Optimizing Average Precision using Weakly Supervised Data

Aseem Behl, Pritish Mohapatra, C.V. Jawahar, M Pawan Kumar

Optimizing Average Precision using Weakly Supervised Data

Aseem Behl, *Student Member, IEEE,* Pritish Mohapatra, C. V. Jawahar, *Member, IEEE,* and M. Pawan Kumar

Abstract—Many tasks in computer vision, such as action classification and object detection, require us to rank a set of samples according to their relevance to a particular visual category. The performance of such tasks is often measured in terms of the average precision (AP). Yet it is common practice to employ the support vector machine (SVM) classifier, which optimizes a surrogate 0-1 loss. The popularity of SVM can be attributed to its empirical performance. Specifically, in fully supervised settings, SVM tends to provide similar accuracy to AP-SVM, which directly optimizes an AP-based loss. However, we hypothesize that in the significantly more challenging and practically useful setting of weakly supervised learning, it becomes crucial to optimize the right accuracy measure. In order to test this hypothesis, we propose a novel latent AP-SVM that minimizes a carefully designed upper bound on the AP-based loss function over weakly supervised samples. Using publicly available datasets, we demonstrate the advantage of our approach over standard loss-based learning frameworks on three challenging problems: action classification, character recognition and object detection.

Index Terms—Weakly supervised learning, Average precision, Latent SVM

1 Introduction

Several problems in computer vision can be formulated as ranking tasks, that is, sorting a set of samples according to their relevance to a query. As a running example throughout this paper, we will consider the task of action classification, where the input is a set of images and the desired output is a ranking of the images. An accurate output is one that ranks an image containing a person performing an action of interest (such as 'jumping' or 'walking') higher than the images that do not contain a person performing the action of interest. Ranking is often reformulated as binary classification, for which there exist several learning frameworks. Among the most popular binary classifiers in computer vision is the support vector machine (SVM) [27]. During training, an SVM minimizes a convex regularized upper bound on the misclassification error over a fully supervised dataset, which consists of positive (that is, relevant) and negative (that is, non-relevant) samples. During testing, the samples are sorted in descending order of the scores provided by the SVM.

As the most commonly used accuracy measure for ranking tasks in computer vision is the average precision (AP) [6], and not misclassification error, the choice of SVM may appear surprising. The case for its use appears even weaker when we consider that there already exists a related learning framework (henceforth referred to as AP-SVM) that optimizes an AP-based loss function (henceforth referred to as the AP loss) [37]. However, a closer look at the empirical evidence reveals the reasoning behind this choice: SVM can be trained more efficiently, and provides comparable accuracy to AP-SVM.

The above observation suggests that we should continue to collect fully supervised datasets and use simple loss functions. If the supervision involves labelling each sample with its class (positive or negative), then this task does not appear to be daunting. However, recent research has shown that the key to achieving high classification and ranking accuracy is to provide additional annotations for each sample, which can guide the learner towards the correct output [3], [9], [18], [33], [35]. Going back to the example of action classification, it would be helpful to not only know the class information of each image but the exact location of the person in the image.

The need for complex additional annotations makes supervised learning impractical. To overcome this deficiency, researchers have started exploring weakly supervised learning [1], [5], [9], [14], [15], [19], [22], [21], [23], [29], [30], where the annotations of some or all the samples contain missing information. Not surprisingly, the convenience of using partial annotations comes at the cost of a significantly more challenging machine learning problem. Specifically, weakly supervised learning typically requires us to solve a non-convex optimization problem, which makes it prone to converge to a bad local minimum. Given the inherent difficulty of the problem, we hypothesize

- *A. Behl, P. Mohapatra and C.V. Jawahar are with the Centre for Visual Information Technology, IIIT Hyderabad, India, 500032. E-mail: see http://researchweb.iiit.ac.in/aseem.behl/*
- *M. Pawan Kumar is with Ecole Centrale Paris & INRIA Saclay*

This work is partially funded by the European Research Council under the European Communitys Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement number 259112, and the INRIA International Internship Programme.

that the choice of the loss function becomes crucial in such settings. In order to provide empirical evidence for our hypothesis, we propose a novel latent AP-SVM framework that models the missing additional annotations using latent variables.

Our formulation differs from the standard latent structured SVM (latent SSVM) [36] for general loss functions in three significant aspects. First, it uses a more intuitive two-step prediction criterion, where the first step consists of choosing the best latent variable for each sample and the second step consists of ranking the samples. This is in contrast to the latent SSVM formulation, which requires the joint optimization of the latent variables and the ranking. For example, in 'jumping' action classification, our latent AP-SVM formulation would first pick out the bounding box that is most likely to contain a 'jumping' person in each image, and then rank them. In contrast, the latent SSVM formulation would require us to simultaneously classify the samples as positive or negative, while picking out the best bounding box for the positive images (bounding box that is most likely to contain a 'jumping' person) and the worst bounding box for the negative images (bounding box that is least likely to contain a 'jumping' person). Second, using the above prediction criterion, the parameters of latent AP-SVM are learned by minimizing a tighter upper bound on the AP loss compared to latent SSVM. Third, unlike latent SSVM, latent AP-SVM lends itself to efficient optimization during learning, which is guaranteed to provide a local minimum or saddle point solution. While the first of the aforementioned differences makes our approach more intuitive, the latter two differences provide a sound theoretical justification for its superiority to latent SSVM. In order to demonstrate that the theoretical superiority also translates to better empirical results, we provide a thorough comparison of latent AP-SVM with the baseline methods for three challenging problems: action classification, character recognition and object detection. For the sake of clarity, we defer the details that are not essential for the understanding of the paper to the appendices. To facilitate the use of latent AP-SVM, we have made our code and data available online at http://cvit.iiit.ac.in/projects/lapsvm/.

2 Related Work

The popularity of the support vector machine (SVM) [27] can be gauged by its numerous applications in computer vision including, image classification [17], [32], action classification [3], [18], [35] and object detection [2], [28]. The main advantages of SVM are its well-understood connections to statistical learning theory [27] and the availability of efficient algorithms to learn its parameters [11], [12], [24].

One of the disadvantages of SVM is that it optimizes the 0-1 loss instead of the average precision (AP) over the training dataset. This disadvantage can be addressed by using the AP-SVM [37] to optimize an upper bound on the AP loss over the training samples. However, empirically, the performance of SVM is comparable to AP-SVM. Furthermore, SVM requires less training time compared to AP-SVM.

Another important disadvantage of SVM is its inability to handle missing information in the annotations. This problem is alleviated by latent SVM [9], which models missing annotations as latent variables. The 0-1 loss based latent SVM can be thought of as a special case of latent structured SVM (latent SSVM) [25], [36], which optimizes a general loss function. Latent SSVM has received considerable attention in the computer vision community [9], [15], [16], [29], [30], [31], [34], on tasks ranging from binary classification (such as object detection) to structured output prediction (such as semantic segmentation and indoor scene understanding). While it can be employed to optimize the AP loss, we will provide both theoretical and empirical arguments for the superiority of our novel latent AP-SVM formulation.

3 Preliminaries

Notation. We use a similar notation to [37]. The training dataset consists of n samples $\mathbf{X} = \{\mathbf{x}_i, i = 1, \cdots, n\}$ together with their class information. The indices for the positive and negative samples are denoted by $\mathcal{P}$ and $\mathcal{N}$ respectively. In other words, if $i \in \mathcal{P}$ and $j \in \mathcal{N}$ then $\mathbf{x}_i$ belongs to the positive class and $\mathbf{x}_j$ belongs to the negative class. Furthermore, for each sample $\mathbf{x}$, the dataset can also provide additional annotations, which we denote by $\mathbf{h}$. For example, in action classification each sample represents an image and the additional annotation $\mathbf{h}$ can represent the bounding box of the person in the image. To simplify the discussion in this section, we will assume that the additional annotations $\mathbf{h}$ are known for all samples. In the next section, we will describe the setting where the additional annotations are latent. We denote the set of all additional annotations for the positive and negative samples by $\mathbf{H}_P = \{\mathbf{h}_i, i \in \mathcal{P}\}$ and $\mathbf{H}_N = \{\mathbf{h}_j, j \in \mathcal{N}\}$ respectively.

The desired output is a ranking matrix $\mathbf{Y}$ of size $n \times n$, such that (i) $\mathbf{Y}_{ij} = 1$ if $\mathbf{x}_i$ is ranked higher than $\mathbf{x}_j$; (ii) $\mathbf{Y}_{ij} = -1$ if $\mathbf{x}_i$ is ranked lower than $\mathbf{x}_j$; and (iii) $\mathbf{Y}_{ij} = 0$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned the same rank. The ground-truth ranking matrix $\mathbf{Y}^*$ is defined as: (i) $\mathbf{Y}^*_{ij} = 1$ and $\mathbf{Y}^*_{ji} = -1$ for all $i \in \mathcal{P}$ and $j \in \mathcal{N}$; (ii) $\mathbf{Y}^*_{ii'} = 0$ and $\mathbf{Y}^*_{jj'} = 0$ for all $i, i' \in \mathcal{P}$ and $j, j' \in \mathcal{N}$.

