Weak to Strong Detector Learning for Simultaneous Classification and Localization

Xiaopeng Zhang, Hongkai Xiong, *Senior Member, IEEE*, Weiyao Lin, *Senior Member, IEEE*, and Qi Tian, *Fellow, IEEE*

***Abstract*— This paper aims at learning discriminative part detectors with only image-level labels. To this end, we need to develop effective technologies for both pattern mining and detection learning. Different from previous methods, which train part detectors in one step, we divide the detector learning process into two stages and formulate it as a weak to strong learning framework. In particular, we first learn exemplar detectors from the unaligned patterns and perform a detector-based spectral clustering to produce weak detectors that are only responsible for a few discriminative patterns. In this way, the weak detectors are able to offer right initial patterns for strong detector learning. Second, we learn strong detectors with patterns discovered from the weak detectors, which we formulate as a confidence-loss sparse multiple instance learning (cls-MIL) task. The cls-MIL considers the diversity of positive samples while avoiding drifting away from the well localized ones by assigning a confidence value to each positive sample. The responses of the learned detectors produce an effective mid-level image representation for both image classification and object localization. Experiments conducted on benchmark data sets well demonstrate the superiority of our method over existing approaches.**

***Index Terms*— Mid-level pattern mining, exemplar-SVM, multiple instance learning, image classification, object localization.**

I. Introduction

OBJECT parts that capture crucial characteristics of an image are important in a variety of object recognition and related applications. For instance, in Deformable Part Model (DPM) [13], an object is modeled as a set of deformable parts organized in a tree structure. In face verification [46], a part-based feature representation is learned under the supervision of face identities through a deep model. In fine-grained recognition [48], distinctive parts such as the head of birds are detected out to enable part-based representation. Nevertheless, obtaining informative parts usually requires object-level [13] or even part-level annotations [2], which is tedious and costly for large-scale datasets. Accordingly, it is desirable to discover these parts with minimal human supervision.

The success of Convolutional Neural Network (CNN) [19] has shed light on the possibility of automatically discovering object parts. It has been revealed that the CNN filters at different layers are sensitive to patches with varying receptive fields, *i.e.*, from low-level cues such as the edges and corners in earlier layers to semantically meaningful parts or even the whole object in deeper layers. From the point of detection, the outputs of the convolutional layers can be interpreted as detection scores of multiple detectors. However, the CNN is trained from ImageNet dataset where objects are typically centered in images, while in practical applications such as PASCAL VOC [12] and MS COCO [25] datasets, images usually contains multiple objects embedded in complex scenes. Due to the significant differences in image statistics, CNN pretrained on ImageNet performs poor on these datasets. An intuitive method is to perform network fine-tuning to enhance the representative ability of the pretrained CNN. However, fine-tuning methods can well handle images with well-aligned objects, but may get relatively inaccurate predictions over multi-label images where objects often suffer severe mis-alignment or partial occlusion. As a result, fine-tuning methods are short of automatically discovering discriminative patterns from images with complex scenes.

An alternative method of discovering informative parts automatically is to learn detectors *explicitly*, which we refer to weakly supervised detector learning. As shown in Fig. 1, the standard approach for detector learning requires initial patterns (object parts) for detector initialization, and an optimization strategy for detector learning. It is of vital importance to develop effective pattern mining and detector learning strategies to enhance the representative ability. However, learning part detectors automatically is a classical chicken-and-egg problem: without an accurate appearance model, examples of a part cannot be discovered, while an accurate appearance model cannot be learned without appropriate part exemplars. To solve this challenge, this paper proposes to learn detectors in a weak to strong framework. Since consistent patterns are hard to obtain, we first learn exemplar detectors from the unaligned

Manuscript received June 19, 2017; revised December 5, 2017 and December 26, 2017; accepted January 16, 2018. Date of publication January 25, 2018; date of current version February 5, 2019. This work was supported in part by the National Science Foundation of China under Grant 61425011, Grant 61720106001, Grant 61529101, Grant 61622112, Grant 61472234, Grant 61471235, and Grant 61429201. The work of H. Xiong was supported by the Program of Shanghai Academic Research Leader under Grant 17XD1401900. The work of W. Lin was supported by Shanghai The Belt and Road Young Scholar Exchange under Grant 17510740100. The work of Q. Tian was supported in part by ARO under Grant W911NF-15-1-0290 and in part by the Faculty Research Gift Awards by NEC Laboratories of America and Blippar. This paper was recommended by Associate Editor G. L. Foresti. *(Corresponding author: Hongkai Xiong.)*

X. Zhang, H. Xiong, and W. Lin are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zxphistory@sjtu.edu.cn; xionghongkai@sjtu.edu.cn; wylin@sjtu.edu.cn).

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2797923

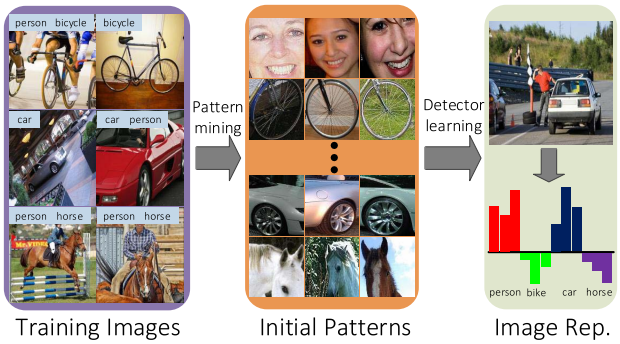


Fig. 1. Image representation based on the part responses. Given a set of training images which are only provided with image-level labels, our goal is to mine mid-level patterns (object parts) that capture crucial aspects of an object, and learn a set of part detectors for image representation.

patterns, producing weak detectors that are only responsible for a few discriminative patterns. Then, we learn strong detectors with patterns discovered from the weak detectors. The motivation is that we hope the weak detectors offer right initial patterns for strong detector learning, which makes the final detectors more generalized.

For weak detector learning, the **first contribution** of our method is the use of spectral clustering for consistent and discriminative pattern mining. To this end, a selection strategy is first utilized to sample discriminative patches of the corresponding category, followed by exemplar-SVM [26] detector training for each sampled patch, finally, these exemplar-SVM detectors are grouped via a spectral clustering strategy for pattern mining. Comparing with traditional clustering methods which are conducted on the original patches, the clustered detectors are able to focus on discriminative details, and discover the most representative patterns based on the detection scores. Furthermore, an entropy coverage criterion is utilized to measure the discriminativeness of each cluster, which enables us to greedily select clusters for detector learning.

As a **second contribution**, we develop a confidence loss sparse Multiple Instance Learning (cls-MIL) strategy for strong detector learning. Different from conventional MIL methods which represents each positive image with a single instance and treats each image equally important, cls-MIL represents each positive image as a sparse linear combination of its member instances, and considers the diversity of the positive images, while avoid drifting away the well localized ones by assigning a confidence value to each positive image. The responses of the learned detectors formulate an effective mid-level image representation for recognition. Another interesting finding is that different from most previous methods which treat image classification [40], [54] and object localization [3], [6], [23] separately, the proposed approach is able to effectively integrate the two tasks into a whole framework. Benefit from the pattern mining process, we are able to perform spectral clustering with reduced number and learn the corresponding discriminative part detectors accordingly. As a result, the detector responses by our approach are able to indicate the locations of the objects. Experiments conducted on benchmark datasets demonstrate the superiority of the proposed representation.

The rest of this paper is organized as follows. Sec. II reviews related works on weakly supervised detector learning. The details of our proposed detector learning method are elaborated in Sec. III. In Sec. IV, we apply the learned detectors for classification and localization. Experiments and discussions are given in Sec. V. Finally, Sec. VI concludes the paper.

II. Related Works

Over the past years there has been a lot of researches aiming at learning part models in an unsupervised or weakly supervised way. Most methods target at improving the two modules: pattern mining technologies for model initialization, and optimization strategies for detector learning. The learned part models offer a promising way for feature representation, which is beneficial for image recognition and other related applications. In the following, we organize the discussions related to part model learning with the above aspects.

A. Pattern Mining Methods

Since the ground truth annotations are not available in a weakly supervised paradigm, a number of strategies have been proposed to discover the discriminative patches for model initialization [52]. A simple method, taken in [29], [38], and [40], starts by randomly sampling a large pool of patches, and employs unsupervised clustering to generate initial patterns for detector learning. Such methods are clumsy and most returned clusters are with inhomogeneous appearances. Hence, many pattern mining technologies are developed to offer better initialization. Song *et al.* [39] formulate a constrained submodular algorithm to identify discriminative configurations of visual patterns. Wang *et al.* [43] propose to discover these latent parts via a probabilistic latent semantic analysis on the windows of positive samples and further employ these clusters as subcategories. Li *et al.* [24] combine the activations of CNN with the association rule mining technique to discover the representative mid-level patterns. Doersch *et al.* [10] formulate part discovery from the perspective of the well-known mean-shift algorithm to maximize the density ratio in the feature space. There is a special case in which we do not need to worry about exemplar alignment, *i.e.*, a training set consisting exactly of one part instance [26]. However, training detectors based on a single exemplar is with limited discriminative power, and the number of detectors scales with the training samples, which is tremendous for large-scale datasets.

