbf{x}|z) = \sum_p \phi^{\text{unary}}(x_p, \mathbf{z}_p) + \sum_{p_1, p_2 \in \mathcal{N}} \phi^{\text{pair}}(x_{p_1}, x_{p_2}, \mathbf{z}_p) \quad (2)$$

$\phi^{\text{unary}}$  and  $\phi^{\text{pair}}$  representing standard unary and pairwise terms in a base CRF energy (with  $\mathcal{N}$  the neighborhood relation) and:

$$\phi^{\text{hard-MIL}}(\mathbf{x}, y) = \begin{cases} 0 & \text{if } \max_p x_p = y \\ \infty & \text{otherwise} \end{cases} \quad (3)$$

Eq. 3 ensures the MIL constraint is enforced between  $\mathbf{x}$  and  $y$ , giving finite energy only if it is. Implicitly, this energy model also implies a probability model, where  $Pr(\mathbf{x}, y|\mathbf{z}) \propto \exp(-E(\mathbf{x}, y|\mathbf{z}))$ .

## 2.2 Model 2 (soft constraints)

Model 2 is illustrated in Figure 1b. The only difference with model 1 is that now in place of the binary bag labels  $y$  we have continuous bag labels  $\sigma \in [0, 1]$  which represent the expected number of positives in the bag (we also change the index  $p$  to  $q$  to match the notation in model 3). The bag energy now takes the form:

$$E^2(\mathbf{x}, \sigma | \mathbf{z}) = E_{\text{base}}(\mathbf{x} | \mathbf{z}) + \phi^{\text{soft-MIL}}(\mathbf{x}, \sigma) \quad (4)$$

where  $E_{\text{base}}(\mathbf{x} | \mathbf{z})$  is as in (2), and:

$$\phi^{\text{soft-MIL}}(\mathbf{x}, \sigma) = g(\sigma \cdot Q - \sum_q x_q) \quad (5)$$

Here  $g$  is any convex function, for example the  $l_1$ -distance  $g(\cdot) = |\cdot|$  or  $l_2$ -distance  $g(\cdot) = (\cdot)^2$ . The intended semantics of  $g$  are that it penalizes the difference between the expected number of positives, and the observed count. We note that these choices of function do not require bags labeled  $\sigma = 0$  to have no positives; if we wish to do that, we must use the convex function  $g$  which is 0 at 0, and  $\infty$  elsewhere, enforcing the *exact* counts are observed for all bags. This may be too restrictive in practice.

## 2.3 Model 3 (combined model)

Our third model is illustrated in Figure 1c. This combines the two types of constraints outlined above to form a more complex bag structure, with instances labels on two levels (the first level forming a set of ‘sub-bag’ labels for the second). The overall bag label is again  $\sigma \in [0, 1]$ , which here represents the expected number of positives in the first level of instances,  $x_q^1$ ,  $q = 1 \dots Q$ . Each of the  $x_q^1$ ’s in turn acts as a hard bag label to a subset of instances at the second level,  $x_p^2$ ,  $p = 1 \dots P$ , and we write  $\mathbf{x}_q^2$  to denote the subset associated with  $x_q^1$ , and  $x_{q_p}^1$  for the  $x_q^1$  for which  $x_p^2 \in \mathbf{x}_q^2$  (assuming non-overlapping sub-bags for simplicity). Given this notation, the bag energy is:

$$E^3(\mathbf{x}^1, \mathbf{x}^2, \sigma | \mathbf{z}^1, \mathbf{z}^2) = E_{\text{base}}(\mathbf{x}^1 | \mathbf{z}^1) + E_{\text{base}}(\mathbf{x}^2 | \mathbf{z}^2) + \phi^{\text{soft-MIL}}(\mathbf{x}^1, \sigma) + \sum_q \phi^{\text{hard-MIL}}(\mathbf{x}_q^2, x_q^1) \quad (6)$$

# 3 Inference using Dual Decomposition

We now consider how to perform MAP inference in the above models. In all cases, the base energy  $E_{\text{base}}$  is assumed to be submodular.

## 3.1 Model 1 (hard constraints)

If neither  $\mathbf{x}$  or  $y$  are known, MAP inference in (1) consists simply of performing a graph-cut to find  $\mathbf{x}^* = \text{argmin}_x E_{\text{base}}(\mathbf{x} | \mathbf{z})$ , and setting  $y^* = \max_p x_p^*$ . If  $y$  is known, the only problematic case is when  $y = 1$  and performing a graph-cut returns  $\mathbf{x}^* = \mathbf{0}$ . In this case, we may perform additional graph-cuts forcing each instance  $x_p$  to be positive in turn and take the solution with the lowest energy, or as an approximation, force only the instance with the highest unary response, and take the associated solution.

### 3.2 Model 2 (soft constraints)

With unknown  $\mathbf{x}$  and  $\sigma$ , inference in (4) is again simple, as we must only make a single graph-cut to find  $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} E_{\text{base}}(\mathbf{x}|\mathbf{z})$ , and set  $\sigma^* = (1/Q) \sum_q x_q^*$ . With a known  $\sigma$ , recent methods for MAP inference suggest a *dual decomposition* approach (DD), as in ‘area matching’ over a Marginal Probability Field (MPF) [23]. We give a slightly simpler DD algorithm than [23] below, which we then extend to model 3 in Sec. 3.3.