AP Loss. Given a training dataset, our aim is to learn a ranking framework that provides a high AP measure. Let $\mathrm{AP}(\mathbf{Y}, \mathbf{Y}^*)$ denote the AP of the ranking matrix $\mathbf{Y}$ with respect to the true ranking $\mathbf{Y}^*$. The value of the $\mathrm{AP}(\cdot, \cdot)$ lies between 0 and 1, where 0 corresponds to a completely incorrect ranking $-\mathbf{Y}^*$ and 1 corresponds to the correct ranking $\mathbf{Y}^*$. In order to maximize the AP, we will minimize a

loss function defined as $\Delta(\mathbf{Y}, \mathbf{Y}^*) = 1 - \text{AP}(\mathbf{Y}, \mathbf{Y}^*)$.

Joint Feature Vector. For positive samples, the feature vector of the input $\mathbf{x}_i$ and additional annotation $\mathbf{h}_i$ is denoted by $\Phi_i(\mathbf{h}_i)$. Similarly, for negative samples, the feature vector of the input $\mathbf{x}_j$ and additional annotation $\mathbf{h}_j$ is denoted by $\Phi_j(\mathbf{h}_j)$. For example, in action classification, $\Phi_i(\mathbf{h}_i)$ can represent poselet [18] or bag-of-visual-words [3] features extracted from an image $\mathbf{x}_i$ using the pixels specified by the bounding box $\mathbf{h}_i$. Similar to [37], we specify a joint feature vector of the input $\mathbf{X}$, output $\mathbf{Y}$, and additional annotations $\{\mathbf{H}_P, \mathbf{H}_N\}$ as

$$\Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \mathbf{Y}_{ij} (\Phi_i(\mathbf{h}_i) - \Phi_j(\mathbf{h}_j)). \quad (1)$$

In other words, the joint feature vector is the scaled sum of the difference between the features of all pairs of samples where one sample is positive and the other is negative.

Parameters. The parameter vector of the ranking framework is denoted by $\mathbf{w}$, and is of the same size as the joint feature vector. Given the parameters $\mathbf{w}$, the ranking of an input $\mathbf{X}$ is defined as the one that maximizes the score, that is,

$$\mathbf{Y}_{opt} = \underset{\mathbf{Y}}{\text{argmax}}\, \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}), \quad (2)$$

where $\mathbf{H}$ is the set of all the given additional annotations. Yue *et al.* [37] showed that the above optimization can be performed efficiently by sorting the samples $(\mathbf{x}_k, \mathbf{h}_k)$ in descending order of the score $\mathbf{w}^\top \Phi_k(\mathbf{h}_k)$.

Supervised AP-SVM. Given the input $\mathbf{X}$, ranking matrix $\mathbf{Y}$, and additional annotations $\mathbf{H}_P$ and $\mathbf{H}_N$, we would like to learn the parameters $\mathbf{w}$ such that the AP loss over the training dataset is minimized. However, the AP loss is highly non-convex in $\mathbf{w}$, and minimizing it directly can result in a bad local minimum solution. To avoid this undesirable outcome, Yue *et al.* [37] proposed the AP-SVM formulation, which minimizes a regularized upper bound on the AP loss. Specifically, the model parameters are obtained by solving the following convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\xi, \quad (3) \\ \text{s.t.} \quad & \forall \mathbf{Y} : \{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})\} \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi. \end{aligned}$$

Intuitively, the above problem introduces a margin between the score of the correct ranking and all incorrect rankings. The desired margin is proportional to the difference in their AP values. The hyperparameter C controls the trade-off between the training error and the model complexity.

Problem (3) is specified over all possible rankings $\mathbf{Y}$, which is exponential in the number of training samples. Nonetheless, it can be solved efficiently using a cutting-plane method [37] described in Algorithm 1.

Algorithm 1 *Cutting plane algorithm for solving* AP-SVM.

Require: $\mathbf{X}, \mathbf{Y}^*, \epsilon$

1: Initialize the set of active constraints $\mathcal{W} \Leftarrow \emptyset$

2: $t \Leftarrow 0$

3: **repeat**

4: $t \Leftarrow t + 1$

5: Learn parameters $\mathbf{w_t}, \xi_\mathbf{t}$ by solving the following convex problem over the set of active constraints $\mathcal{W}$,

$$\begin{aligned} & \underset{\mathbf{w}, \xi}{\text{argmin}}\, \frac{1}{2}||\mathbf{w}||^2 + C\xi \\ & \text{s.t. } \forall \mathbf{Y} \in \mathcal{W} : \{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})\} \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi. \end{aligned}$$

6: Find the most violated constraint by solving the following problem,

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\text{argmax}} \{\Delta(\mathbf{Y}^*, \mathbf{Y}) + \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})\} \quad (4)$$

7: Add the most violated constraint to set of active constraints $\mathcal{W}$.

8: **until**

$$\begin{aligned} & \Delta(\mathbf{Y}^*, \hat{\mathbf{Y}}) + \{\mathbf{w_t}^\top \Psi(\mathbf{X}, \hat{\mathbf{Y}}, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & - \mathbf{w_t}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\})\} \leq \xi_t + \epsilon. \end{aligned}$$

Briefly, the algorithm starts by specifying no constraints (step 1 of Algorithm 1: $\mathcal{W}$ is initialized to the null set). At each iteration, it adds a single constraint, which corresponds to the most violated ranking (step 6 of Algorithm 1: solving problem (4)). Intuitively, problem (4) finds a ranking that differs significantly from the ground-truth ranking in terms of AP but has a high score. Having added the most violated constraint, the cutting plane algorithm updates the parameters by solving a convex quadratic program (step 5 of Algorithm 1). The algorithm stops once no constraint can be found that is violated by more than the desired precision ϵ.

The feasibility of the cutting plane algorithm relies on solving problem (4) efficiently. For fixed values of $\mathbf{H}_P$ and $\mathbf{H}_N$ (which is indeed the case for supervised AP-SVM), this can be achieved using the greedy algorithm of Yue *et al.* [37] outlined in Algorithm 2.

Briefly, it starts with the ground-truth ranking, where each positive sample is ranked higher than all the negative samples (step 2 of Algorithm 2). Next, it finds the ranking for each negative sample independently such that the loss-augmented score is maximized (step 3-6 of Algorithm 2). The overall complexity of the above algorithm is $O(n^2)$ (where n is the number of samples). It can be shown to provide the optimal ranking $\mathbf{Y}$, that is, one that maximizes

Algorithm 2 *Finding the most violated constraint for* AP *loss with known values of additional annotations.*

Require: $\mathbf{X}, \mathbf{Y}^*, \mathbf{H}_P, \mathbf{H}_N, \mathbf{w}$

1: Sort positive inputs $\mathbf{x}_i \in \mathcal{P}$ and negative inputs $\mathbf{x}_j \in \mathcal{N}$ in descending order of the score $\mathbf{w}^\top \Phi_i(\mathbf{h}_i)$ and $\mathbf{w}^\top \Phi_j(\mathbf{h}_j)$ respectively.
2: Initialize ranking $\mathbf{Y}$, s.t. all positive samples are ranked higher than negative samples.
3: **for** $j = 1 \rightarrow |\mathcal{N}|$ **do**
4: Find the position of j^{th} ranked negative which maximizes loss-augmented score, $\Delta(\mathbf{Y}^*, \mathbf{Y}) + \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})$.
5: Update $\mathbf{Y}$ with the best position of the j^{th} ranked negative.
6: **end for**
7: return $\mathbf{Y}$.

the AP loss augmented score. We refer the interested reader to [37] for details.

4 Optimizing Average Precision with Weak Supervision

The main deficiency of supervised learning is that it involves the onerous task of collecting detailed annotations for each training sample. Since detailed annotations are also very expensive, such an approach quickly becomes financially infeasible as the size of the datasets grow. In this work, we consider a more pragmatic setting where the additional annotations $\mathbf{H}_P$ and $\mathbf{H}_N$ are unknown. For example, consider 'jumping' action classification, where each input represents an image that can belong to the positive class or the negative class. In order to learn a ranking framework that can distinguish between 'jumping' and 'not jumping' images, we only require image-level annotations instead of the bounding box of the person in each image.

The convenience of not specifying additional annotations comes at the cost of a more complex machine learning problem. Specifically, we need to deal with two confounding factors: (i) since the best value of the additional annotation $\mathbf{h}_i$ for each positive sample $i \in \mathcal{P}$ is unknown, it needs to be imputed automatically; (ii) since a negative sample remains negative regardless of the value of the additional annotation $\mathbf{h}_j$, we need to consider all possible values of $\mathbf{H}_N$ during parameter estimation. In the 'jumping' action classification example, this implies that (i) we have to identify the bounding box of the jumping person in all the positive images, and (ii) ensure that the scores of the identified jumping person bounding boxes are higher than the scores of all possible bounding boxes in the negative images. In the following subsection, we describe how the standard latent SSVM attempts to resolve these confounding factors in order to optimize the AP loss. This will allow us to identify its shortcomings and correct them with our novel formulation in subsection 4.2.