Different from previous approaches which aim at grouping the original patches, this paper performs clustering in terms of the corresponding weak detectors, and makes use of the grouped detectors for pattern mining. In order to generate weak detectors, a selection strategy is first utilized to sample discriminative patches, and each patch is associated with a detector via exemplar-SVM training. Though a single exemplar-SVM detector is weak, a collection of such detectors offer relatively satisfactory localization capacity for pattern mining.

B. Optimization for Detector Learning

Based on these discovered patterns, most methods employ an iterative learning approach to refine the detectors.

Juneja *et al.* [18] employ an LDA accelerated version [16] of the exemplar-SVMs [26], which reduces the training cost substantially comparing with the standard SVM procedure that involves hard negative mining [13]. However, the detectors are trained with only one positive instance, which results in limited discriminative powers. Singh *et al.* [38] split the training set into two disjoint parts, and a part model is refined via an iterative procedure which alternates between clustering on one dataset and training discriminative classifiers on the other to avoid overfitting. Parizi *et al.* [29] propose a jointly training method which optimizes part models and class specific weights iteratively. Sun and Ponce [40] propose a latent SVM model to learn detectors, which tends to select the discriminative parts by enforcing group sparsity regularizer. However, these methods suffer from complex jointly optimization, *e.g.*, [29] takes over five days to train detectors on MIT Indoor-67 [32].

The majority of related works treat weakly supervised detector learning as a Multiple Instance Learning (MIL) task, in which labels are assigned to bags (sets of patterns), instead of individual patterns. MIL is originally introduced to solve a problem in biochemistry [9], and a variety of MIL algorithms have been developed over the years. Andrews *et al.* [1] present a new formulation of MIL as a max-margin SVM problem. Bunescu and Mooney [5] develop an MIL method which is particularly effective when the positive bags are sparse. Standard MIL proceeds by an iterative procedure which alternates between selecting the highest scoring detection per bag as positive instance and refining the detectors. However, such simplified setting is sensitive to initialization and easy to getting stuck in a local minimum.

This paper also formulates the weakly supervised detector learning as a MIL task. Different from previous works, we introduce a confidence loss term in MIL problem when determining the classifier hyperplane. The key insight is that due to the occlusion, illumination variation, and viewpoint variation, it is suboptimal to treat instances from different bags equally important for detector learning. The introduced confidence loss term measures the reliability of each instance for MIL learning. As a result, the detectors are able to focus on more confident samples and downweights those samples with lower reliability. Furthermore, a cross-validation strategy is introduced to avoid overfitting the initial patterns.

C. Mid-Level Image Representation

A collection of detector responses can be used as mid-level image representation. The paradigm is inspired by object bank [22], a pioneering work of using detector responses for image representation. The object bank represents an image as a scale-invariant response map of a large number of pre-trained generic object detectors. Following that, most technologies employ detection scores as image representation, and improve the performance by incorporating part responses [36], [38] or via multiple scale pooling [29], [40], [49].

Over the past years, CNN has become a powerful tool for image representation [50], [51]. Due to the domain mismatch between ImageNet and the target dataset, previous works attempt to enhance CNN representation by transferring learning [27], [14], [48]. However, these methods need substantial object/part annotations of the target dataset, which is tedious and impractical in real applications. Zhang *et. al* [48] propose an alternative method to fine tune the network via saliency-based sampling, which is free of the object annotations. Nevertheless, such method is only limited to datasets with relatively simple backgrounds (such as fine grained dataset [42]). It may obtain limited performance improvement on datasets with complex scenes such as Pascal VOC [12] datasets.

Our approach follows the pipeline of using detector responses as feature representation. Different from previous works which learn a large number of detectors for classification [18], [24], [29] or focus on learning a single detector for localization [6], [39], [43], this paper integrates classification and localization into a whole framework, *i.e.*, we not only solve the problem of whether an object is present in an image, but also focus on where the object (if exists) is. We find that it is possible to use only a few detectors for both classification and localization. Such an integrated framework is beneficial to close the gap between these two tasks.

III. Learning Part Detectors

In this section, we target at learning a collection of discriminative part detectors automatically for image representation. Our framework includes two steps of detector learning: weak detector learning for consistent pattern mining and strong detector learning for feature representation. The weak detector learning module first selects patches which are representative and discriminative, then a series of exemplar-SVM [26] detectors are trained from each selected patch. This is followed by a spectral clustering procedure which groups exemplar-SVM detectors for pattern mining. Furthermore, an entropy coverage criterion is proposed to measure the generalization ability of each cluster. The strong detector learning module formulates the optimization issue as a confidence loss sparse MIL (cls-MIL) task, which considers the reliability of each positive sample via alternating between mining new positive samples and retraining the part model. The whole framework of the proposed approach is illustrated in Fig. 2. In the following, we present the detailed design for each module.

A. Weak Detector Learning for Pattern Mining

Discovering groups of mid-level patterns that are discriminative and representative is crucial for detector learning. To solve this issue, we first introduce a sampling strategy which aims at selecting the discriminative patches, and propose a detector-based spectral clustering approach to mine consistent patterns. Furthermore, we present an entropy coverage criterion to measure the discriminativeness of each cluster, which enables us to greedily select detectors for image representation. These steps are described as follows:

1) Discriminative Patch Selection: For each image, there only exist a few patches that are discriminative for this category (the patches around the target object), and also many patches that are non-discriminative (the cluttered background). Here, we introduce a sampling strategy to first pick out those

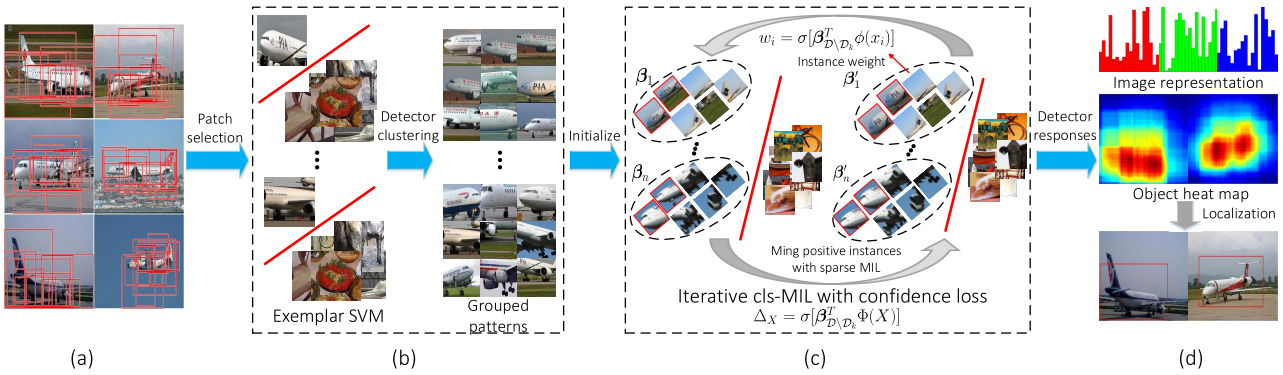


Fig. 2. The proposed weak to strong detector learning (WSDL) framework. Given a set of training images, we first learn a set of exemplar-SVM detectors from the selected patches, followed by detector clustering to discover patterns which are consistent and discriminative. The mined patterns are seeded for strong detector learning, which we formulate as a cls-MIL task. The detector responses are applied for both image classification and object localization. (a) Training images. (b) Weak detectors. (c) Strong detectors. (d) Classification & Localization.

discriminative and representative patches without object/part annotations. Specifically, given an image I with label $y \in \{-1, 1\}^C$, where $y_c = 1$ ($y_c = -1$) indicates the presence (absence) of an object class c, and C is the number of categories, we first generate M region proposals $X = \{x_1, ..., x_M\}$ with edge boxes [53], which probably includes the object of interest with a high recall. Denote the features extracted from region proposals as $\{\phi(I, x_1), ..., \phi(I, x_M)\}$, and the image level representation of I is obtained by sum pooling the features over M regions: $\phi(I) = \frac{1}{M}\sum_{m=1}^{M}\phi(I, x_m)$. For each category c, we select images containing object c (*i.e.*, $y_c = 1$) as the positive instances, and take all other images as the negative ones, and train a one-vs-rest SVM classifier based on the sum pooled features. Benefiting from the non-negativity of CNN features (we extract CNN features after ReLU layer) and the additivity of the linear classifier, we are able to select the patches which contribute the most to the classification score. These patches are denoted as discriminative patches, while other non-discriminative patches usually have low classification scores and are filtered out during the selection process. Specifically, given one category c and its classification model $\boldsymbol{\beta}_c$, the discriminative patch set X_{D_c} of an image I corresponding to category c is denoted as

$$X_{D_c} = \{x_i \mid \boldsymbol{\beta}_c^T \phi(I, x_i) > \tau\},$$

where τ denotes the threshold which enforces selecting the discriminative patches for classification. In all experiments we set τ to 1, *i.e.*, each selected patch strictly follows the SVM classification hyperplane.