We reformulate Eq. 4 as an integer program, introducing a (redundant) variable  $a$  to encode the soft MIL constraint between  $\sigma$  and  $\mathbf{x}$  and dropping the dependencies on  $\mathbf{z}$ , leading to the primal objective:

$$\min_{\mathbf{x}, a} [E_{\text{base}}(\mathbf{x}) + g(a)] \quad (7)$$

where  $g$  is a convex function as in Sec. 2.2, and we have constraints:

$$a = \sum_q x_q - \sigma \cdot Q, \quad \mathbf{x} \in \{0, 1\}^Q, \quad a \in \{-Q \dots Q\} \quad (8)$$

Introducing a Lagrange multiplier  $\lambda$  for the first constraint, the dual objective is:

$$\max_{\lambda} \psi(\lambda) = \max_{\lambda} (\min_{\mathbf{x}} [E_{\text{base}}(\mathbf{x}) - \langle \lambda \cdot \mathbf{1}, \mathbf{x} \rangle] + \min_a [g(a) + \lambda a] + \lambda \sigma Q) \quad (9)$$

where  $\lambda$  is unconstrained. Evaluating the dual requires we solve two minimization problems to give us  $\mathbf{x}^*$  and  $a^*$  for a given  $\lambda$ . The first is straightforward, and requires a single graph-cut on the energy  $E_{\text{base}}(\mathbf{x})$  with  $-\lambda$  added to the positive unary terms. The second can be solved analytically for most convex  $g$ . For instance, for  $g(\cdot) = |\cdot|$  we set:

$$a^* = -Q \text{ if } \lambda > 1, \quad Q \text{ if } \lambda < -1, \quad 0 \text{ otherwise} \quad (10)$$

We can thus maximize (9) by setting  $\lambda = 0$ , and taking steps along the subgradient:

$$\frac{\partial \psi}{\partial \lambda} = a^* - \sum_q x_q^* + \sigma \cdot Q \quad (11)$$

Several methods are available for choosing step sizes which are guaranteed to converge to the solution of the dual (9) (see [4]). A solution to the primal is only guaranteed if the duality gap is closed at this value. However, a reasonable solution here, and also if the algorithm is not run to convergence, is to take the  $x^*$  with the lowest base energy seen across all iterations.

### 3.3 Model 3 (combined model)

We now extend the dual decomposition approach outlined in Section 3.2 to the more complex bag structure of model 3. Here, we note that even in the case that  $\sigma$  is unknown, inference is now difficult since the hard MIL constraints must be enforced between  $\mathbf{x}^1$  and  $\mathbf{x}^2$ . Instead of considering this case separately, we outline below the case for inference with known  $\sigma$ , which is more complex. For  $\sigma$  unknown, we can simply remove the  $\sigma$  terms from the objectives below, solve for  $\mathbf{x}^1$  and  $\mathbf{x}^2$ , and set  $\sigma = (1/Q) \sum_q x_q^1$ .

As above, we reformulate (6) as an integer program, giving the primal objective:

$$\min_{\mathbf{x}^1, \mathbf{x}^2, a} [E_{\text{base}}(\mathbf{x}^1) + E_{\text{base}}(\mathbf{x}^2) + g(a)] \quad (12)$$

We now include additional constraints (2 and 3 below) to encode the hard MIL constraints between  $\mathbf{x}^1$  and  $\mathbf{x}^2$ , giving:

$$\begin{aligned} a &= \sum_q x_q^1 - \sigma \cdot Q & \mathbf{x}^1 &\in \{0 \ 1\}^Q \\ \forall p \quad x_p^2 &\leq x_{q_p}^1 & \mathbf{x}^2 &\in \{0 \ 1\}^P \\ \forall q \quad \sum_{\{p \in \mathbf{x}_q^2\}} x_p^2 &\geq x_q^1 & a &\in \{-Q \dots Q\} \end{aligned} \quad (13)$$

Introducing Lagrangian multipliers, we form the dual objective by relaxing the constraints in (13):

$$\begin{aligned} \max_{\lambda} \psi(\lambda) &= \max_{\lambda} (\min_{\mathbf{x}^1, \mathbf{x}^2, a} [E_{\text{base}}(\mathbf{x}^1) + E_{\text{base}}(\mathbf{x}^2) + g(a) \\ &+ \lambda^1 (a - \sum_q x_q^1 + \sigma Q) + \sum_p \lambda_p^2 (x_p^2 - x_{q_p}^1) + \sum_q \lambda_q^3 (x_q^1 - \sum_{\{p \in \mathbf{x}_q^2\}} x_p^2)]) \end{aligned} \quad (14)$$

subject to:

$$\lambda^1 \text{ unconstrained}, \quad \lambda^2, \lambda^3 \geq 0 \quad (15)$$

We may further group (14) as follows:

$$\begin{aligned} \max_{\lambda} \psi(\lambda) &= \max_{\lambda} (\min_{\mathbf{x}^1} [E_{\text{base}}(\mathbf{x}^1) + \langle (\lambda^3 - \lambda'^2 - \lambda^1 \cdot \mathbf{1}), \mathbf{x}^1 \rangle] \\ &+ \min_{\mathbf{x}^2} [E_{\text{base}}(\mathbf{x}^2) + \langle (\lambda^2 - \lambda'^3), \mathbf{x}^2 \rangle] \\ &+ \min_a [g(a) + \lambda^1 a] + \lambda^1 \sigma Q) \end{aligned} \quad (16)$$

where we have defined  $\lambda_q'^2 = \sum_{\{p \in \mathbf{x}_q^2\}} \lambda_p^2$  and  $\lambda_p'^3 = \lambda_{q_p}^3$ .