4.1 Latent SSVM Formulation

Given an input $\mathbf{X}$, the prediction rule of a latent SSVM requires us to maximize the score jointly over the output $\mathbf{Y}$ and the additional annotations $\mathbf{H}$, that is,

$$(\mathbf{Y}_{opt}, \mathbf{H}_{opt}) = \underset{(\mathbf{Y},\mathbf{H})}{\operatorname{argmax}} \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}). \tag{5}$$

The parameters $\mathbf{w}$ of a latent SSVM are learned by minimizing a regularized upper bound on the training loss. Specifically, the parameters are obtained by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\xi, \\ \text{s.t.} \quad & \forall \mathbf{Y}, \mathbf{H} : \max_{\hat{\mathbf{H}}} \{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \hat{\mathbf{H}})\} \\ & -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}) \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi. \end{aligned} \tag{6}$$

Intuitively, the above problem introduces a margin between the maximum score corresponding to the ground-truth output and all other pairs of output and additional annotations. Similar to the supervised setting, the desired margin is proportional to the AP loss.

There are three main drawbacks of the standard latent SSVM formulation in the case of AP loss optimization. The first drawback is the prediction rule. While existing latent SVMs for binary loss [9] first obtain the score of each sample by maximising over the additional annotations and then ranking them according to their scores, this does not hold true for existing latent SVMs optimising a general structured loss function such as AP loss. This is specified by problem (5), which requires us to simultaneously label the samples as positive or negative (optimize over $\mathbf{Y}$) and find the highest scoring additional annotations for the positive samples and the lowest scoring additional annotations for the negative samples (optimize over $\mathbf{H}$) in order to maximize the score. This is in stark contrast to the prediction rule of existing weakly supervised binary classifiers, which first obtain the score of each sample by maximizing over the additional annotations (regardless of whether they will be labelled as positive or negative), and then ranking them according to their scores. For example, in action classification, we rank the images according to the highest scoring bounding box of a person in each image. In other words, we never compare the scores of particular choice of additional annotations with a different set of additional annotations. The second drawback is the learning formulation. This is specified by problem (6), which provides a very loose upper bound on the AP loss. The third drawback is the optimization. Specifically, to the best of our knowledge, the local optimum solution of problem (6) cannot be found efficiently due to the lack of an

appropriate cutting plane algorithm. For the details on the difficulty of optimization of latent SSVM, as well as an approximate algorithm used in our experiments, we refer the reader to Appendix C.

4.2 Latent AP-SVM Formulation

We now describe a novel latent AP-SVM formulation that overcomes the three drawbacks of the standard latent SSVM framework discussed in the previous section. Specifically, latent AP-SVM uses an intuitive prediction rule, provides a tighter upper bound on the AP loss, and lends itself to efficient optimization.

4.2.1 Intuitive Prediction

We use a two-step prediction rule. In the first step, we obtain the value of the additional annotations for each sample by maximizing the score, that is,

$$\mathbf{h}_{opt} = \operatorname*{argmax}_{\mathbf{h}} \mathbf{w}^\top \Phi(\mathbf{h}). \tag{7}$$

Next, we obtain the optimal ranking $\mathbf{Y}_{opt}$ for the additional annotations $\mathbf{H}_{opt}$, that is,

$$\mathbf{Y}_{opt} = \operatorname*{argmax}_{\mathbf{Y}} \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}_{opt}), \tag{8}$$

where $\mathbf{H}_{opt}$ is the set of all the additional annotations obtained by solving problem (7) for all samples. Similar to the supervised setting, the optimal ranking is computed by sorting the samples in descending order of their scores. Note that our prediction rule is the same as the ones used in conjunction with the current weakly supervised binary classifiers [1], [9].

4.2.2 Tighter Bound on the AP Loss

We learn the parameters of latent AP-SVM by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\xi, \\ \text{s.t.} \quad & \forall \mathbf{Y}, \mathbf{H}_N : \max_{\mathbf{H}_P} \{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})\} \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi. \end{aligned} \tag{9}$$

Intuitively, the above problem finds the best assignment of values for the additional annotations $\mathbf{H}_P$ of the positive samples such that the score for the correct ranking (which places all the positive samples above the negative samples) is higher than the score for an incorrect ranking, regardless of the choice of the additional annotations $\mathbf{H}_N$ of the negative samples.

It is worth noting the significant difference between the optimization corresponding to latent AP-SVM and the standard latent SSVM. Specifically, in the constraints of problem (6), the values of the additional annotations for a correct and incorrect ranking are independent of each other. In contrast, the constraints of problem (9) are specified using the same values of the additional annotations. The following proposition provides a sound theoretical justification for preferring problem (9) over problem (6).

Proposition 1. *The latent* AP-SVM *formulation provides a tighter upper bound on the* AP *loss compared to the standard latent* SSVM *formulation (proof in Appendix A).*

4.2.3 Efficient Optimization

The local minimum or saddle point solution of problem (9) can be obtained using the CCCP algorithm [38], as described in Algorithm 3. The algorithm involves two main steps. In the first step (step 3 of Algorithm 3), it imputes the best additional annotations $\mathbf{H}_P$ of the positive samples given the current estimate of the parameters. In the second step (step 4 of Algorithm 3), given the imputed values of $\mathbf{H}_P$, CCCP updates the parameters by solving the resulting convex optimization problem. We discuss both these steps in detail below.

Algorithm 3 *The* CCCP *algorithm for parameter estimation of latent* AP-SVM.

Require: $\mathbf{X}, \mathbf{Y}^*, \mathbf{w}_0, \epsilon$

1: $t \Leftarrow 0$
2: **repeat**
3: For the current set of parameters $\mathbf{w}_t$, obtain the value of the latent variables $\mathbf{H}_P^*$ that minimizes the objective function value of problem (9).
4: Update $\mathbf{w}_{t+1}$ by fixing the latent variables to $\mathbf{H}_P^*$ and solving the resulting convex problem.
5: $t \leftarrow t + 1$
6: **until** Objective function cannot be decreased below tolerance ϵ

Imputing the Additional Annotations. For a given parameter $\mathbf{w}$, we need to obtain the values of the additional annotations $\mathbf{H}_P$ for the positive samples such that it minimizes the objective function of problem (9). Since $\mathbf{w}$ is fixed, the first term of the objective function (that is, the squared ℓ_2 norm of $\mathbf{w}$) cannot be modified. Instead, we need to minimize the slack ξ, which is equivalent to solving the following problem:

$$\begin{aligned} \min_{\mathbf{H}_P} \max_{\mathbf{Y}, \mathbf{H}_N} \{ & \Delta(\mathbf{Y}^*, \mathbf{Y}) - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & + \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})\}. \end{aligned} \tag{10}$$

We refer to the above problem as output-consistent inference (since it fills in the missing information under the constraint that it is consistent with the output, that is, the optimal ranking). Although problem (10) contains $\mathbf{Y}$ and $\mathbf{H}_N$, the following proposition shows that it can be optimized easily with respect to $\mathbf{H}_P$.

Proposition 2. *Problem (10) can be solved efficiently by independently choosing the latent variable for each positive sample using the following criterion:*

$$\mathbf{h}_i^* = \operatorname*{argmax}_{\mathbf{h}_i} \mathbf{w}^\top \Phi_i(\mathbf{h}_i), \forall i \in \mathcal{P} \tag{11}$$

(proof in Appendix B).

Updating the Parameters. Given the imputed latent variables $\mathbf{H}_P^*$, the parameters are updated by solving the following convex problem:

$$\begin{aligned}&\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C\xi, \qquad (12)\\ &\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P^*, \mathbf{H}_N\}) - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P^*, \mathbf{H}_N\})\\ &\geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi, \forall \mathbf{Y}, \mathbf{H}_N.\end{aligned}$$

Similar to supervised AP-SVM, the above problem can be solved using a cutting plane algorithm. The computational feasibility of the cutting plane algorithm relies on being able to efficiently compute the most violated constraint. In our case, the most violated constraint is found by solving the following problem:

$$\begin{aligned}\hat{\mathbf{Y}}, \hat{\mathbf{H}}_N = \underset{\mathbf{Y}, \mathbf{H}_N}{\operatorname{argmax}} \{&\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P^*, \mathbf{H}_N\})\\ &- \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P^*, \mathbf{H}_N\}) + \Delta(\mathbf{Y}^*, \mathbf{Y})\}. \qquad (13)\end{aligned}$$

We refer to the above problem as loss-augmented inference (since it augments the score of the ranking with its AP loss). Note that, in contrast to supervised AP-SVM, we not only need to optimize over the ranking $\mathbf{Y}$, but also the variables $\mathbf{H}_N$. The following proposition allows us to perform the joint optimization efficiently. Here, we use the following shorthand:

$$s_j(\mathbf{h}_j) = \frac{1}{|\mathcal{P}||\mathcal{N}|} \mathbf{w}^\top \Phi_j(\mathbf{h}_j). \qquad (14)$$

Proposition 3. *Problem (13) can be solved by first maximizing over* $\mathbf{H}_N$ *using the criterion,* $\mathbf{h}_j = \operatorname{argmax}_{\mathbf{h}_j} s_j(\mathbf{h}_j)$. *(proof in Appendix B).*

Using Proposition 3, problem (13) can be solved in two steps. In the first step we maximize the loss-augmented score over $\mathbf{H}_N$ by maximizing the score of each negative sample independently. The second step is to maximize the loss-augmented score over $\mathbf{Y}$, which is achieved using the optimal greedy algorithm described in Algorithm 2.