In order to avoid the classifier overfitting the training set $\mathcal{D}$, we equally divide $\mathcal{D}$ into K disjoint and complementary subsets $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_K\}$. The classifier is trained on $K-1$ subsets and validated on the rest one. Fig. 3 illustrates some discriminative patches selected on Pascal VOC 2007 dataset. It can be seen that the selected patches probably locate around the object of interest, and skip other irrelevant backgrounds.

2) Detector-Based Clustering: The patch selection process usually generates tens of thousands of patterns per category, and most of them are highly correlated, *e.g.*, there exists some patches describing the head of dogs, and some

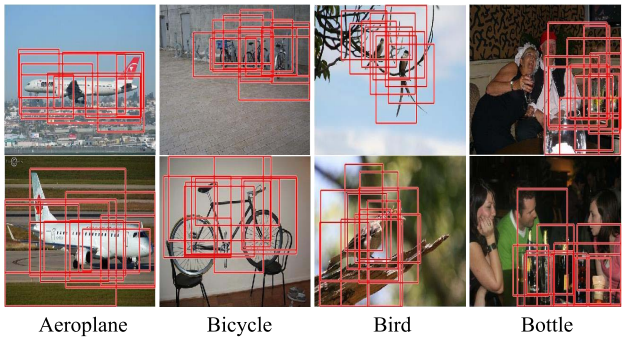


Fig. 3. Examples of the selected discriminative patches (shown in red bounding boxes) on Pascal VOC 2007 [12].

others describing the legs of dogs. It is necessary to cluster these patterns into smaller and representative groups for detector initialization. To this end, an alternative method is to employ some form of unsupervised clustering such as k-means [29], [38], [40]. However, k-means behaves poorly in high dimensional space due to the disturbance of unimportant entries, and often produces clustered instances which are in no way visually similar. Instead of clustering the original patches, this paper proposes a detector-based spectral clustering strategy, which discovers similar patterns via the grouped detectors.

Inspired from exemplar-SVMs [26], we start learning detectors from only one instance, which avoids the issue of exemplar misalignment. For each selected discriminative patch x from Eq. (1), we train an exemplar-SVM detector $\boldsymbol{d}$. Since the negative samples are too large, standard hard mining method [13] is quite expensive. We use instead Linear Discriminant Analysis (LDA) [16] to train a detector, which is an accelerated version of the exemplar-SVMs. Specifically, the detector template $\boldsymbol{d}$ is learned simply by:

$$\boldsymbol{d} = \Sigma^{-1}(\overline{\boldsymbol{x}}_p - \boldsymbol{\mu}_0), \tag{1}$$

where $\overline{\boldsymbol{x}}_p$ is the mean features of the positive examples, $\boldsymbol{\mu}_0$ denotes the mean of the features in the whole dataset, and $\sum$ is the corresponding covariance matrix. Since each exemplar-SVM detector is supposed to fire only on visually similar examples, we cannot expect it to generalize too

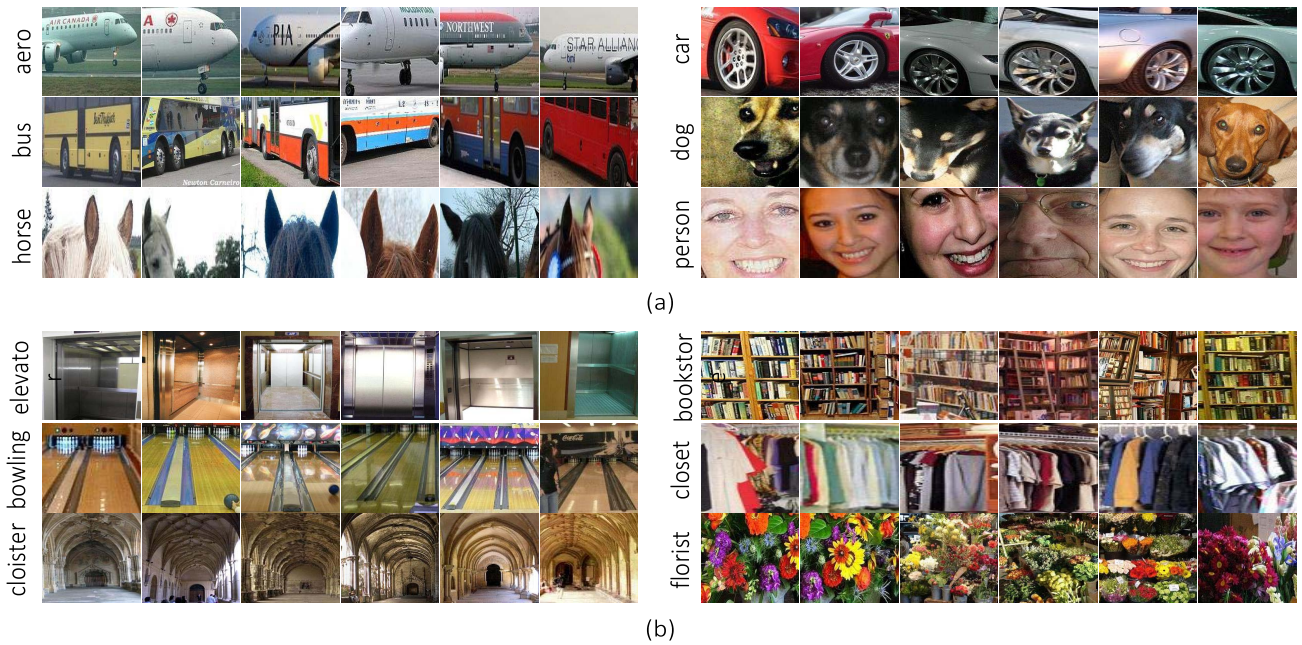


Fig. 4. Examples of the discovered mid-level patterns with clustered weak detectors on (a) Pascal VOC 2007 [12] and (b) MIT Indoor-67 [32].

much. To solve this issue, we follow an iterative procedure [18] which adds new positive samples to enhance the exemplar detectors. At each round, we run the current detector on all other images with the same label, and retrain it by augmenting the positives with the top scored patches. The idea behind this process is using detection score as a similarity metric, which emphasizes the distinctive details and suppresses those irrelevant ones.

Using exemplar-SVMs, each selected patch is associated with a detector. The key insight of the proposed strategy is that instead of clustering the original patches, we group the corresponding detectors. Specifically, given n_c exemplar detectors $\{\boldsymbol{d}_i\}_{i=1}^{n_c}$ trained from one certain class c, we perform spectral clustering on the similarity matrix S generated from the detectors $\{\boldsymbol{d}_i\}_{i=1}^{n_c}$, and obtain $\mathcal{K}$ clusters $\{C_k\}_{k=1}^{\mathcal{K}}$ corresponding to category c, where $S(i, j)$ denotes the cosine similarity of $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$. Thus, detectors sharing similar response distributions are grouped together. Inspired by the boosting strategy [41], each cluster acts as an integrated detector to discover similar patterns, *i.e.,* the detection score of a patch x with respect to a cluster C_k is denoted as

$$s(C_k|x) = \sum_{\boldsymbol{d}_k \subseteq C_k} \boldsymbol{d}_k^T \phi(x). \tag{2}$$

For simplicity, each category shares the same number of clusters, and as a result, we obtain $\mathcal{K}C$ detectors in total. As an illustration, Fig. 4 shows some examples of the discovered patterns using the clustered detectors. It can be shown that although a single detector is weak, a collection of such detectors offer satisfactory localization capacity. Another advantage of the detector-based pattern mining method is that we can select the most discriminative and representative patterns according to the top responses of the grouped detectors.