Evaluating the dual then requires solving the 3 minimization problems in (16) to give  $\mathbf{x}^{1*}$ ,  $\mathbf{x}^{2*}$  and  $a^*$  for a given  $\lambda$ . The first two can be solved by single graph cuts on  $\mathbf{x}^1$  and  $\mathbf{x}^2$ , where we simply add to the positive unaries the values implied by the inner products in (16).  $a^*$  can again be found for  $g(\cdot) = |\cdot|$  as in (10). Using these solutions, we derive a subgradient at  $\lambda$  as:

$$\frac{\partial \psi}{\partial \lambda^1} = a^* - \sum_q x_q^{1*} + \sigma \cdot Q, \quad \frac{\partial \psi}{\partial \lambda_p^2} = x_p^{2*} - x_{q_p}^{1*}, \quad \frac{\partial \psi}{\partial \lambda_q^3} = x_q^{1*} - \sum_{\{p \in \mathbf{x}_q^2\}} x_p^{2*} \quad (17)$$

As before, we maximize the dual by starting with  $\lambda = \mathbf{0}$  and taking steps along this subgradient (using a scheme as in [4]). Several additional operations are required compared to model 2. First, we must check with each update that the constraints (15) are satisfied, projecting  $\lambda$  if not. Further, a rounding step is required to produce a valid solution, as  $\mathbf{x}^{1*}$  and  $\mathbf{x}^{2*}$  for any given  $\lambda$  may not respect the inter-layer MIL constraints. A simple scheme may be used, such as setting  $\mathbf{x}_q^2 = 0$  for each  $x_q^1 = 0$ , and treating positive sub-bags violating the constraints using the inference techniques in Sec. 3.1. The rounded primal energy is checked at each iteration, and the minimum selected.

## 4 Training using Deterministic Annealing

We describe here how the deterministic annealing (DA) approach used for training MI-Forests in [12] may be adapted to train structured bag models of the kind proposed. No assumptions are made about the nature of the unary classifiers (which are the main objects to be trained), and the algorithms given may be applied for instance to boosted classifiers, SVMs or random forests (RFs).

### 4.1 Model 1 (hard constraints)

We outline the DA algorithm first in detail for the simplest case of model 1. Modifications for models 2 and 3 are discussed in Sections 4.2 and 4.3. We consider we have a training set of  $I$  bags. We assume for simplicity we have only the bag labels,  $y_{i=1\dots I}$  and observations,  $\mathbf{z}_{i=1\dots I}$ , since additional instance observations are easily incorporated. Since we are working with a structured bag model, we formulate our loss in terms of the score (which will be derived from the energy) given by the model to the whole bag, which we write as  $F(\mathbf{x}, y|\mathbf{z})$ . Given this notation, and a loss function  $l$ , we may write the overall training objective as:

$$(F^*, \mathbf{x}^*) = \operatorname{argmin}_{F, \mathbf{x}} \sum_i l(F(\mathbf{x}_i, y_i|\mathbf{z}_i)) \quad (18)$$

As discussed in [12], optimizing such an objective is difficult due to the large search space for  $\mathbf{x}$ . We thus propose to optimize Eq. 18 by DA, producing an auxiliary objective by introducing a distribution over the instance labels,  $\pi(\mathbf{x})$ , and a ‘temperature’  $T$ :

$$(F^*, \pi^*) = \operatorname{argmin}_{F, \pi} \sum_i \sum_{\mathbf{x}_i} \pi(\mathbf{x}_i) l(F(\mathbf{x}_i, y_i|\mathbf{z}_i)) - T \sum_i \mathcal{H}(\pi_i) \quad (19)$$

where  $\mathcal{H}(\pi_i) = -\sum_{\mathbf{x}_i} \pi(\mathbf{x}_i) \log(\pi(\mathbf{x}_i))$  is the entropy of the distribution over bag  $i$ . Eq. 19 can be minimized by starting at a high  $T$ , and alternately updating  $F$  and  $\pi$  while reducing  $T$ . When  $T = 0$ , Eq. 19 reduces to Eq. 18, and hence we are optimizing the original objective.

Since [12] was working with independent classifier responses, updating  $\pi$  in (19) was automatically a convex problem. In our case, since we are dealing with responses for full bags, we must be careful how we formulate  $F$  to maintain tractability. The straightforward option of basing  $F$  on the CRF distribution implied by (1) (i.e.  $F = Pr(\mathbf{x}_i, y_i|\mathbf{z}_i) \propto \exp(-E_i)$ ) will not work, as it involves estimating the partition function. Instead, we choose to base  $F$  on a simplified distribution which is easier to work with. We factorize this distribution as follows:  $Pr'(\mathbf{x}_i, y_i|\mathbf{z}_i) = Pr'(y_i|\mathbf{z}_i)Pr'(\mathbf{x}_i|y_i, \mathbf{z}_i)$ . The first term, we derive from the energy of the mode (MAP estimate) of (1) when  $y_i$  is fixed to the specified value. We write these modes as  $\mathbf{f}_i^{y_i} = \operatorname{argmin}_{\mathbf{x}_i} E^1(\mathbf{x}_i, y_i|\mathbf{z}_i)$  (noting that  $\mathbf{f}_i^0 = \mathbf{0}$  automatically), and let  $Pr'(y_i|\mathbf{z}_i) = (1/Z_i) \exp(-E^1(\mathbf{f}_i^{y_i}, y_i|\mathbf{z}_i))$ , where we note that normalization is now only over two terms,  $Z_i = \sum_{y_i=\{0,1\}} \exp(-E^1(\mathbf{f}_i^{y_i}, y_i|\mathbf{z}_i))$ . For the second term,  $Pr'(\mathbf{x}_i|y_i, \mathbf{z}_i)$ , we introduce an uncertainty parameter  $\beta$ , which indicates the probability that an instance label may be flipped from the