5 Experiments

The previous section shows the theoretical benefit of latent AP-SVM over the standard latent SSVM formulation, namely that it minimizes a tighter upper bound on the AP loss and allows for efficient inference, while using an intuitive prediction rule. We now show that the theoretical benefits translate to improved empirical performance using three important and challenging problems in computer vision.

5.1 Action Classification

Dataset. We use the PASCAL VOC 2011 [6] action classification dataset, which consists of 4846 images depicting 10 action classes. The dataset is divided into two subsets: 2424 'trainval' images for which we are provided the bounding boxes of the persons in the image together with their action class; and 2422 'test' images for which we are only provided with the person bounding boxes.

Recall that our main hypothesis is that the challenging nature of weakly supervised learning makes it essential to use the right loss function during training. In order to test this hypothesis, we use the 'trainval' images to create five types of datasets that vary in their level of supervision. Specifically, each type of dataset provides the ground-truth additional annotations[1] for S percent of the positive and the negative samples, where $S \in \{0, 25, 50, 75, 100\}$. The additional annotations for the remaining $100 - S$ percent of the samples are treated as latent variables. The putative values of each latent variable are restricted to the top $T = 20$ boxes obtained by a standard person detector [8]. During testing, we use the learned parameters to rank the given person bounding boxes in the 'test' dataset. The performance is measured by submitting the scores of all the bounding boxes to the PASCAL VOC evaluation server.

Features. Given a bounding box $\mathbf{h}_i$ of the image $\mathbf{x}_i$, we use the standard poselet-based feature vector [18] to specify $\Phi_i(\mathbf{h}_i)$. It consists of 2400 activation scores of action-specific poselets and 4 object activation scores. In addition, we use the score of the person detector [8], which results in a 2405 dimensional feature vector.

Methods. We compare our latent AP-SVM formulation with the baseline latent SVM that is commonly used in computer vision. Latent SVM consists of two hyperparameters: (i) C, the trade-off between the regularization and the loss; and (ii) J, the relative weight of the positive samples. In order to further strengthen the baseline, we add robustness to outliers using a further hyperparameter c. Specifically, we prevent the classifier from considering the most confusing $c\%$ bounding boxes in the negative samples under the constraint that at least one bounding box is used per negative image. We obtain the best settings of the hyperparameters via a 5-fold cross validation, where the 'trainval' set is split into 1940 training images and 484 validation images. We consider the following putative values: $C \in \{10^{-3}, 10^{-2}, \ldots, 10^{4}\}$, $J \in \frac{|\mathcal{P}|+|\mathcal{N}|}{|\mathcal{P}|} \times \{10^{-4}, 10^{-3}, \ldots, 10^{1}\}$ and $c \in \{0, 0.1, \ldots, 0.9\}$ (note that, when $c = 0$, the resulting baseline is the standard latent SVM without robustness). In addition, we also compare the performance of our latent AP-SVM with latent SSVM. For the latent AP-SVM and latent SSVM, we only need to specify a single hyperparameter C, whose value is also obtained via 5-fold cross-validation. In order to mitigate the effects of initialization, we use 5 random seeds and choose the one that provides the minimum objective value for each method independently.

Complexity. The running time of weakly supervised learning algorithms is dominated by computation of

1. Additional annotation provided is the bounding-box obtained by a standard person detector overlapping most with the ground-truth bounding box in PASCAL VOC.

the most violated constraint. Empirically, we found that computation of most violated constraint in latent AP-SVM is around 5 times slower and 100 times faster compared to latent SVM and latent SSVM respectively. However, latent AP-SVM does not require an extra hyperparameter J (the relative weight of the positive samples). Hence, the time taken for crossvalidation by both latent SVM and latent AP-SVM is comparable.

Results. Figure 1 shows the best mean AP value over all 10 action classes obtained during 5-fold cross validation. Note that as the amount of supervision decreases, the gap between our method and the two baselines steadily increases. In the fully supervised setting, that is, $S = 100$, latent AP-SVM provides statistically significant improvements over latent SVM for only 4 out of 10 classes (using paired t-test with p-value less than 0.05), with an overall improvement of less than 3%. Note that, for fully supervised datasets, both latent AP-SVM and latent SSVM are equivalent to the AP-SVM, and hence provide the same results. However, in the more interesting weakly supervised setting, that is, $S = 0$, latent AP-SVM provides statistically significant improvements over latent SVM for 6 out of 10 classes, and an overall improvement of more than 5%. By cross-validating c, instead of choosing the default value of $c = 0$ (standard latent binary SVM), we improve the performance of the baseline by 2.4%. Latent AP-SVM also provides statistically significant improvements over latent SSVM for 7 out of 10 classes and an overall improvement of more than 4%.

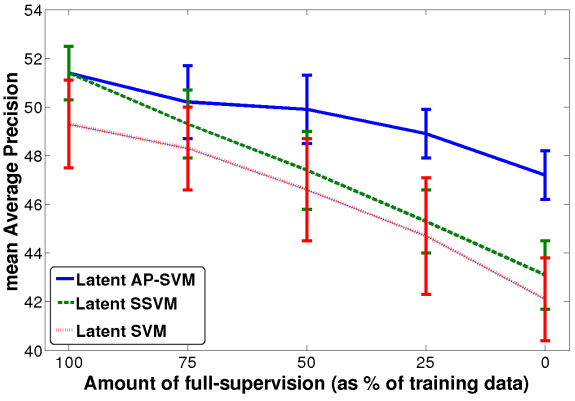

Fig. 1: *The best mean average precision over all 10 action classes obtained during 5-fold cross validation. The x-axis corresponds to the amount of supervision provided. The y axis corresponds to the mean average precision. As the amount of supervision decreases, the gap in the performance of latent AP-SVM and the baseline methods increases, thereby illustrating the importance of using the correct loss function and the correct learning formulation for weakly supervised learning.*

Table 1 shows the comparison of our latent AP-SVM with latent SVM and latent SSVM on the test set. Note that we use 5 different random seeds for each method. The hyperparameters are set using 5-fold cross-validation. Latent AP-SVM performs better than latent SVM for all 10 classes with significant increase in performance for 4 classes. Overall, we get an improvement of 5.1% on the test performance compared to latent SVM. Similarly, latent AP-SVM performs better than latent SSVM for 8 out of 10 classes. Overall, we get an improvement of 3.7% on the test performance compared to latent SSVM.

In order to analyse the better performance of latent AP-SVM in comparison to latent SSVM, we compute the AP loss and AP loss upper-bound values across iterations during the training of latent AP-SVM and latent SSVM. Figure 2 shows the corresponding plots. We obtain lower AP loss on the training set after completion of training of the latent AP-SVM compared to that of latent SSVM. This happens even though the value for the latent SSVM upper bound on the training set is higher for latent AP-SVM training compared to that for latent SSVM training. From this we can conclude that the lower AP loss obtained for latent AP-SVM is due to the latent AP-SVM upper bound being a better surrogate for AP loss compared to the latent SSVM upper bound.

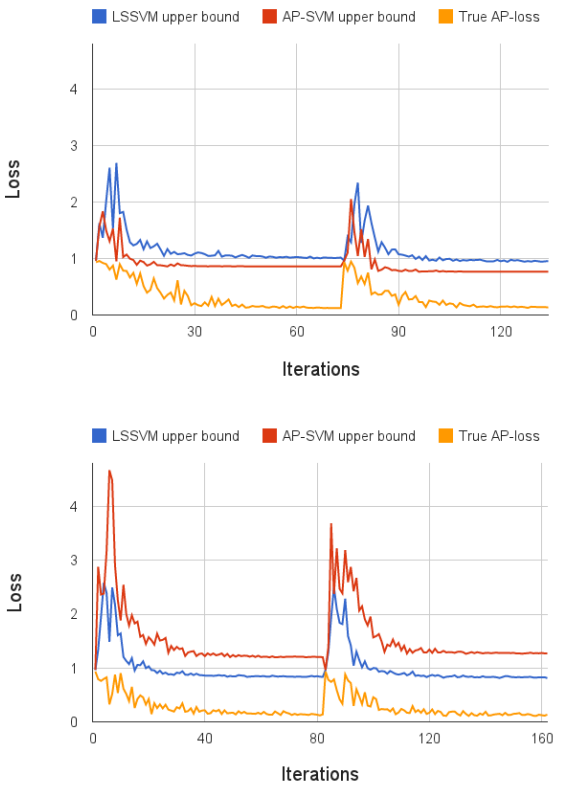

Fig. 2: *A comparison between the AP loss and AP loss upper-bound values computed across iterations during training of latent AP-SVM (top) and latent SSVM (bottom).*

5.2 Character Recognition in Natural Images

Dataset. We use the IIIT 5K-WORD [20] scene text dataset, which consists of 5000 cropped word images from scene texts and born-digital images, which are divided into 2000 'trainval' images and 3000 'test' images. Each image is annotated with the corresponding word, that is, a string where each character is an

Method	Jump	Use phone	Play instrument	Read	Ride bike	Ride horse	Run	Take photo	Use computer	Walk	Overall
Latent AP-SVM	45.7	30.5	34.0	21.1	75.5	74.9	76.0	15.7	24.6	47.5	44.6
Latent SSVM	37.6	26.5	33.9	22.5	71.2	66.7	66.8	17.4	21.9	44.8	40.9
Latent SVM	36.9	28.0	32.2	20.6	65.3	68.2	63.5	13.4	21.6	45.7	39.5

TABLE 1: *The average precision of latent* AP-SVM *and the baseline latent* SVM *and latent* SSVM *methods under weak supervision. The training is performed over the entire 'trainval' dataset with $S = 0$ using the best hyperparameters obtained during 5-fold cross-validation. The testing is performed on the 'test' dataset and evaluated on the* PASCAL VOC *server. The last column ('Overall') shows the mean average precision over all ten action classes.*

upper case letter ('A' to 'Z'), a lower case letter ('a' to 'z'), or a number ('0' to '9'). In addition, the dataset also provides the bounding boxes for each character of the word, which we discard during learning. Instead, we treat the bounding box of the characters as latent variables whose putative values are restricted to $T = 20$ boxes obtained by a standard character detector [2]. Using this dataset, we perform ranking for the 22 classes that contain at least 150 samples in the 'trainval' dataset.