3) Entropy Coverage: The detector-based clustering generates a series of clusters with varying discriminative capacities. The notation of discriminative clusters is that the detectors within a cluster should be trained from as many images as possible. Such clusters include detectors corresponding to repeated patterns among varying images. We propose an entropy coverage criterion to measure the discriminativeness of each cluster. Given N images $\{I_i\}_{i=1}^{N}$ belonging to the same class and the corresponding $\mathcal{K}$ clustered detectors $\{C_k\}_{k=1}^{\mathcal{K}}$, the entropy coverage of cluster C_k is defined as:

$$\mathcal{H}(C_k) = -\sum_{i=1}^{N} p(I_i|C_k) \log_2 p(I_i|C_k), \tag{3}$$

where $p(I_i|C_k)$ denotes the probability of detectors coming from image I_i. $\mathcal{H}(C_k)$ is large if the clustered detectors within C_k are trained from diverse images, and reaches its maximum when the detectors are trained from patterns with equal distribution. The larger $\mathcal{H}(C_k)$ is, the more frequent patterns the detectors in C_k could find. Such an entropy coverage criterion enables us to greedily select clusters for detector initialization, while not worrying about choosing appropriate number of clusters. In the experimental section, we would find that the optimal number of clusters is determined by the classification performance.

B. Strong Detector Learning via cls-MIL

The clustered exemplar detectors are rather weak, since each exemplar detector is quite specific to its exemplar, and only performs well on visually similar examples. We cannot expect these weak detectors to generalize well on all examples that have the same part. In order to enhance the detectors, we develop a confidence loss sparse Multiple Instance Learning (cls-MIL) strategy for strong detector learning. Compared with standard MIL, cls-MIL makes three improvements:

• First, the standard MIL mines a single instance for positive image representation, which is highly dependent on the detectors and is not robust. To tackle this issue, we introduce a pooling strategy which represents the positive image as a weighted linear combination of its mined positive instances.

• Second, the standard MIL treats each mined instance equally important, which is often not the case. Due to occlusion, illumination variation, and viewpoint variation, the same part from different images suffers varying confidence of being positive. Comparatively, we introduce a regularized term to measure the confidence of each bag containing the positive samples, which considers the diversity of the positive samples while avoiding drifting away from the well localized ones.

• Third, the standard MIL alternatively selects the highest detection per image as the positive instance and refines the detection model within the same dataset. This would easily make the detectors latch on to the initial patches they are trained from and prefer them during the following iterations. In contrast, we introduce a multi-fold cross-validation to avoid overfitting the initial training samples in cls-MIL.

1) Problem Formulation: To use MIL for detector learning, each image is considered as a bag, and the patches within it as instances. Given a set of training images, we treat images of one particular category as positive bags, and the rest images as negative bags. For each image, if it is labeled as positive, then at least one patch within it should be treated as a positive instance; when it is labeled as negative, then all patches within it should be treated as negative instances. Specifically, let $\mathcal{X}$ be the set of bags used for training, which consists of a set of positive bags $\mathcal{X}_p$ and negative bags $\mathcal{X}_n$, *i.e.*, $\mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_n$. Denote X as a bag of images, and $\tilde{\mathcal{X}}_p = \{x | x \in X \subseteq \mathcal{X}_p\}$ and $\tilde{\mathcal{X}}_n = \{x | x \in X \subseteq \mathcal{X}_n\}$ as the set of instances from positive bags and negative bags, respectively. For any instance $x \in X$ from a bag $X \subseteq \mathcal{X}$, let $\phi(x)$ be the feature representation of x (for brevity, we include the bias term into feature representation). The cls-MIL problem can be formulated as solving the following objective:

$$\begin{aligned} \min\ & \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{X\subseteq\mathcal{X}_p}\Delta_X\xi_X + C\sum_{x\in\tilde{\mathcal{X}}_n}\xi_x \\ s.t.\ & \boldsymbol{\beta}^T\Phi(X) \geq 1-\xi_X, \quad \forall X \subseteq \mathcal{X}_p, \\ & \boldsymbol{\beta}^T\phi(x) \leq -1+\xi_x, \quad \forall x \in \tilde{\mathcal{X}}_n, \\ & \xi_X \geq 0, \quad \xi_x \geq 0, \quad \forall X \subseteq \mathcal{X}_p, \forall x \in \tilde{\mathcal{X}}_n, \end{aligned} \tag{4}$$

where $\Phi(X)$ is the feature representation of positive bag X, Δ_X is the latent variable which measures the positiveness of a bag $X \subseteq \mathcal{X}_p$, ξ_X, ξ_x are the slack variables, and C is the control parameter of the loss term. For positive bag representation, $\Phi(X)$ is denoted as the weighted linear combination of its mined top scored positive instances,

$$\Phi(X) = \frac{\sum_{m\in s(X)} w_m\phi(x_m)}{\sum_{m\in s(X)} w_m}, \tag{5}$$

where w_m is a weight assigned to each instance and is determined by previous round detector scores, and $s(X)$ is an indicator which denotes the selected patterns as the positive "witness" in a positive bag X. In our experiments, only a few instances per positive bag are selected.

2) Optimization: The cls-MIL leads to a non-convex optimization problem due to the introduction of implicit feature representation $\Phi(X)$ for the positive bags and the latent confidence variables Δ_X. However, this problem is semi-convex since optimization problem becomes convex once these latent variables are fixed. In the following, we solve Eq. (4) via an iterative procedure which alternates between fixing the latent variables and optimizing the detectors. In order to avoid focusing on the initial positive samples, the optimization procedure is processed via cross-validation. Specifically, the training set $\mathcal{D}$ is equally divided into K disjoint and complementary subsets $\{\mathcal{D}_1, ..., \mathcal{D}_K\}$. Starting from the patterns discovered by the clustered exemplar-SVM detectors, the detector $\boldsymbol{\beta}$ is optimized via iteratively **Updating** the latent variables and **Optimizing** Eq. (4). In the **Updating** step, the latent variables in $\mathcal{D}_k$ are determined by detectors $\boldsymbol{\beta}_{\mathcal{D}\setminus\mathcal{D}_k}$ trained on $\{\mathcal{D} \setminus \mathcal{D}_k\}$, *i.e.*, each instance weight w_m of $\Phi(X)$ is updated by: $w_m = \sigma[\boldsymbol{\beta}^T_{\mathcal{D}\setminus\mathcal{D}_k}\phi(x_m)]$, and the confidence loss term $\Delta_X = \sigma[\boldsymbol{\beta}^T_{\mathcal{D}\setminus\mathcal{D}_k}\Phi(X)]$, where σ is a sigmoid function which maps the value into the range of $(0, 1)$. In the **Optimizing** step, the detector is optimized according to the updated latent variables via hard negative mining [13].

Proposition: The solution $\boldsymbol{\beta}$ of Eq. (4) is a linear combination of the positive instances $\phi(X)$ and the negative instances $\phi(x)$, i.e., $\boldsymbol{\beta} = \sum_{X\subseteq\mathcal{X}_p}\alpha_X\phi(X) + \sum_{x\in\tilde{\mathcal{X}}_n}\alpha_x\phi(x)$, *where the coefficients α_X and α_x are bounded by: $0 \leq \alpha_X \leq C\Delta_X$, $0 \leq \alpha_x \leq C$, respectively.*

Proof: The constrained minimization problem in Eq. (4) can be solved with a classical Lagrangian method. The Lagrangian operator can be represented as:

$$\begin{aligned} \mathcal{L} = & \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{X\subseteq\mathcal{X}_p}\Delta_X\xi_X + C\sum_{x\in\tilde{\mathcal{X}}_n}\xi_x \\ & + \alpha_x(\boldsymbol{\beta}^T\phi(x) + 1 - \xi_x) - \sum_{x\in\tilde{\mathcal{X}}_n}\gamma_x\xi_x \\ & - \alpha_X(\boldsymbol{\beta}^T\Phi(X) - 1 + \xi_X) - \sum_{X\subseteq\mathcal{X}_p}\gamma_X\xi_X, \end{aligned} \tag{6}$$

where α_X, α_x, γ_X, and γ_x denote Lagrange multipliers. The minimization of Lagrangian operator in Eq. (6) with respect to $\boldsymbol{\beta}, \xi_X, \xi_x$ is obtained:

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} = 0 \Rightarrow \boldsymbol{\beta} = \displaystyle\sum_{X\subseteq\mathcal{X}_p}\alpha_X\phi(X) - \sum_{x\in\tilde{\mathcal{X}}_n}\alpha_x\phi(x), \\ \dfrac{\partial\mathcal{L}}{\partial\xi_X} = 0 \Rightarrow \gamma_X = \Delta_X C - \alpha_X, \\ \dfrac{\partial\mathcal{L}}{\partial\xi_x} = 0 \Rightarrow \gamma_x = C - \alpha_x. \end{cases} \tag{7}$$