mode estimate. We base  $F$  directly on the joint probability of  $\mathbf{x}_i$  and  $y_i$  in this altered distribution:

$$F(\mathbf{x}_i, y_i | \mathbf{z}_i) = (1/Z_i) \exp(-E^1(\mathbf{f}_i^{y_i}, y_i | \mathbf{z}_i)) \prod_p \beta^{[f_{ip}^{y_i} \neq x_{ip}]} (1 - \beta)^{[f_{ip}^{y_i} = x_{ip}]} \quad (20)$$

We note that by this approximation we have introduced a (small) probability that the MIL constraints will be violated: this is not a problem though if we are concerned with MAP inference, as the modes  $\mathbf{f}^{y_i}$  will not violate the constraints, so at  $T = 0$  all the probability mass in (19) will be placed on a valid labeling. Finally, we take the loss to be the negative log-likelihood,  $l(\cdot) = -\log(\cdot)$ .

The advantage of the factorized distribution form used in (20) is that the losses incurred by the instances become independent given  $F$  and the bag labels  $\mathbf{y}$ . This means that we know the optimal  $\pi^*$  will factor as  $\pi^* = \prod_{ip} \pi_{ip}^*$ , where  $\pi_{ip}^*$  is the probability that  $x_{ip}$  is positive, and finding these factors reduces to solving a series of convex problems. Differentiating (19) with respect to  $\pi_{ip}$  and setting to zero yields:

$$\pi_{ip}^* = \frac{1}{1 + \exp(l_{ip}/T)} \quad (21)$$

where  $l_{ip}$  is the loss that would be incurred by labeling  $x_{ip}$  as positive, which, as in [12] is the negative log of the positive margin:

$$l_{ip} = -\log \frac{\beta^{[f_{ip}^{y_i}=0]} (1 - \beta)^{[f_{ip}^{y_i}=1]}}{\beta^{[f_{ip}^{y_i}=1]} (1 - \beta)^{[f_{ip}^{y_i}=0]}} \quad (22)$$

Optimizing (19) with respect to  $F$  for fixed  $\pi$  and  $\mathbf{y}$  is more involved, and in general will depend on the form of classifiers used in the potentials of the bag CRF models. As in [12], we first sample instance labels  $x_{ip}$  across all positive bags according to  $\pi_{ip}^*$  (forcing at least one instance to be positive). We then update the bag parameters to maximize the joint probability of  $\mathbf{x}_i$  and  $y_i$ . We note that the problem of CRF training is simplified in our case given the factorized form of (20), and a reasonable piecewise method is to (1) optimize the unary potentials using the current instance samples, and (2) optimize remaining parameters, i.e.  $\beta$  in (20) and possibly a weight  $\alpha_{\text{pair}}$  on the base pairwise potential, by 1-d line searches, where MAP inference only is required at each step to evaluate (19), which can be performed as in Sec. 3.1. In practice, step (2) can be omitted, and parameters such as  $\alpha_{\text{pair}}$  and  $\beta$  treated as fixed hyperparameters set by cross-validation over several runs of DA.<sup>1</sup>

## 4.2 Model 2 (soft constraints)

The DA algorithm for model 2 closely mirrors that for model 1, with some exceptions which we draw attention to below. The overall and auxiliary training objectives are as

<sup>1</sup> Our approach in Sec. 4.1 is similar to the *pseudolikelihood* (see [5]) in that we ensure the overall objective (Eq. 18) factorizes across the losses on each variable (Eq. 22). However, unlike the *pseudolikelihood*, which bases these losses on the true marginals (via each variable's Markov blanket), we base ours on the altered distribution  $Pr'(\mathbf{x}, y | \mathbf{z})$ , which is centered on the modes of  $Pr(\mathbf{x} | y, \mathbf{z})$ , and factorizes into a product of marginals by construction.

in Eqs. 18 and 19, where we simply replace the binary bag labels  $y_i$  with continuous labels  $\sigma_i$ . Again, we use a simplified form of distribution to generate the bag responses  $F$  as in (20). However, we adopt a slightly different form for this distribution to cater for the more complex energy form of model 2. We begin by quantizing  $\sigma$  into a number of levels,  $\sigma = k$ ,  $k \in \{k_1, k_2 \dots k_K\}$ , effectively forming a set of classes. As before, we factorize the distribution  $P_{r'}(\mathbf{x}_i, \sigma_i | \mathbf{z}_i) = P_{r'}(\sigma_i | \mathbf{z}_i) P_{r'}(\mathbf{x}_i | \sigma_i, \mathbf{z}_i)$ , where the first term is now dependent on the MAP solutions of  $E^2$  (Eq. 4) when  $\sigma_i$  is set to each value  $k$  in turn:  $P_{r'}(\sigma_i = k | \mathbf{z}_i) = (1/Z_i) \exp(-E^2(\mathbf{f}_i^k | \mathbf{z}_i))$ , writing  $\mathbf{f}_i^k$  for the mode when  $\sigma_i$  is set to  $k$  (i.e.  $\text{argmin}_{\mathbf{x}_i} E^2(\mathbf{x}_i, k | \mathbf{z}_i)$ ), which can be inferred by MAP inference as in Sec. 3.2. We note that exact MAP inference is not required (and cannot be guaranteed with dual decomposition): we need only a deterministic way of finding modes to approximate the full CRF distribution, which can be achieved by, say, using a fixed number of subgradient updates, and always initializing to  $\lambda = \mathbf{0}$ . Further, normalization is again tractable,  $Z_i = \sum_k \exp(-E^2(\mathbf{f}_i^k, k | \mathbf{z}_i))$ . The bag responses take the form:

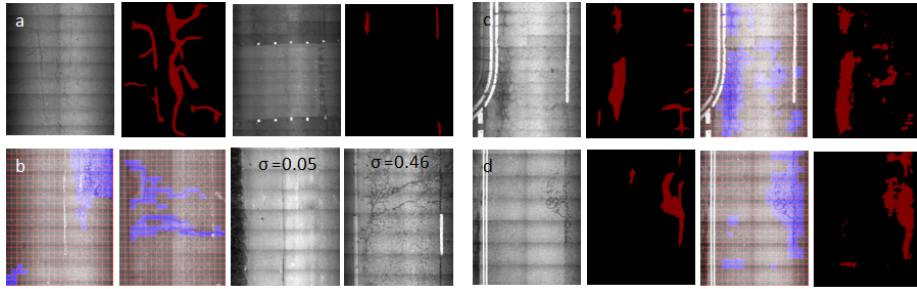
$$F(\mathbf{x}_i, \sigma_i = k | \mathbf{z}_i) = (1/Z_i) \exp(-E^2(\mathbf{f}_i^k, k | \mathbf{z}_i)) \prod_q \beta^{[f_{iq}^k \neq x_{iq}]} (1 - \beta)^{[f_{iq}^k = x_{iq}]} \quad (23)$$

The factorized form of (23) means that again it is straightforward to optimize with respect to  $\pi$  for fixed  $F$  and  $\sigma$ , and the optimal updates again take the same form as Eq. 21, where  $l_{iq}$  is defined as was  $l_{ip}$  in (22), but substituting  $f_{iq}^{\sigma_i}$  for  $f_{ip}^{y_i}$ . Optimizing the CRF parameters can again be done in a piecewise fashion, after sampling new instances labels from  $\pi$ . There is now potentially an extra weight  $\alpha_{\text{soft}}$  for the soft-MIL potential, which can be set by line search, or treated as a hyperparameter.

A further issue arises in model 2 regarding inference when this model is trained by the above DA procedure. Since the model was trained under the factorized approximation of the CRF energy implied by (23), the correct procedure for MAP inference for a bag with unknown  $\sigma$  is to set  $\sigma = k$  for each of the  $\{k_1, k_2 \dots k_K\}$  quantized levels in turn, run dual decomposition on each and take the solution with the minimum energy. This contrasts with the simpler procedure mentioned in Sec. 3.2 of simply performing a single cut on the base energy and setting  $\sigma = (1/Q) \sum_q x_q$ , which would assume our approximation of the original energy during training good enough to permit switching at test time. In our experimentation we use the former approach, and leave investigation of relationship between these solutions to future work.

### 4.3 Model 3 (combined model)

The DA algorithm for model 3 is essentially the same as for model 2, and again involves quantizing the continuous bag labels  $\sigma$  into a set of discrete levels. The modes  $\mathbf{f}_i^k$  are now the (approximate) MAP solutions across both levels  $\mathbf{x}^1$  and  $\mathbf{x}^2$  when  $\sigma = k$ , which are found through the DD algorithm of Sec. 3.3 with appropriate rounding to respect the hard MIL constraints between levels.  $F$  again takes the same form as in (23), where  $P_{r'}(\mathbf{x}_i | \sigma_i, \mathbf{z}_i) = \prod_q \beta^{[f_{iq}^k \neq x_{iq}^1]} (1 - \beta)^{[f_{iq}^k = x_{iq}^1]} \cdot \prod_p \beta^{[f_{ip}^k \neq x_{ip}]} (1 - \beta)^{[f_{ip}^k = x_{ip}]}$  is simply adapted to run across both levels. This implies the same updates can be used for  $\pi$  (an extra check is needed though in sampling to ensure all the constraints are satisfied), and piecewise training of the CRF model can proceed by training the unary potentials on each level separately, and setting the remaining parameters to optimize (19) directly.



**Fig. 2.** Example annotations and results on the LabelCracks dataset [22] for road crack detection. Shown are (a) fully labeled data at the pixel level, (b) weak MIL annotations at the patch level (blue indicates patches containing cracks), and image level ( $\sigma$  denotes proportion of cracking), (c-d) results from model 1 (DA) showing original image, ground truth, patch-level estimates, pixel-level estimates (see Sec. 5). Best viewed in color.