Features. Given a character bounding box $\mathbf{h}_i$ of the word image $\mathbf{x}_i$, we use the histogram of oriented gradients (HOG) [2] features to specify $\Phi_i(\mathbf{h}_i)$. The HOG features are computed by resizing the bounding box to 48×48 pixels.

Methods. We compare our latent AP-SVM formulation with the baseline latent SVM and latent SSVM with AP loss. Similar to the action classification experiments, we set the hyperparameters of all the methods using 5-fold crossvalidation by splitting the 'trainval' dataset into 80%/20% folds. In order to avoid errors due to initialization, we use 3 different random seeds for each method and pick the one corresponding to the minimum objective value.

Results. Figure 3 and 5 show the best AP values for all the classes where the performance of latent AP-SVM is statistically different from that of latent SVM and latent SSVM respectively (using paired t-test with p-value less than 0.05). Latent SVM and latent SSVM provides statistically significant improvements over latent AP-SVM for only 1 class. In contrast, latent AP-SVM improves the performance for 4 and 3 classes over latent SVM and latent SSVM respectively. In terms of the mean AP value, latent AP-SVM provides an improvement of 3.2% and 2.7% over latent SVM and latent SSVM respectively.

Figure 4 and 6 show the AP values for the statistically significant characters on the 'test' set. Similar to the cross-validation results, latent AP-SVM outperforms latent SVM and and latent SSVM for 4 and 3 classes respectively. In terms of the mean AP value on the 'test' set, latent AP-SVM provides an improvement of 2.7% and 2.5% over latent SVM and latent SSVM respectively.

The detailed results over all 22 classes during cross validation and testing are provided in Figure 7 and Figure 8 respectively.

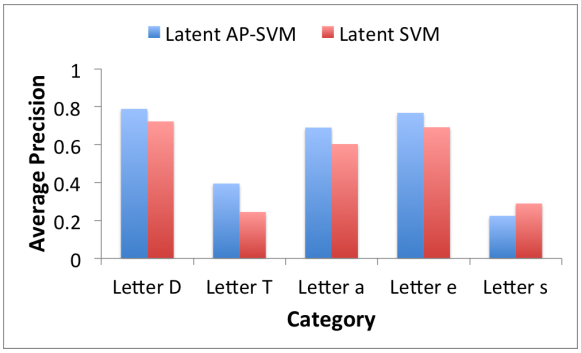

Fig. 3: *The best average precision values for the 5 statistically significant character classes obtained during 5-fold cross validation on the 'trainval' set of the* IIIT 5K-WORD *dataset. The x-axis corresponds to the characters. The y axis corresponds to the average precision. Latent* AP-SVM *provides statistically significant improvements over latent* SVM *for 4 out of the 5 characters.*

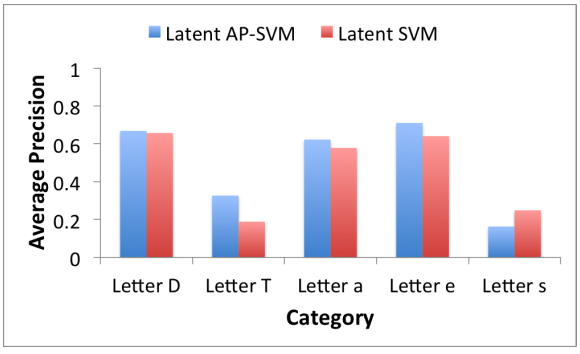

Fig. 4: *The average precision values for the 5 statistically significant characters obtained on the 'test' set of the* IIIT 5K-WORD *dataset. The x-axis corresponds to the character categories. The y axis corresponds to the average precision.*

5.3 Object Detection

Dataset. We use the PASCAL VOC 2007 [7] object detection dataset, which consists of a total of 9963 images. The dataset is divided into a 'trainval' set of 5011 images and a 'test' set of 4952 images. All the images are labeled to indicate the presence or absence of the instances of 20 different object categories. In addition, we are also provided with tight bounding boxes around the object, which we ignore during training and testing. Instead, we treat the location of the objects as a latent variable. In order to reduce the latent variable space, we use the selective-search

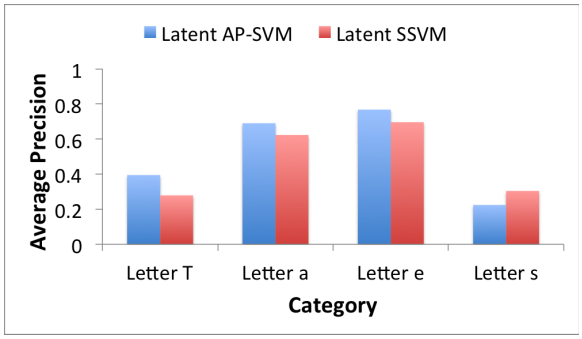

Fig. 5: *The best average precision values for the 4 statistically significant character classes obtained during 5-fold cross validation on the 'trainval' set of the* IIIT 5K-WORD *dataset. The x-axis corresponds to the characters. The y axis corresponds to the average precision. Latent* AP-SVM *provides statistically significant improvements over latent* SSVM *for 3 out of the 4 characters.*

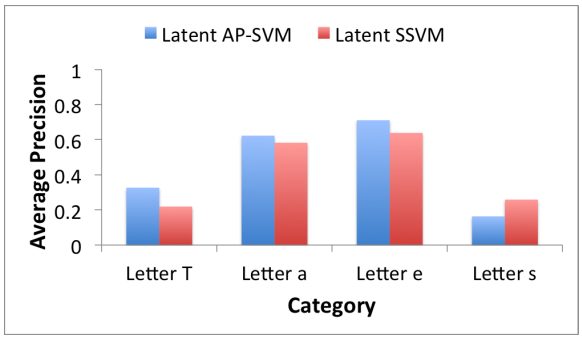

Fig. 6: *The average precision values for the 4 statistically significant characters obtained on the 'test' set of the* IIIT 5K-WORD *dataset. The x-axis corresponds to the character categories. The y axis corresponds to the average precision.*

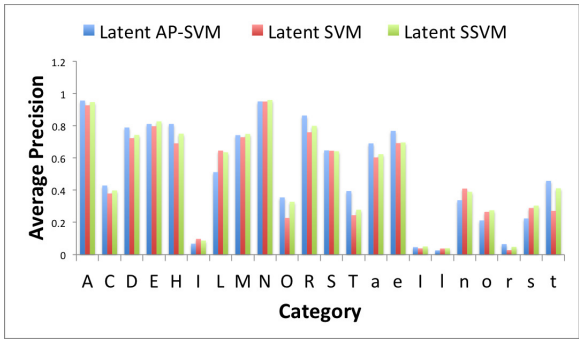

Fig. 7: *The best average precision values for all 22 character classes obtained during 5-fold cross validation on the 'trainval' set of the* IIIT 5K-WORD *dataset. The x-axis corresponds to the characters. The y axis corresponds to the average precision.*

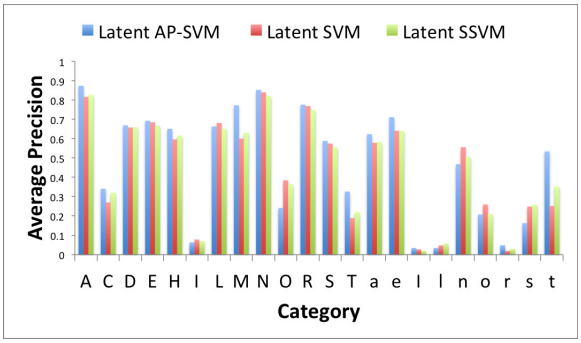

Fig. 8: *The average precision values for all 22 characters obtained on the 'test' set of the* IIIT 5K-WORD *dataset. The x-axis corresponds to the character categories. The y axis corresponds to the average precision.*

algorithm [26] in its fast mode, which generates an average of 2000 candidate windows per image.

Features. For each of the candidate windows, we use a feature representation that is extracted from a trained Convolutional Neural Network (CNN). Specifically, we pass the image as input to the CNN and use the activation vector of the penultimate layer of the CNN as the feature vector. Inspired by the work of Girshick *et al.* [10], we use the CNN that is trained on the ImageNet dataset [4], by rescaling each candidate window to a fixed size of 224×224. The length of the resulting feature vector is 4096.