Due to the nonnegativity of γ_X and γ_x, we have $0 \leq \alpha_X \leq C\Delta_X$ and $0 \leq \alpha_x \leq C$. Given a test example $\tilde{x}$, the detection score can be represented as:

$$f(\tilde{x}) = \left(\sum_{X\subseteq\mathcal{X}_p}\alpha_X\phi(X) - \sum_{x\in\tilde{\mathcal{X}}_n}\alpha_x\phi(x)\right)\phi(\tilde{x}), \tag{8}$$

It can be seen that the final detection score $f(\tilde{x})$ is a weighted combination of the inner product between training features $\phi(X)$, $\phi(x)$ and test feature $\phi(\tilde{x})$, and is only determined by samples with nonzero coefficients α_i $(i = X, x)$. These α_is are called support vectors, since they are the only training samples necessary to define the separating hyperplane. Note that for positive samples, the coefficient α_X is bounded by $C\Delta_X$, with KKT conditions, it is also possible to see when an example is a support vector, this happens only if the example is on the margin, or it does not respect the separation conditions in Eq. (4). According to [7], the coefficient α_X for positive samples in different locations is defined as:

$$\begin{cases} \alpha_X = 0, & \boldsymbol{\beta}^T\phi(X) > 1, \\ \alpha_X = C\Delta_X, & \boldsymbol{\beta}^T\phi(X) < 1, \\ 0 < \alpha_X < C\Delta_X, & \boldsymbol{\beta}^T\phi(X) = 1. \end{cases} \tag{9}$$

For positive bags which do not respect the classification hyperplane, the corresponding coefficient α_X is bounded by $C\Delta_X$, which takes the reliability of X into consideration. The regularized term Δ_X helps to boost the detection performance. If a positive bag X is not reliable at previous round, its contribution to the classification hyperplane at current round would be lowered. As a result, MIL introduces diverse samples for detector learning, while the confidence loss term encourages the detector focusing on positive instances which are good enough and downweighting those instances with lower reliability. The whole procedure of the proposed weakly supervised detector learning algorithm is summarized in Algorithm 1.

Algorithm 1 Weakly Supervised Detector Learning

Input: Positive bags $\mathcal{X}_p$, negative bags $\mathcal{X}_n$, the number of spectral clusters $\mathcal{K}$ per category, and iterations T;

Pattern Mining with Weak Detectors: For instances in the positive bags $\mathcal{X}_p$, mining patterns with weak detectors.

a). Select discriminative patches $\{x_i\}_{i=1}^m$ with Eq. (1) via cross-validation.

b). For each selected patch $x_p \in \{x_i\}_{i=1}^m$, learn exemplar-SVM detector via Eq. (1).

c). Spectral clustering of detectors $\{\boldsymbol{d}_i\}_{i=1}^m$ into $\mathcal{K}$ clusters $\{C_k\}_{k=1}^{\mathcal{K}}$.

d). For each cluster, pattern mining on $\mathcal{X}_p$ according to scores $s(C_k|x) = \sum_{\boldsymbol{d}_k \subseteq C_k} \boldsymbol{d}_k\phi(x)$.

Strong Detector Learning via cls-MIL: For each cluster, given initial patterns discovered by weak detectors, solving cls-MIL in Eq. (4) via iteratively updating and optimizing.

For iteration t=1 to T

a). Updating: Updating the latent variables via cross-validation. The latent variables in $\mathcal{D}_k$ are determined by detectors $\boldsymbol{\beta}_{\mathcal{D}\setminus\mathcal{D}_k}$ trained on $\{\mathcal{D}\setminus\mathcal{D}_k\}$, *i.e.*, updating instance weights w_m of $\Phi(X)$ by: $w_m = \sigma[\boldsymbol{\beta}_{\mathcal{D}\setminus\mathcal{D}_k}^T\phi(x_m)]$, and the confidence loss term $\Delta_X = \sigma[\boldsymbol{\beta}_{\mathcal{D}\setminus\mathcal{D}_k}^T\Phi(X)]$.

b). Optimizing: solving Eq. (4) via hard negative mining on negative bags $\mathcal{X}_p$ with the updated latent variables $\Phi(X)$ and Δ_X.

end

Output: Detector set $\{\boldsymbol{\beta}_k\}_{k=1}^{\mathcal{K}}$.

IV. Applications: Image Classification and Object Localization

The learned detectors are discriminative for the corresponding category, and an ensemble of the detectors across different categories offers an effective mid-level image representation. In this section, we apply such mid-level representation for image classification and object localization.

A. Image Classification

Unsupervised clustering methods have been used for feature representation [31], [44]. Since our learned detectors can be considered as the true visual patterns corresponding to a certain category (as opposed to the clustered ambiguous visual letters in [31] and [44]), it makes sense to apply such detectors for image coding. Denote all the learned detectors across different categories as $\Gamma = \{\boldsymbol{\beta}_i\}_{i=1}^K$, where K is the total number of detectors. Our mid-level feature representation is based on the maximal responses of a collection of detectors. Specifically, given an image I and the corresponding region proposals X, the feature representation is denoted as: $f(I, \Gamma) = [\boldsymbol{\beta}_1^T\phi(I, z_1), \ldots, \boldsymbol{\beta}_K^T\phi(I, z_K)]$, where z_k is a latent variable indicating the region with maximum response corresponding to detector $\boldsymbol{\beta}_k$, *i.e.*, $z_k = \text{argmax}_{z\in X}\boldsymbol{\beta}_k^T\phi(I, z)$. An illustration of image representation is shown in Fig. 5.

Given the image representation, a conventional SVM classifier is performed to produce the final classification results. Note that the complexity of the feature coding using detector responses is very low, which includes no more than a dot product operation once the features (*e.g.*, CNN) are extracted. On the other hand, we greedily select detector responses based on the entropy coverage criterion, and find that the performance saturates as the first few detectors are added in, which decreases the feature dimension by one order. In the experimental section, we will demonstrate the effectiveness of the proposed feature coding approach.

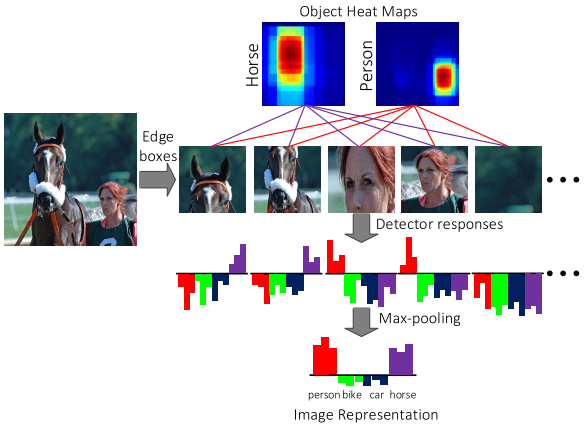


Fig. 5. An illustration of how to compute image representation and object heat maps according to the detector responses.

B. Object Localization

The learned part detectors are discriminative for the corresponding category, and a collection of them offers rough

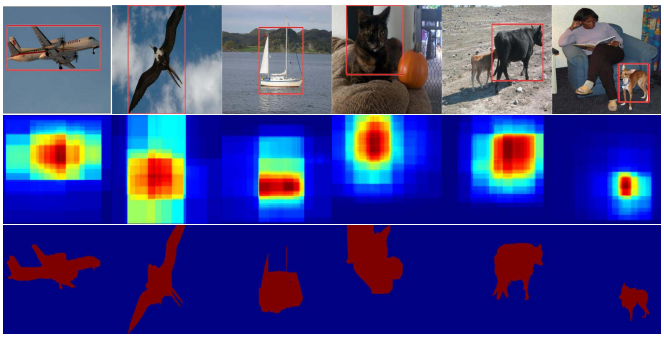


Fig. 6. Examples of localization process on Pascal VOC 2007 *trainval* split. We generate the object heat map and perform grabcut [34] to obtain segmentation mask of the object. Then a tight object bounding box (shown in red) is obtained via enclosing the segmentation mask.

position of the object of interest. In this section, we present a simple object localization technology based on the learned part detectors. The basic idea is to accumulate the part responses into a whole object heat map, which indicates the potential position of an object. Specifically, starting from a collection of part detectors $\{\boldsymbol{\beta}_k\}_{k=1}^{\mathcal{K}}$ corresponding to a category, we first define a part map $\mathcal{O}_k$ based on detector $\boldsymbol{\beta}_k$, the confidence of a pixel p which is contained in an object part is denoted as:

$$\mathcal{O}_k(p) = \frac{\sum_{x_m \subset \Omega_p} \sigma[\boldsymbol{\beta}_k^T \phi(I, x_m)]}{Z}, \tag{10}$$

where Ω_p denotes the patch set that includes pixel p, σ is a sigmoid function, and Z is a normalization constant such that $\max \mathcal{O}_k(p) = 1$. Finally, the object map is a weighted linear combination of the part maps obtained by all part detectors, *i.e.*, $\mathcal{O}(p) = \sum_{k=1}^{\mathcal{K}} \frac{w_k \mathcal{O}_k(p)}{\sum_k w_k}$, where w_k is a weight factor which denotes the reliability of each detector, and is given by $w_k = \max_{m \in X} \sigma[\boldsymbol{\beta}_k^T \phi(I, x_m)]$. Fig. 5 illustrates examples of how to compute the object heat maps.