	Image Level (% correct)				Patch Level (union-intersect)				Pixel Level (union-intersect)			
	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
Supervision	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
Model 1	36.1	39.5	33.1	34.9	0.232	0.225	0.246	<b>0.248</b>	0.123	0.126	<b>0.140</b>	<b>0.147</b>
Model 2	40.7	40.7	39.0	<b>45.9</b>	0.152	0.181	0.182	0.171	-	-	-	-
Model 3	44.8	45.4	47.7	45.4	0.192	0.216	0.212	0.209	0.085	0.118	0.127	0.118
Model 1 (DA)	30.8	32.6	34.3	34.9	<b>0.244</b>	<b>0.249</b>	<b>0.253</b>	<b>0.248</b>	<b>0.131</b>	<b>0.137</b>	0.138	<b>0.147</b>
Model 2 (DA)	<b>45.9</b>	44.2	42.4	<b>45.9</b>	0.164	0.158	0.166	0.171	-	-	-	-
Model 3 (DA)	<b>45.9</b>	<b>47.7</b>	<b>50.6</b>	45.4	0.169	0.202	0.198	0.209	0.083	0.106	0.108	0.118

**Table 1.** Summary of road crack detection results across all models. Notice particularly (1) the generally higher performance of the models trained with deterministic annealing (DA) (exceptions discussed in text), (2) the higher performance of the more complex bag structures (models 2 and 3) at the image level, and (3) the higher performance of model 1 at patch and pixel levels.

## 5 Semi-supervised Segmentation with weak MIL Annotations (Road Crack Detection)

We first make a comparison of the models proposed using a binary segmentation task involving detecting cracked regions in road surface imagery. We use the LabelCracks dataset from [22], which provides 100 such fully annotated images. To these, we add 572 further images which we weakly annotate by overlaying each with a grid of  $25 \times 20$  squares, and marking a square as positive if it contains cracking (example annotations shown in Figure 2a-b.) The patch and pixel annotations correspond directly to the levels  $\mathbf{x}^1$  and  $\mathbf{x}^2$  in model 3 above, and  $\sigma$  labels for all images are derived via the proportion of positive-labeled patches. These annotation types all correspond to types of output with application interest. We divide original and new images 0.75/0.25 for training/testing.

Several design choices are held constant across models. We extract from the images HOG and Texton features, and use boosted classifiers at both pixel and patch levels on these features for the unary potentials in our models. These use weak-learners which threshold feature responses in regions of random size/offset from the pixel/patch-

Model 1 (DA)	Patch Level				Pixel Level			
Proportion labeled	0.25	0.5	0.75	1	0.25	0.5	0.75	1
Precision (Low)	0.267	0.270	0.283	<b>0.288</b>	0.144	0.149	0.156	<b>0.172</b>
Recall	0.744	0.756	0.709	0.639	0.596	0.627	0.537	0.503
Precision (High)	0.471	0.473	<b>0.504</b>	0.460	0.293	0.319	<b>0.386</b>	0.259
Recall	0.176	0.161	0.164	0.131	0.084	0.084	0.100	0.054

**Table 2.** Precision and Recall values for model 1 (DA) at patch and pixel levels. These provide alternatives to the union-intersection metric used in Table 1, and results show how these scores can be traded off by varying parameter settings.

center in the manner of TextonBoost [17]. Further, our pairwise terms are based on contrast-sensitive Potts models at the pixel-level (see [8]), and an Ising model with a single parameter at the patch-level. Dual decomposition is run for 5 iterations, using the sub-gradient update procedure outlined in [9], and deterministic annealing is run for 5 rounds, starting at  $T = 5$  and using the update  $T^{n+1} = T^n / 1.5$  each round.

We compare the performance of the models under several conditions. We divide the training sets (pixel and patch annotations) in four equal parts, and test the models when the full annotations are provided for 0.25, 0.5, 0.75 and all of the training data. In the first three cases, for the remaining data only the bag labels are visible, consisting of patch-level annotations for pixel-data, and  $\sigma$  annotations for the patch-data. Each model is trained under two scenarios: (a) using only the supervised data, and (b) using both the supervised subset and the extra bag labeled data. In only the latter case is the DA training used, but dual decomposition is used at testing for models 2 and 3 in both training scenarios (i.e. they must propose a consistent labeling across levels). We quantize  $\sigma$  into  $K = 3$  levels for models 2 and 3 (covering equal probability mass in the training distribution, giving  $\sigma \in \{0.04, 0.1, 0.2\}$ ), and measure performance in terms of %-correct classification at the image level, and the union-intersection score at patch and pixel levels (using the true/false positive/negative counts:  $TP/(TP+FP+FN)$ ).

**Results:** Results are shown across models in Table 1. Several points can be noted. First, as expected, increasing the percentage of supervised training data generally increases performance. A notable feature though is that in several cases, e.g. Model 3 (DA) image level, the maximum is achieved at 0.5 or 0.75 supervision, suggesting that the weaker-labeled MIL annotations may be providing a buffer to overfitting in these cases. Second, we note that in most cases, adding the MIL-annotations for the DA-trained models does appear to improve performance, as is particularly notable in models 2 and 3 at the image level, and model 1 at the patch and pixel levels. The cases where the opposite is true (model 1 at the image level, 2 and 3 at the patch level, and 3 at the pixel level) may be explained as follows: In each case, the level for which the model is optimizing the loss through DA does improve (i.e. image level for models 2 and 3, and patch level for model 1), but in other levels performance may decrease as it is not specifically targeted. In addition, in the case of model 1, there is a separate issue in that the extra annotations we added to the fully labeled LabelCracks set may have subtly different distribution properties, so there is an issue of transfer learning when testing at the image and patch levels for this model. Indeed, it performs at around chance ( $= 33.3\%$ ) at the image level. Nevertheless, it generalizes surprisingly at the patch level. Finally we note