Methods. We compare our latent AP-SVM formulation with latent SVM for 20 detection tasks, corresponding to the 20 object categories. Note that this experiment places high computational demands due to the size of the dataset (5011 'trainval' images), as well as the size of the latent space (2000 candidate windows per image). Hence, we were unable to run the expensive latent SSVM baseline. For both latent AP-SVM and latent SVM, we determine the value of the hyperparameters using 5-fold cross-validation. During testing, we evaluate each candidate window generated by selective search, and use non-maxima suppression to prune highly overlapping detections.

Results. We report the detection AP for all the 20 object categories obtained by latent SVM and latent AP-SVM. For all object categories other than 'bottle', latent AP-SVM does better than latent SVM on the test set. For 15 of the 20 object categories, we get statistically significant improvement with latent AP-SVM over latent SVM (using paired t-test with p-value less than 0.05). While latent AP-SVM gives an overall improvement of 7.12% compared to latent SVM, for 5 classes it gives an improvement of more than 10%. The bottom 2 classes with the least improvement obtained by latent AP-SVM, 'chair' and 'bottle' seem to be difficult object categories to detect, with detectors registering very low detection APs. The superior performance of latent AP-SVM compared to latent SVM can be partially attributed to the better localization

Object category	latent SVM	latent AP-SVM
Aeroplane	46.60	48.18
Bicycle	48.53	61.45
Bird	33.31	36.73
Boat	15.23	19.66
Bottle	6.10	1.01
Bus	37.01	49.51
Car	61.28	66.78
Cat	38.12	40.77
Chair	2.71	3.23
Cow	21.06	38.52
Dining-table	14.20	39.53
Dog	33.55	36.25
Horse	46.14	53.86
Motorbike	29.97	34.81
Person	29.58	30.41
Potted-plant	21.27	23.03
Sheep	11.65	32.20
Sofa	36.66	42.03
Train	29.71	37.10
TV-monitor	27.31	37.26

TABLE 2: Object category wise detection AP (%) on Pascal VOC2007 test set.

of objects by latent AP-SVM during training. Figure 9 shows this difference in localization performance of the two methods.

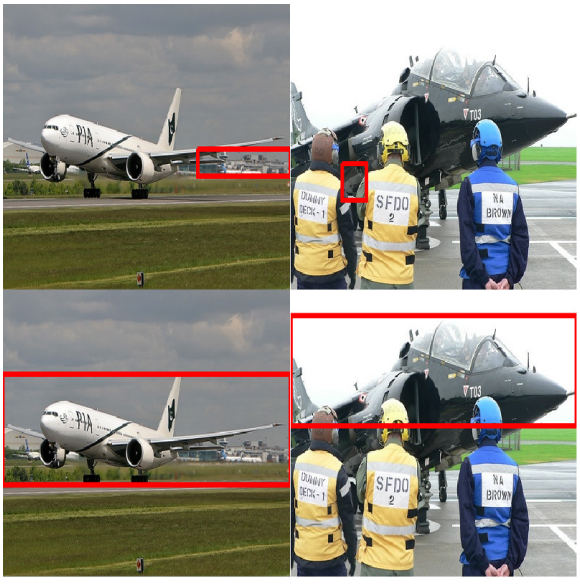

Fig. 9: *Localization results for some training set images. First row corresponds to latent* SVM *and the second row corresponds to latent* AP-SVM.

6 Discussion

We proposed a novel latent AP-SVM formulation that obtains accurate ranking by minimizing a carefully designed difference-of-convex upper bound on the AP loss. We showed the advantage of our approach over latent SVM and the standard latent SSVM for action classification, character recognition and object detection using standard, publicly available datasets.

An interesting direction of future research would be to extend the latent AP-SVM formulation to learn from images that have been labelled by noisy tags. This will allow us to exploit the large, freely available datasets provided by photo-sharing websites (for example, Flickr or Picasa). The large size of such datasets would also make it necessary to improve the efficiency of the CCCP algorithm for latent AP-SVM.

Appendix A
Latent AP-SVM as an Upper Bound of AP Loss

Before deriving the proof for our proposition that the latent AP-SVM formulation provides a tighter upper bound on the AP loss compared to the standard latent SSVM formulation, it would be helpful to prove the following lemma.

Lemma 1. *For a given value of the ranking matrix* $\mathbf{Y}$ *and the additional annotations of the negative samples* $\mathbf{H}_N$*, consider the following optimization problem:*

$$\mathbf{H}_P^* = \underset{\mathbf{H}_P}{\operatorname{argmin}}\, \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}), \tag{15}$$

where $\mathbf{Y}^*$ *is the optimal ranking matrix. The above optimization problem can be solved optimally as follows regardless of the choice of* $\mathbf{Y}$ *and* $\mathbf{H}_N$*:*

$$\mathbf{h}_i^* = \underset{\mathbf{h}_i}{\operatorname{argmax}}\, \mathbf{w}^\top \Phi_i(\mathbf{h}_i). \tag{16}$$

Proof: We use the following shorthand:

$$s_i(\mathbf{h}_i) = \frac{1}{|\mathcal{P}||\mathcal{N}|}\mathbf{w}^\top \Phi_j(\mathbf{h}_j),$$
$$s_j(\mathbf{h}_j) = \frac{1}{|\mathcal{P}||\mathcal{N}|}\mathbf{w}^\top \Phi_j(\mathbf{h}_j). \tag{17}$$

Using the above shorthand, we can rewrite the expression in the RHS of problem (15) as follows:

$$\begin{aligned} & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) - \\ & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ = & \sum_{i \in \mathcal{P}} s_i(\mathbf{h}_i) \left(\sum_{j \in \mathcal{N}} (Y_{ij} - Y_{ij}^*) \right) + \\ & \sum_{j \in \mathcal{N}} s_j(\mathbf{h}_j) \left(\sum_{i \in \mathcal{P}} (Y_{ij}^* - Y_{ij}) \right). \end{aligned} \tag{18}$$

By definition $\sum_{j \in \mathcal{N}} Y_{ij}^* \geq \sum_{j \in \mathcal{N}} Y_{ij}$ since $Y_{ij}^* = 1$ for all $i \in \mathcal{P}$ and $j \in \mathcal{N}$. Thus, the coefficient of the score term $s_i(\mathbf{h}_i)$ is negative for all positive samples. Therefore, in order to minimize problem (15) over all possible choices of $\mathbf{H}_P$, we should maximize the score $s_i(\mathbf{h}_i)$ for each positive sample. This is exactly the solution proposed in equation (16), which completes the proof. □

We are now ready to prove the following proposition.

Proposition 1. *The latent* AP-SVM *formulation provides a tighter upper bound on the* AP *loss compared to the standard latent* SSVM *formulation.*

Proof: We compare the optimization problems corresponding to the standard latent SSVM formulation and the latent AP-SVM formulation. The parameters of latent SSVM are estimated by solving the following problem:

$$\begin{aligned}&\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C\xi, \\ &\text{s.t. } \forall \mathbf{Y}, \hat{\mathbf{H}}_N, \mathbf{H}_P, \mathbf{H}_N : \\ &\max_{\hat{\mathbf{H}}_P} \{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \hat{\mathbf{H}}_P, \hat{\mathbf{H}}_N)\} \\ &-\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}_P, \mathbf{H}_N) \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi.\end{aligned} \tag{19}$$

The parameters of latent AP-SVM are estimated by solving the following problem:

$$\begin{aligned}\min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\xi, \\ \text{s.t.} \quad & \forall \mathbf{Y}, \mathbf{H}_N : \max_{\mathbf{H}_P} \{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})\} \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi.\end{aligned} \tag{20}$$

Note that the constraints specified in problem (20) are a subset of the constraints specified in problem (19). In other words, the feasible region of problem (20) is a superset of the feasible region of problem (19). It follows that the optimal objective function value of problem (20) is guaranteed to be less than or equal to the optimal objective function value of problem (19). In other words, minimizing problem (20) provides a quantity that is guaranteed to be less than or equal to the regularized upper bound on the AP loss that is optimized by the standard latent SSVM. Thus, in order to prove proposition 1 it is sufficient to show that problem (20) minimizes a valid regularized upper bound on the AP loss.