The object heat map indicates the most discriminative details of an object, and usually focuses on object parts (*e.g.*, the head of dogs), instead of the whole object. Inspired from [30] which casts localization as a segmentation task, we perform grabcut [34] on the object heat map to generate the segmentation mask. The goal is to propagate the discriminative part details to the whole object with color continuity cues. To this end, the foreground and background are set to be gaussian mixture models. The foreground model is estimated from heat map values higher than 0.8, and the background model is estimated from values lower than 0.2. Finally, we take the bounding box that covers the largest connected component in the generated segmentation mask as localization result. Some example localization processes are shown in Fig. 6. In the experimental section, we will show that as a byproduct of the learned discriminative detectors, such localization technique achieves satisfactory localization performance.

V. Experiments

In this section, we present an evaluation of the proposed weakly supervised image classification and object localization framework. We also perform ablation study to understand how various design choices impact the recognition performance.

A. Datasets and Evaluation Metrics

We evaluate the proposed approach on four publicly available benchmarks, ranging from different scales. The details of the datasets are briefly summarized as follows:

Pascal VOC 2007 and 2012: The Pascal VOC datasets are widely used benchmarks for multi-label image classification and object localization. We choose VOC 2007 and VOC 2012 for evaluation. The VOC 2007 [12] contains a total of 9,963 images spanning 20 generic object classes, of which 5,011 images are used for *trainval* and the rest 4,952 images for *test*. The VOC 2012 [11] is an extended version of the Pascal VOC 2007, which contains a total of 22,531 images, including 11,540 images for *trainval* and 10,991 images for *test*. For image classification, we choose *trainval* split as training set and *test* split as test set, and the evaluation metrics is mean Average Precision (mAP).

MIT Indoor-67: The MIT Indoor-67 [32] dataset consists of 15,620 images belonging to 67 categories of indoor scenes. It is challenging because of the large ambiguities between categories. We follow the standard *train/test* split as in [32], *i.e.*, approximately 80 images per class for *train* and 20 images per class for *test*. The evaluation metric for MIT Indoor dataset is the mean classification accuracy.

MS COCO 2014: The MS COCO [25] is a large scale dataset which contains over 135k images spanning 80 categories. We choose the 80k *train* split for training and the 40k *val* split for test. The evaluation criterion is mean Average Precision (mAP).

In addition to classification, we also evaluate the localization performance of the proposed approach. For PASCAL VOC, we evaluate the performance on *trainval* set with CorLoc [8]. While for MS COCO, a point-based object localization metric [28] is chosen. This metric is widely used on MS COCO and we choose it for fair comparisons with previous works.

B. Implementation Details

1) Models and Features: We choose two widely used CNN models for feature extraction, a typical network CaffeNet [17] and a more accurate but deeper one VGG-VD [37] (the 16-layer model). We extract features from the fc6 layer (FC-CNN) after the rectified linear unit (ReLU), which is a 4096-d nonnegative vector for each region. Edge boxes [53] are used for generating candidate region proposals. In addition to region proposals, edge boxes also provide an objectness score for each region. For computation efficiency, we disregard regions which occupy less than 1% areas of an image, and retain the top scored 500 region proposals as candidates.

2) Parameter Settings: In pattern mining, the number of spectral clustering per category $\mathcal{K}$ is set as 50, and the top scored 100 patches per clustered detectors are selected as patterns for detector initialization. In detector optimization, the number of iterations T is set as 3, as we find that the performance of the detectors do not need more to converge. For simplicity, the number of detectors is equally selected among categories, and the optimal number of detectors per category is obtained by cross validation on the training set.

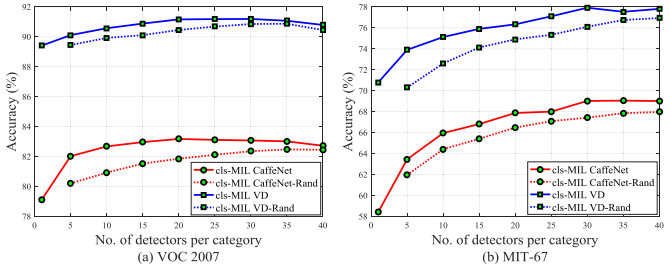


Fig. 7. The classification performance with respect to the number of detectors per category on (a) VOC 2007 and (b) MIT Indoor-67. The solid lines indicate the greedily selected detectors via entropy coverage criterion, while the dashed lines indicate the randomly selected detectors.

TABLE I

RECOGNITION PERFORMANCE ON VOC 2007 WITH DIFFERENT NUMBER OF REGION PROPOSALS. RESULTS ARE BASED ON MODEL CAFFENET

NO. of proposals	300	500	1000	2000
mAP	82.4%	83.2%	83.4%	83.7%

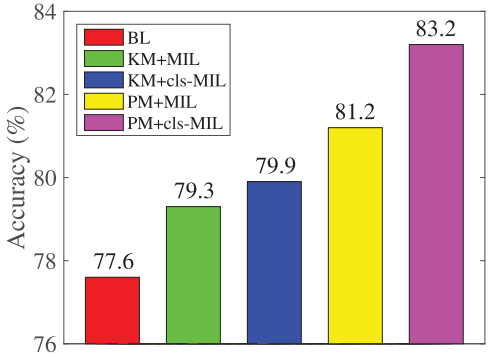


Fig. 8. The classification performance comparisons with different configurations on Pascal VOC 2007 *test* split. BL refers to baseline which max pooling CNN region features, KM is short for standard k-means pattern initialization algorithm, PM denotes the proposed pattern mining approach, MIL stands for standard multiple instance learning, and cls-MIL is the proposed confidence loss sparse MIL method. Results are based on model CaffeNet.

For all situations where cross-validation is needed, we use typical 5-fold cross-validation.

C. Ablation Study

To better understand the relative contribution of each module, we analyze the performance of our approach with different configurations. As the localization can be regarded as a byproduct of the learned detectors, we mainly measure how different designs affect the discriminativeness of the detectors in terms of classification performance.

1) Effects of Number of Detectors: An advantage of the proposed approach is that detectors are trained from patterns with different entropy coverages. This enables us to greedily select detectors based on the entropy coverage criterion. As shown in Fig. 7, we add detectors orderly to probe how the number of detectors affects the classification performance. It can be seen that the performance improves fast when a small number of detectors are added. For example, on PASCAL VOC 2007 with CaffeNet model, the performance boosts from 79.1% to 82.7% when the number of detectors per category increases from 1 to 10, which demonstrates the discriminativeness of the learned detectors. Comparatively, the performance of randomly selected detectors is inferior to our greedily selected detectors with the same dimension. Notably, we achieve an accuracy of 79.1% when the number of detectors per category is only 1 (total feature dimension is 20), which demonstrates that the entropy coverage criterion is able to select more discriminative detectors first. The performance tends to be stable and even drops sightly when more detectors are added. This is mainly because the subsequent detectors are not discriminative enough for classification. Similar results can be found on MIT-67 dataset. We obtain accuracy of 69.0% with CaffeNet model, and 77.9% with VGG-VD model when the number of detectors is 30 per category.

2) Effects of Number of Region Proposals: In order to probe the performance with respect to the number of candidate region proposals, we select the number of region proposals in different settings. Table I shows the results on VOC 2007 by varying the number of region proposals. The performance are relatively stable (from 2000 to 300 region proposals, only 1.3% drop). Considering the performance and computational efficiency, we choose the number of region proposals as 500.

3) Effects of Different Modules: We now compare the results with different configurations to analysis how each module affect the final classification performance. Different modules are summarized as follows:

- BL: This is the baseline method which directly max pooling multiple region proposal features for classification. It is introduced to help understand how the proposed approach improve the discriminative power of the detectors.
- PM & KM: PM denotes the proposed pattern mining method in Sec. III A, while KM is the standard k-means clustering method that is widely used for detector initialization in previous algorithms [29], [38], [40]. For fair comparisons, we perform k-means clustering on the selected patches with the number of clusters setting as 20.
- MIL & cls-MIL: MIL stands for standard multiple instance learning method which mines new positive sample without considering the confidence of each bag, and cls-MIL is the confidence loss sparse MIL detector optimization strategy proposed in Sec. III B.