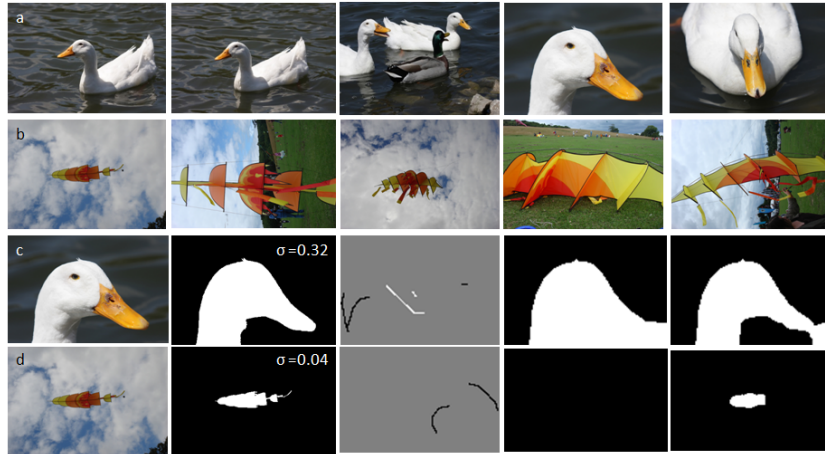
that the image level results clearly show the utility of using bag models with greater amounts of structure, where the two level model (model 3) outperforms both models 1 and 2. This confirms that model 3 is able to use the weak MIL annotations at both levels, and combine all the strong annotations, whose distribution characteristics are not necessarily identical, to improve the overall estimate of the image bag labels.

A number of further points about the results should be noted. While the actual numerical results may seem quite low, this is in part due to the metrics being used, as well as the difficulty of the task. Indeed, as shown in Figure 2c-d, the qualitative results are often reasonable, although the cracks are often poorly localized, which is partly a function of the underlying TextonBoost classifier. The intersection-union score is typically a harsh metric, and the pixel and patch level scores in Table 1 are comparable to those of the state-of-the-art in many classes of interest in for instance the PASCAL VOC challenge (see [11]). Table 2 gives precision/recall performance for model 1 (DA) at both high and low recall by varying a global weight on the positive unary potentials, where the high recall results correspond to the intersection-union scores in Table 1. We note that at low recall we are able to achieve precisions comparable or better than the 0.37-0.40 precisions reported in [22]. We use the stricter union-intersection metric in Table 1 so as to have a single measurement to compare across models, since the precision/recall regimes vary in different models according to the loss being optimized (models 2 and 3 impose lower recall rates at the patch and pixel levels, while model 1 favors higher recall at all levels). Finally, the 3-way categorization enforced by the quantization levels of  $\sigma$  may not reflect the total global information in the models, and a comparison of inference techniques (including continuous estimates of  $\sigma$ , see Secs. 4.2 and 4.3) is desirable. In this respect we also note that we have not made use of expected correlations between  $\sigma$  values on neighboring sections of road, and the same sections of road captured at different times. With these qualifications, the results generally show promise for combining various levels of annotations on this task using the models provided.

## 6 MIL-based Interactive Cosegmentation

For our second task, we consider interactive cosegmentation, using the iCoseg dataset of [2]. We use the 37 groups of 5 images provided, each containing a highly related set of images, where the task is to extract out the foreground objects in each group (example groups are shown in Figure 3a-b). We duplicate the machine-based scenario explored in [2], where ‘scribbles’ are automatically generated across all images of the group, totaling between 120 and 1200 pixels, mimicking user foreground/background annotations (see Figure 3c-d for examples). The task is to provide a binary segmentation of each image from these annotations.

We over-segment all images using the method of [1], and use the over-segmentations to augment the annotations by labeling segments containing scribbles by their mode scribble-class. In addition, we extract HOG, Color-HOG and Texton features from all images, using a similar quantization into visual words as in [10]. For our unary classifiers, we simply build histograms of these features for both foreground and background classes to learn a set of multinomial models (after adding a constant to each bin), and take the combined negative log-likelihood across features as the unary classifier re-



**Fig. 3.** Examples and results from the iCoseg dataset [2]. Shown are (a-b) two example subgroups of images from the dataset, and (c-d) examples of results achieved, showing the original image, ground truth and  $\sigma$  annotation (proportion of positives), scribble annotations, output of our baseline CRF, output of our model 2 (MIL-DA), using the provided  $\sigma$  annotations.

# pixels annotated	120	240	360	480	600	720	840	960	1080	1200
iCoseg (machine) [2]	61.3	73.5	79.9	84.0	86.6	88.3	89.6	91.1	91.8	92.7
Our baseline CRF (no MIL)	85.7	87.0	88.9	89.6	89.6	91.3	92.4	93.1	93.2	93.7
Model 2 (MIL-DA)	<b>90.1</b>	<b>91.7</b>	<b>92.5</b>	<b>92.5</b>	<b>93.3</b>	<b>94.2</b>	<b>95.1</b>	<b>95.0</b>	<b>95.0</b>	<b>95.3</b>

**Table 3.** Summary of MIL-based interactive cosegmentation results. We outperform [2] with our baseline CRF, and adding the  $\sigma$  annotations (proportion of positives) allows us to improve performance still further using our MIL model 2 with deterministic annealing.

sponse at each pixel. To create a base CRF, we also introduce contrast-sensitive Potts pairwise potentials as in Sec. 5. We compare the performance of this baseline CRF with model 2 as outlined earlier, which also has access to global  $\sigma$  values (proportion of positive labels), extracted from the ground truth. This mimics an interaction scenario where the user may provide not only scribbles on the images, but also an estimate of the size of the object relative to the image (which may be revised). We use the DA algorithm of Sec. 4.2 to alternate between sampling labels for the unannotated pixels in the group, and re-estimating the unary models, while maintaining the soft MIL constraints.