In order to show that problem (20) minimizes a regularized upper bound on the AP loss, we introduce the following notation. Given a set of parameters $\mathbf{w}$ for the latent AP-SVM formulation, we denote the predicted additional annotations by $H(\mathbf{w}) = (\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N(\mathbf{w}))$. Recall that, for each sample $\mathbf{x}_i$, the additional annotation $\mathbf{h}_i(\mathbf{w})$ is predicted using the following rule:

$$\mathbf{h}_i(\mathbf{w}) = \operatorname*{argmax}_{\mathbf{h}_i} \mathbf{w}^\top \phi_i(\mathbf{h}_i). \tag{21}$$

Similarly, we denote the predicted output (that is, the ranking matrix) as $\mathbf{Y}(\mathbf{w})$, which is obtained using the following rule:

$$\mathbf{Y}(\mathbf{w}) = \operatorname*{argmax}_{\mathbf{Y}} \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}(\mathbf{w})). \tag{22}$$

Using the above notation, the AP loss over the training dataset $\Delta(\mathbf{Y}^*, \mathbf{Y}(\mathbf{w}))$ can be upper bounded as follows:

$$\begin{aligned}& \Delta(\mathbf{Y}^*, \mathbf{Y}(\mathbf{w})) \\ = \quad & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}(\mathbf{w}), \{\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N(\mathbf{w})\}) \\ & +\Delta(\mathbf{Y}^*, \mathbf{Y}(\mathbf{w})) \\ & -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}(\mathbf{w}), \{\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N(\mathbf{w})\}) \qquad (23) \\ \leq \quad & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}(\mathbf{w}), \{\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N(\mathbf{w})\}) \\ & +\Delta(\mathbf{Y}^*, \mathbf{Y}(\mathbf{w})) \\ & -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N(\mathbf{w})\}) \qquad (24) \\ \leq \quad & \max_{\mathbf{Y}, \mathbf{H}_N} \left\{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N\}) \right. \\ & +\Delta(\mathbf{Y}^*, \mathbf{Y}) \\ & \left. -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P(\mathbf{w}), \mathbf{H}_N\})\right\}. \qquad (25)\end{aligned}$$

Expression (24) follows from the fact that the score for optimal ranking $\mathbf{Y}^*$ must be less than or equal to the score for the predicted ranking $\mathbf{Y}(\mathbf{w})$ (since the predicted ranking is obtained by maximizing the score as shown in equation (22)). Expression (25) follows from the fact that instead of using the prediction, we maximize over all possible rankings and additional annotations of the negative samples.

Note that expression (25) still contains the predicted value of the additional annotations of the positive samples. In order to further simplify the upper bound, we make use of lemma (1), which implies that for any value of $\mathbf{Y}$ and $\mathbf{H}_N$, the following holds true:

$$\begin{aligned}\mathbf{H}_P(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{H}_P} & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) - \\ & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}),\end{aligned} \tag{26}$$

since the predicted value of $\mathbf{H}_P(\mathbf{w})$ is exactly equal to that specified by equation (16). Therefore, it follows that

$$\begin{aligned}\mathbf{H}_P(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{H}_P} & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) - \\ & \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) + \Delta(\mathbf{Y}^*, \mathbf{Y}),\end{aligned} \tag{27}$$

since $\Delta(\mathbf{Y}^*, \mathbf{Y})$ is independent of $\mathbf{H}_P$. For example, consider the action classification task, where $\mathbf{Y}$ denotes the ranking of the images and $\mathbf{H}_P$ is the bounding boxes in the positive samples. Regardless of the choice of the positive samples, the loss is computed based on the ranking $\mathbf{Y}$ and not the bounding boxes $\mathbf{H}_P$. Using the fact that the above equation holds true for any choice of $\mathbf{Y}$ and $\mathbf{H}_N$, expression (25) can be simplified as

$$\begin{aligned}\min_{\mathbf{H}_P} \max_{\mathbf{Y}, \mathbf{H}_N} & \left\{\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) \right. \\ & +\Delta(\mathbf{Y}^*, \mathbf{Y}) \\ & \left. -\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\})\right\}.\end{aligned} \tag{28}$$

We note that this is exactly the value of the slack variable ξ in problem (20). This proves that problem (20) minimizes a valid regularized upper bound of the AP loss, which in turn proves the proposition. □

APPENDIX B
EFFICIENT INFERENCE FOR LEARNING LATENT AP-SVM

Proposition 2. *Problem (10) can be solved efficiently by independently choosing the latent variable for each positive sample using the following criterion:*

$$\mathbf{h}_i^* = \underset{\mathbf{h}_i}{\operatorname{argmax}}\, \mathbf{w}^\top \Phi_i(\mathbf{h}_i), \forall i \in \mathcal{P}. \tag{29}$$

Proof: For any given value of $\mathbf{Y}$ and $\mathbf{H}_N$, lemma (1) proves the following:

$$\mathbf{H}_P^* = \underset{\mathbf{H}_P}{\operatorname{argmin}}\, \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}), \tag{30}$$

where $\mathbf{H}_P^*$ is computed as suggested in the above proposition, that is, by maximizing the scores of positive samples independently over their choice of additional annotations. Therefore, it follows that

$$\mathbf{H}_P^* = \underset{\mathbf{H}_P}{\operatorname{argmin}}\, \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) + \Delta(\mathbf{Y}^*, \mathbf{Y}), \tag{31}$$

since $\Delta(\mathbf{Y}^*, \mathbf{Y})$ is independent of $\mathbf{H}_P$. For example, consider the action classification task, where $\mathbf{Y}$ denotes the ranking of the images and $\mathbf{H}_P$ is the bounding boxes in the positive samples. Regardless of the choice of the positive samples, the loss is computed based on the ranking $\mathbf{Y}$ and not the bounding boxes $\mathbf{H}_P$. Using the fact that the above equation holds true for any choice of $\mathbf{Y}$ and $\mathbf{H}_N$, it follows that

$$\begin{aligned}\mathbf{H}_P^* = \quad & \underset{\mathbf{H}_P}{\operatorname{argmin}} \max_{\mathbf{Y}, \mathbf{H}_N} \left\{ \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) \right. \\ & + \Delta(\mathbf{Y}^*, \mathbf{Y}) \\ & \left. - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\mathbf{H}_P, \mathbf{H}_N\}) \right\}. \end{aligned} \tag{32}$$

This proves the proposition. □

Proposition 3. *Problem (13) can be solved by first maximizing over $\mathbf{H}_N$ using the criterion, $\mathbf{h}_j = \operatorname{argmax}_{\mathbf{h}_j} s_j(\mathbf{h}_j)$.*

Proof: Problem (13) can be written as

$$\begin{aligned}\underset{\mathbf{Y}, \mathbf{H}_N}{\operatorname{argmax}} \{ & \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} (\mathbf{Y}_{ij}^* - \mathbf{Y}_{ij}) s_j(\mathbf{h}_j) \\ & + \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} (\mathbf{Y}_{ij} - \mathbf{Y}_{ij}^*) s_i(\mathbf{h}_i^*) + \Delta(\mathbf{Y}^*, \mathbf{Y}) \}. \end{aligned} \tag{33}$$

By definition, $\mathbf{Y}_{ij}^* = 1$ for all $i \in \mathcal{P}, j \in \mathcal{N}$. Thus, the coefficient of the score term $s_j(\mathbf{h}_j)$ is non-negative for all negative samples regardless of the value of $\mathbf{Y}$. Therefore, problem (13) can be maximized first over $\mathbf{H}_N$ by maximizing the score of each negative sample independently. This proves the proposition. □

APPENDIX C
OPTIMIZATION FOR LATENT SSVM

The parameters $\mathbf{w}$ of a latent SSVM are learned by minimizing a regularized upper bound on the training loss. Specifically, the parameters are obtained by solving the following optimization problem:

$$\begin{aligned} & \min_{\mathbf{w}} \frac{1}{2} ||\mathbf{w}||^2 + C\xi, \\ & \text{s.t. } \forall \mathbf{Y}, \hat{\mathbf{H}}_N, \mathbf{H}_P, \mathbf{H}_N : \\ & \max_{\hat{\mathbf{H}}_P} \{ \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \hat{\mathbf{H}}_P, \hat{\mathbf{H}}_N) \} \\ & - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \mathbf{H}_P, \mathbf{H}_N) \geq \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi. \end{aligned} \tag{34}$$

The above problem belongs to a special class of non-convex optimization problems called difference-of-convex programs. Specifically, its feasible region can be viewed as the difference of two convex sets. Thus, we can employ a concave-convex procedure (CCCP) [38] to obtain an approximate solution, as described in Algorithm 4. Briefly, the CCCP algorithm starts with an initial set of parameters and iterates over two steps until convergence. In the first step, it fixes the parameters and finds the best set of additional annotations $\hat{\mathbf{H}}_P^*$ of the positives samples for the ground-truth output $\mathbf{Y}^*$. In the second step, it fixes the additional annotations $\hat{\mathbf{H}}_P^*$ and updates the parameters by solving the resulting convex optimization problem.

In more detail, the imputation of the additional annotations for the positive samples requires us to minimize the slack ξ, which is carried out by solving the following problem:

$$\begin{aligned} \underset{\hat{\mathbf{H}}_P}{\operatorname{argmin}} \{ & \Delta(\mathbf{Y}^*, \mathbf{Y}) - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\hat{\mathbf{H}}_P, \hat{\mathbf{H}}_N\}) \\ & + \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) \}. \end{aligned} \tag{35}$$

As the loss term $\Delta(\mathbf{Y}^*, \mathbf{Y})$ and score for incorrect ranking $\Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\})$ are independent of $\hat{\mathbf{H}}_P$, problem (35) reduces to the following:

$$\hat{\mathbf{H}}_P^* = \underset{\hat{\mathbf{H}}_P}{\operatorname{argmax}}\, \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\hat{\mathbf{H}}_P, \hat{\mathbf{H}}_N\}). \tag{36}$$

We refer to the above problem as output-consistent inference (since it fills in the missing information under the constraint that it is consistent with the output, that is, the optimal ranking).