As shown in Fig. 8, both k-means and multiple instance learning do help to improve the classification performance, nevertheless with limited gains. The proposed pattern mining and cls-MIL method surpass the counterparts consistently, *e.g.*, pattern mining improves the accuracy from 79.9% (k-means) to 81.2%, and cls-MIL obtains an accuracy improvement of 2% (83.2% vs 81.2%) comparing with standard MIL. We also find that detector initialization really counts for multiple instance learning, even for the modified cls-MIL (79.9% with k-means, and 83.2% with pattern mining). This is widely discussed in previous approaches which aim to develop efficient pattern mining methods [24], [3] for detector initialization. However, few works emphasis detector optimization. We demonstrate that both modules are essential, and a combination of them achieves considerable performance improvement.

TABLE II

RECOGNITION AVERAGE PRECISION (%) ON VOC 2007 *Test* SPLIT. WE REPORT PERFORMANCE WITH TWO MODELS: CAFFENET [17] AND VGG-VD [37]

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
MR-CaffeNet [17]	90.4	87.0	87.2	84.1	40.5	76.4	86.9	87.5	60.7	70.5	75.7	82.7	89.4	80.4	93.9	53.9	76.6	66.6	90.9	71.5	77.6
MR-VD [37]	98.3	95.3	96.0	95.0	70.0	90.1	93.8	94.9	73.7	84.6	85.9	94.5	95.4	92.0	97.5	70.6	90.6	79.7	98.1	86.7	89.1
MILinear [33]	83.8	79.3	75.3	47.8	55.6	77.5	87.8	70.6	60.8	60.8	65.1	73.4	80.6	84.8	90.0	61.2	65.9	65.2	78.8	75.0	72.0
Transfer Alex* [27]	88.5	81.5	87.9	82.0	47.5	75.5	90.1	87.2	61.6	75.7	67.3	85.5	83.5	80.0	95.6	60.8	76.8	58.0	90.4	77.9	77.7
HCP Alex* [45]	95.4	90.7	92.9	88.9	53.9	81.9	91.8	92.6	60.3	79.3	73.0	90.8	89.2	86.4	92.5	66.9	86.4	65.6	94.4	80.4	82.7
HCP VD* [45]	98.6	97.1	**98.0**	95.6	75.3	**94.7**	95.8	97.3	73.1	90.2	80.0	**97.3**	96.1	**94.9**	96.3	78.3	**94.7**	76.2	97.9	91.5	90.9
WSDDN VD [4]	93.3	93.9	91.6	90.8	**82.5**	91.4	92.9	93.0	**78.1**	**90.5**	82.3	95.4	92.7	92.4	95.1	**83.4**	90.5	80.1	94.5	89.6	89.7
WSDL CaffeNet	94.6	92.0	90.4	89.3	56.9	81.9	93.0	90.8	67.9	71.7	77.0	84.9	89.7	86.4	97.1	71.8	80.7	69.4	93.8	84.3	83.2
WSDL VD	98.6	97.7	97.2	96.0	78.4	92.0	95.8	96.9	76.5	86.9	82.4	94.1	95.3	93.5	98.6	79.4	94.5	80.1	98.6	92.2	**91.3**
WSDL+MR-VD	**99.3**	**97.8**	97.6	**96.4**	79.1	92.9	**95.9**	**97.3**	78.0	88.5	**87.1**	95.4	**96.1**	94.4	**98.7**	80.0	94.6	**82.9**	**99.0**	**92.2**	**92.2**

The method marked with * are those using additional training images

TABLE III

RECOGNITION AVERAGE PRECISION (%) ON VOC 2012 TEST.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
Transfer Alex* [27]	93.5	78.4	87.7	80.9	57.3	85.0	81.6	89.4	66.9	73.8	62.0	89.5	83.2	87.6	95.8	61.4	79.0	54.3	88.0	78.3	78.7
Weak Sup.* [28]	96.7	88.8	92.0	87.4	64.7	91.1	87.4	94.4	74.9	89.2	76.3	93.7	95.2	91.1	97.6	66.2	91.2	70.0	94.5	83.7	86.3
HCP Alex* [45]	97.7	83.2	92.8	88.5	60.1	88.7	82.7	94.4	65.8	81.9	68.0	92.6	89.1	87.6	92.1	58.0	86.6	55.5	92.5	77.6	81.8
HCP VD* [45]	**99.1**	**92.8**	**97.4**	**94.4**	**79.9**	**93.6**	**89.8**	**98.2**	78.2	**94.9**	79.8	**97.8**	**97.0**	93.8	96.4	**74.3**	**94.7**	71.9	96.7	88.6	**90.5**
WSDL CaffeNet	96.2	84.9	90.7	87.1	61.8	89.9	83.4	92.1	71.1	77.8	73.4	89.6	88.1	89.8	96.4	63.6	82.9	63.7	93.1	82.2	82.9
WSDL VD	99.0	90.7	95.5	93.7	78.9	93.2	88.6	97.3	**80.5**	91.3	**81.6**	96.0	96.1	**95.2**	**97.9**	70.0	93.6	**72.3**	**97.5**	**89.0**	89.9

The method marked with * are those using additional training images. available at http://host.robots.ox.ac.uk:8080/anonymous/ukzvbm.html and http://host.robots.ox.ac.uk:8080/anonymous/cd25ho.html

D. Image Classification

1) PASCAL VOC: Table II and III show the object recognition results of the proposed approach on Pascal VOC 2007 and 2012 *test* splits, respectively. In order to make fair comparisons, we extract CNN features from multiple region proposals, and max-pooling the region features into a final representation, which we refers to MR-CNN. Then the only difference between MR-CNN and our method is the detectors since they make use of the same region proposals. From Table II we can see that the proposed detectors improve the classification performance considerably, achieving accuracies of 83.2% with CaffeNet, and 91.3% with very deep model, which bring 5.6% and 2.2% gains comparing with using CNN features.

There exist many previous approaches that report classification results on Pascal VOC dataset, and we compare our results with some most recent ones. Most of previous approaches that achieve high classification results are based on network fine tuning [4], [27], [45]. Since network fine tuning is hard for multi-label images, previous works [27] rely on object annotations to find category specific patches. In [45], the authors proposed a weakly supervised classification framework via two-steps of network fine tuning, while it is trained using extended ILSVRC datasets, enriched with additional categories semantically close to the ones in PASCAL VOC. Our result (91.3%) is slightly better than the best performing one (90.9%) [45], demonstrating that the traditional optimization approaches are able to achieve competing results with CNN fine tuning. Furthermore, the proposed features are complementary with CNN features, and achieve an accuracy of 92.2% when combined. For VOC 2012, our method obtains an accuracy of 89.9%, which is slightly worse than [45] (90.5%) that makes use of additional training images and late fusion with a complex hand-crafted method [47]. The reason lies in that CNN-based methods are powerful as the training data grow, while MIL-based methods are relatively robust to the amount of data.

TABLE IV

COMPARISONS OF RECOGNITION PERFORMANCE ON MIT INDOOR-67. CLUSTERED DETECTORS REFER TO DIRECTLY USING CLUSTERED EXEMPLAR-SVM DETECTOR RESPONSES AS FEATURES

Method	Dimension	Accuracy (%)
DMS [10]	67K	64.0
DSFL [54]	42K	76.2
MOP-CNN CaffeNet [15]	13K	68.9
MDPM VD [24]	3.3K	77.6
FC-CNN CaffeNet [17]	4K	60.3
MR-CNN CaffeNet [17]	4K	65.1
Weak Detectors CaffeNet	2K	66.3
WSDL CaffeNet	2K	69.0
WSDL VD	2K	77.9
WSDL VD+ [37]	6K	**80.1**

2) MIT-67: Table IV compares the recognition results on MIT Indoor-67. MR-CNN denotes max-pooling multiple region features for representation, and FC-CNN refers to directly extract a single global feature from the whole image. Weak detectors denote the method which relies on the responses of the clustered exemplar-SVM detectors as features. From Table IV we observe that: 1) MR-CNN is much better than FC-CNN. Using CaffeNet model, the accuracy is 65.1% with MR-CNN, and 60.3% with FC-CNN. This demonstrate that local features are crucial for scene recognition. 2) The features using clustered detector responses (66.3%) is better than MR-CNN (65.1%), even with half dimension (2K versus 4K). This is mainly because CNN is primarily trained from the object centric images, instead of the scene centric images.