**Results:** Quantitative results are given in Table 2. Our baseline CRF already outperforms [2] (which also uses a pairwise CRF) at all levels, while adding the extra image-level annotations (as expected) gives a marked further increase. Qualitative comparisons of results before and after introduction of the global constraint are shown in Figure 3c-d. While to an extent a toy task, these results suggest that certain types of user interaction currently under-explored may be opened up by these models. Along similar lines, [23] explored a scenario for single image segmentation with an ‘area constraint’ (equivalent to  $\sigma$ ), and a similar DD algorithm used for inference. Comparison with our algorithm, which includes the extra DA optimization, is left to future work.

## 7 Discussion

We have introduced a number of techniques in this paper for using multiple-instance learning in settings where bags may be highly structured, drawing on recent dual decomposition approaches from the CRF literature, and also training techniques from the semi-supervised/MIL literature (deterministic annealing). As suggested, the use of more highly structured bags may be particularly beneficial in vision applications, and we provided two example tasks which demonstrate the potential of the models proposed.

A number of avenues for further research are suggested by this work. Most directly, we envisage such techniques to be ideally suited to general scene segmentation tasks if extended to a multilabel setting (as is straightforward), enabling us to utilize labelings of varying strengths (image-wise, pixel-wise, image regions etc.). The advantage of this approach over others (e.g. [7, 24]) is that CRF designs of arbitrary complexity can be incorporated (e.g. [10]), in which our DA algorithm requires only that we can perform approximate MAP inference. In addition, we have already mentioned the desirability of a closer comparison of different possible inference strategies in the models proposed at test time, and this might be extended to related sampling based approaches (e.g. Swendsen-Wang cuts) to compare the relative merits. Finally, we might also consider incorporating stronger methods of MAP-CRF learning such as structured max-margin approaches as components within the deterministic annealing framework.

**Acknowledgments.** This work was supported by Yotta DCL, and the IST Programme of the European Community, under the PASCAL2 Network of Excellence. P. H. S. Torr is in receipt of a Royal Society Wolfson Research Merit Award.

## References

1. P. Arbelaez, M. Maire, C. Fowlkes and J. Malik. From Contours to Regions: An Empirical Evaluation. *CVPR*, 2009.
2. D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. iCoseg: Interactive Cosegmentation with Intelligent Scribble Guidance. *CVPR*, 2010.
3. B. Babenko, M.H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. *CVPR*, 2009.
4. D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.
5. J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179-195, 1975.
6. P.V. Gehler, and O. Chapelle. Deterministic Annealing for Multiple-Instance Learning. *AISTATS*, 2007.
7. S. Gould, T. Gao, and D. Koller. Region-based Segmentation and Object Detection. *NIPS*, 2009.
8. P. Kohli, L. Ladicky and P.H.S. Torr. Robust Higher Order Potentials for Enforcing Label Consistency. *IJCV*, 2009.
9. N. Komodakis, N. Paragios, and G. Tziritas. MRF Optimization via Dual Decomposition: Message-passing Revisited. *ICCV*, 2005.
10. L. Ladicky, C. Russell, P. Kohli, and P.H.S. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. *ICCV*, 2009.
11. L. Ladicky, C. Russell, P. Kohli, P.H.S. Torr. Graph Cut Based Inference with Co-occurrence Statistics. *ECCV*, 2010.
12. C. Leistner, A. Saffari and H. Bischof. MIForests: Multiple-Instance Learning with Randomized Trees. *ECCV*, 2009.
13. C. Leistner, A. Saffari, J. Santner and H. Bischof. Semi-supervised Random Forests. *ICCV*, 2009.
14. K. Rose. Deterministic annealing, constrained clustering, and optimization. *IJCNN*, 1998.
15. C. Rother, V. Kolmogorov, T. Minka, and A. Blake. Cosegmentation of Image Pairs by Histogram Matching - Incorporating a Global Constraint into MRFs. *CVPR*, 2006.
16. A. Saffari, C. Leistner, M. Godec, J. Santner, and H. Bischof. On-line Random Forests. *OLCV*, 2009.
17. J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TexonBoost*: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. *ECCV*, 2006.
18. A. Vezhnevets, and J. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. *CVPR*, 2010.
19. S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. *ICCV*, 2009.
20. S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation Revisited: Models and Optimization. *ECCV*, 2010.
21. P. Viola, J.C. Platt, and C. Zang. Multiple instance boosting for object detection. *NIPS*, 2005.
22. J. Warrell, S. Prince, and P.H.S. Torr. StyP-Boost: A Bilinear Boosting Algorithm for Learning Style-Parameterized Classifiers. *BMVC*, 2010.
23. O.J. Woodford, C. Rother, and V. Kolmogorov. A Global Perspective on MAP inference for Low-Level Vision. *ICCV*, 2009.
24. Z. Zha, X. Hua, T. Mei, J. Wang, G. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. *CVPR*, 2008.
25. Z. Zhou, and M. Zang. Multiple-Instance Multi-Label Learning with application to Scene Classification. *NIPS*, 2006.
26. Z. Zhou, Y. Sun, and Y. Li. Multiple-Instance Learning by Treating Instances as Non-I.I.D. Samples. *ICML*, 2009.