Given the imputed latent variables $\hat{\mathbf{H}}_P^*$, the parameter can be updated using the cutting plane algorithm. The computational feasibility of the cutting plane algorithm relies on being able to efficiently compute the most violated constraint. In our case, the most violated constraint is found by solving the following problem:

$$\begin{aligned} \hat{\mathbf{Y}}, \hat{\mathbf{H}}_N, \mathbf{H}_P, \mathbf{H}_N = & \\ & \underset{\mathbf{Y}, \hat{\mathbf{H}}_N, \mathbf{H}_P, \mathbf{H}_N}{\operatorname{argmax}} \{ \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) \\ & - \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\hat{\mathbf{H}}_P^*, \hat{\mathbf{H}}_N\}) + \Delta(\mathbf{Y}^*, \mathbf{Y}) \}. \end{aligned} \tag{37}$$

We refer to the above problem as loss-augmented inference (since it augments the score of the ranking with its AP loss).

Algorithm 4 *The* CCCP *algorithm for parameter estimation of latent* SSVM.

Require: $\mathbf{X}, \mathbf{Y}^*, \mathbf{w}_0, \epsilon$

$t \Leftarrow 0$

repeat

Impute the latent variables by solving problem (36).

Update $\mathbf{w}_{t+1}$ by fixing the latent variables to $\hat{\mathbf{H}}_P^*$ and solving the following convex problem,

$$\begin{aligned} \mathbf{w}_{t+1} = \text{argmin}_{\mathbf{w}} \tfrac{1}{2}||\mathbf{w}||^2 + C\xi, \\ \text{s.t. } \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\hat{\mathbf{H}}_P^*, \hat{\mathbf{H}}_N\}) - \\ \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}, \{\mathbf{H}_P, \mathbf{H}_N\}) \geq \\ \Delta(\mathbf{Y}^*, \mathbf{Y}) - \xi, \forall \mathbf{Y}, \mathbf{H}_N, \mathbf{H}_P, \mathbf{H}_N \end{aligned}$$

$t \leftarrow t + 1$

until Objective function cannot be decreased below tolerance ϵ

To summarize, CCCP requires us to solve two problems: problem (36), that is, output-consistent inference and problem (37), that is, loss-augmented inference. We now describe how both the problems can be solved efficiently to obtain an accurate set of parameters for latent SSVM.

Output-Consistent Inference. Problem (36) can be solved efficiently by independently choosing the additional annotation for each positive sample using the following criterion:

$$\mathbf{h}_i^* = \underset{\mathbf{h}_i}{\text{argmax}}\, s_i(\mathbf{h}_i). \tag{38}$$

Loss-Augmented Inference. To the best of our knowledge, the second step of the CCCP algorithm cannot be solved optimally due to the lack of an efficient algorithm that computes the most violating constraint for the corresponding cutting plane algorithm. We now provide details of an approximate optimization algorithm for problem (37) that was used in our experiments.

An approximate solution of problem (37) can be obtained in two steps. In the first step we minimize the positive score $\mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{Y}^*, \{\hat{\mathbf{H}}_P^*, \hat{\mathbf{H}}_N\})$ over $\hat{\mathbf{H}}_N$. Since, the positive score is independent of the loss, it decomposes over all negative samples and thus can be solved independently by choosing the additional annotation $\mathbf{h}_j$ for each negative sample using the following criterion:

$$\hat{\mathbf{h}}_j = \underset{\mathbf{h}_j}{\text{argmax}}\, s_j(\mathbf{h}_j). \tag{39}$$

The second step is to maximize the loss-augmented negative score. Note that loss-augmented inference requires us to not only obtain the ranking of the samples, but also the additional annotations of each sample. Let us denote the set of all the ranks that are occupied by a positive sample by $\mathcal{R}^+$. Similarly, the set of all the ranks that are occupied by a negative sample is denoted by $\mathcal{R}^-$. We first focus on the problem of finding the additional annotations of each sample for a given ranking, that is, for fixed sets $\mathcal{R}^+$ and $\mathcal{R}^-$. We will later describe how we approximately optimize over the rankings. Let us consider the task of obtaining the additional annotations for the positive samples. To solve this task, we construct a $|\mathcal{P}| \times |\mathcal{P}|$ matrix $\mathbf{S}_P$, such that $\mathbf{S}_P(i, a)$ is the score of assigning the sample $i \in \mathcal{P}$ to the rank $a \in \mathcal{R}^+$. Formally,

$$\mathbf{S}_P(i, a) = \max_{\mathbf{h}_i} c_a s_i(\mathbf{h}_i), \tag{40}$$

where

$$c_a = \max_{b \in \mathcal{R}^-} \delta(b < a) - \delta(b > a). \tag{41}$$

$\delta(.)$ returns the number of positive, negative example pairs satisfying the condition in the argument. The best assignment of additional annotations for the positive samples can be found efficiently by applying the dynamic Hungarian algorithm [13] to the matrix $\mathbf{S}_P$. Similarly, to solve the task of obtaining the additional annotations for the negative samples, we construct a $|\mathcal{N}| \times |\mathcal{N}|$ matrix $\mathbf{S}_N$, such that $\mathbf{S}_N(j, a)$ is the score of assigning the sample $j \in \mathcal{N}$ to the rank $a \in \mathcal{R}^-$, that is,

$$\mathbf{S}_N(j, a) = \max_{\mathbf{h}_j} c_a s_j(\mathbf{h}_j), \tag{42}$$

where

$$c_a = \max_{b \in \mathcal{R}^+} \delta(b > a) - \delta(b < a). \tag{43}$$

Once again, the best assignment of additional annotations for the negative samples can be found efficiently by applying the dynamic Hungarian algorithm [13] to the matrix $\mathbf{S}_N$.

The above argument shows that, for a given ranking, the optimal assignment of additional annotations for both the positive and the negative samples can be obtained in a computationally feasible manner. However, the optimization over the ranking itself poses a difficult problem. In order to obtain an approximate solution, we employ the following greedy strategy, which is a natural extension of the algorithm proposed by Yue *et al.* [37] for the supervised learning case. We start with the perfect ranking, where all the positive samples are ranked higher than all the negative samples. Next, we consider shifting the highest negative rank up the ranking while keeping all other ranks fixed. For each such ranking, we compute the value of the loss. Furthermore, we also compute the assignment of additional annotation to the ranks such that the score is maximized. We pick the ranking that maximizes the loss augmented score among all such rankings. Next, we consider shifting the second highest negative rank up the ranking, and find the ranking that maximizes the loss augmented

negative score. We continue this procedure until we have considered shifting the lowest negative rank up the ranking.

Non-Optimality of Loss-Augmented Inference The above algorithm can be shown to be non-optimal using a counter-example. For example, consider table 3 with two positive and negative samples each. Each sample has two possible values for additional annotation. The first row of the table represents the identifier of each sample. The second and third rows represent the score of the sample upon choosing first or second value of additional annotation respectively. The bottom row contains the label of each sample.

Sample ID	0	1	2	3
Score_0	4	7	1	0
Score_1	8	6	9	7
Label	-1	1	-1	1

TABLE 3: *Counter-example with two positive and negative samples each. Each sample has two possible values for additional annotation. The first row of the table represents the identifier of each sample. The second and third rows represent the score of the sample upon choosing first or second value of additional annotation respectively. The bottom row contains the label of each sample.*

We represent the ranking as a list of ordered pairs (a, b). The first entry of the ordered pair is the Sample-ID and second entry is the index of additional annotation selected:

The global optimum for the example in Table 3 is the following ranking,

$$(0, 1); (2, 1); (3, 0); (1, 1),$$

which yields a loss augmented score of is 22.6. In contrast, the greedy algorithm described above provides the following ranking,

$$(1, 0); (3, 1); (2, 0); (0, 0),$$

which yields a loss augmented score of 18.

REFERENCES

[1] M. Blaschko, A. Vedaldi, and A. Zisserman. Simultaneous object detection and ranking with weak supervision. In *NIPS*, 2010.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010.

[6] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[8] P. Felzenszwalb, R. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[11] T. Joachims. Making large-scale SVM learning practical. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods*. MIT Press, 1999.

[12] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009.

[13] G. A. Korsah, A. T. Stentz, and M. B. Dias. The dynamic hungarian algorithm for the assignment problem with changing costs. Technical report, Robotics Institute, 2007.

[14] M. P. Kumar, B. Packer, and D. Koller. Modeling latent variable uncertainty for loss-based learning. In *ICML*, 2012.

[15] M. P. Kumar, H. Turki, D. Preston, and D. Koller. Learning specific-class segmentation from diverse data. In *ICCV*, 2011.

[16] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *PAMI*, 2012.

[17] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and svm training. In *CVPR*, 2011.

[18] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[19] K. Miller, M. P. Kumar, B. Packer, D. Goodman, and D. Koller. Max-margin min-entropy models. In *AISTATS*, 2012.

[20] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.

[21] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[22] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 2012.

[23] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[24] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *ICML*, 2009.

[25] A. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *AISTATS*, 2005.

[26] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2), 2013.

[27] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[28] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[29] A. Vezhnevets, J. Buhmann, and V. Ferrari. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012.

[30] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. In *ECCV*, 2010.

[31] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

[32] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[33] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for object category recognition. In *CVPR*, 2008.

[34] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010.

[35] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[36] C.-N. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009.

[37] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR*, 2007.

[38] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 2003.