TABLE V

CLASSIFICATION (*val* SPLIT) AND POINT LOCALIZATION (*train* SPLIT) PERFORMANCE ON MS COCO 2014. THE MODEL IS BASED ON CAFFENET

Methods	Classification(%)	Point Localization (%)
MR-CNN	58.1	36.8
DeepMIL [28]	62.8	42.1
MultiFC [20]	60.4	**45.8**
WSDL	62.2	44.7
WSDL+MR-CNN	**63.9**	-

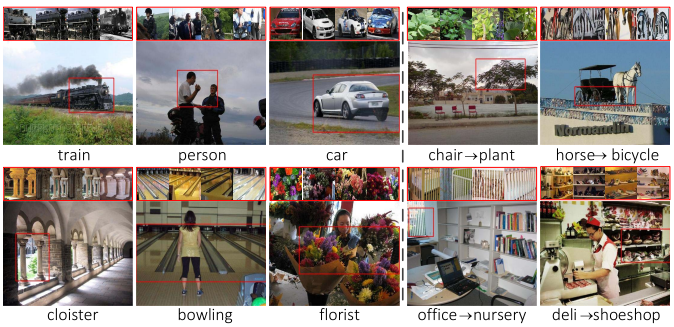


Fig. 9. Some visualizations of the correct and incorrect classification. We show the top detection that makes it look like the corresponding category, and some patches that the detectors are trained on.

As a result, the weak exemplar-SVM detectors still outperform MR-CNN due to the data specific representation. 3) The proposed WSDL is much better than the features with clustered responses. Benefit from the detector optimization strategy, our method obtains an accuracy of 69.0%, which brings about a 2.7% improvement comparing with the clustered responses. The performance is boosted to 77.9% when switching to the very deep model. Another observation is that the proposed features are complementary with CNN features, and achieve an accuracy of 80.1% when combined.

There are some approaches which also aim at learning discriminative part detectors for recognizing indoor scenes. The method of [29] integrates detector learning and classification by jointly training, and [10] poses mid-level pattern discovery as discriminative mode seeking via developing an extension of the classic mean-shift algorithm to density ratio estimation. Our method is closely related to [24], which also makes use of CNN activations for pattern mining. Our method achieves slightly better result comparing with the best performing method (77.9% vs 77.6%). There exists a majority of algorithms which employ multiple region pooling for final feature representation. A typical representation is MOP-CNN [15] which uses VLAD to encode CNN activations into bag of words representation, and achieves an accuracy of 68.9%, our results (69%) is comparable with [15] using the same model, but with much lower dimension (2K vs 13K).

3) MS COCO: We also conduct experiments on MS COCO to test whether the proposed method scales well on large scale datasets. We simply follow the settings of PASCAL VOC 2007, *i.e.,* setting the number of detectors per category as 20. Table V shows the classification results on MS COCO with CaffeNet model. From this table we find that our proposed WSDL method achieves a classification accuracy of 62.2%, which is about 4.1% improvement over MR-CNN, and is comparable with [28] (62.8%) which uses a more deep Overfeat [35] model. Again, WSDL is complementary to the MR-CNN and the result is 63.9% when combining them. The results demonstrate that WSDL scales well on the large scale dataset MS COCO.

4) Visualizing Mid-Level Patterns: As an illustration, Fig. 9 shows some discovered patterns on VOC 2007 (top row) and MIT-67 (bottom row) *test* splits. We show the highest activation region per image, which offers a clue indicating why it is classified as the corresponding category. Specifically, given a test image and the category label that the image is classified with (no matter correct or not), we employ category specific detectors to find which region responds most to the given category, and show some patches that the detector is trained on. For correctly classified images, there often exist discriminative patches that respond significantly to the corresponding detectors, *e.g.*, on VOC 2007, the head of a train is important for recognizing the trains, and the upper body of a person is important for recognizing the persons. Similar results can be found on MIT-67, it is the pillar of a cloister that makes it look like a cloister, and the slide rail that makes bowling look like bowling. It is helpful to investigate why incorrect results happen, on VOC 2007, a classifier misclassifies chair as the plant, or horse as bicycle, probably because there exist corresponding details, *e.g.*, the wheel of the carriage is similar with bicycle wheels. Similar results can be found on MIT-67, the window of the office is misclassified as the bar of the baby bed, which is most discriminative for recognizing nursery. Actually, these details look similar, and it is hard to recognize them. However, these observations offer a direction to further improve the recognition performance.

E. Object Localization

1) PASCAL VOC: Table VI shows the image localization results on Pascal VOC 2007 *trainval* split. Benefit from the learned part detectors, the proposed localization strategy (47.7%) is better than recent methods that is specifically designed for localization [3], [23], and is comparable with [43] (48.5%) which uses latent category learning for object localization. Another observation is that different from recognition, using deeper model does not bring about localization improvement (46.9%). This can be explained with the fact that deeper models frequently focus on parts of the object instead of the whole object. Note that all these comparing methods are designed for localization, which often makes use of context information for better localization, while we rely on detectors which are learned for classification to uncover the connection between these two basic tasks. The results demonstrate that image classification and localization can be done simultaneously.

2) MS COCO: Table V shows the localization results on MS COCO. In our implementation, the point-based localization is obtained by selecting pixels with the maximal response of the object heat map (as shown Fig. 6). WSDL obtains an accuracy of 44.7%, which is about 8 point improvement over MR-CNN.

TABLE VI
OBJECT LOCALIZATION PRECISION (%) ON VOC 2007 TRAINVAL IMAGES IN TERMS OF CORLOC METRIC

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
Mimick [23]	**73.1**	45.0	43.4	**27.7**	6.8	53.3	58.3	45.0	6.2	48.0	14.3	47.3	**69.4**	66.8	24.3	12.8	51.5	25.5	**65.2**	16.8	40.0
Con-Clust [3]	66.4	59.3	42.7	20.4	21.3	63.4	**74.3**	**59.6**	21.1	**58.2**	14.0	38.5	49.5	60.0	19.8	39.2	41.7	30.1	50.2	44.1	43.7
MMIL [6]	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	**36.7**	18.7	56.5	13.2	54.9	59.4	38.8
PLSA [43]	80.1	**63.9**	**51.5**	14.9	21.0	55.7	74.2	43.5	**26.2**	53.4	16.3	**56.7**	58.3	69.5	14.1	38.3	58.8	47.2	49.1	**60.9**	**48.5**
WSDL CaffeNet	60.8	55.3	43.8	16.5	**29.4**	64.5	69.3	49.4	12.6	52.7	**29.7**	39.1	58.2	**81.1**	34.0	**39.6**	58.8	**47.8**	59.3	53.1	47.7
WSDL VD	60.8	58.8	40.8	17.6	24.8	**67.0**	68.1	50.0	12.2	48.6	27.4	36.5	58.2	78.7	29.7	36.6	**63.9**	44.4	58.9	55.6	46.9

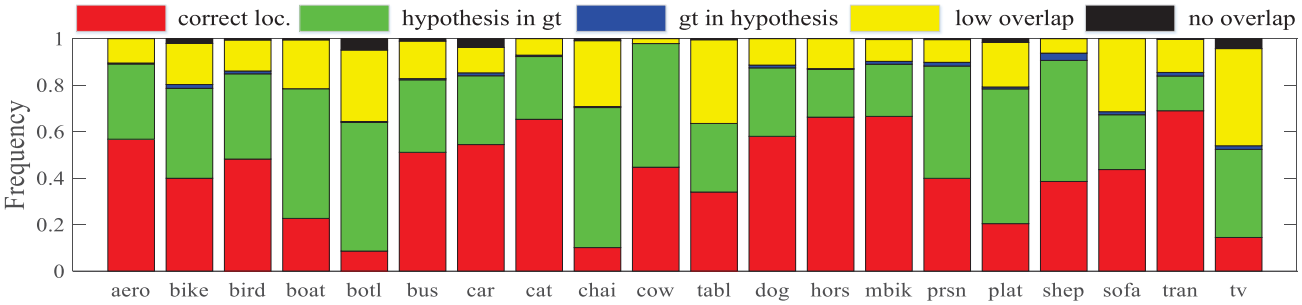


Fig. 10. An illustration of the error distribution of the proposed localization method on Pascal VOC 2007 *trainval* split.

Fig. 11. Examples of localization results on Pascal VOC 2007 *trainval* split. The correct localization are marked with red bounding boxes, while the failed ones are marked with green. The failed results often come from localizing object parts or grouping multiple objects from the same class.

This is achieved by the enhanced localization ability of the learned detectors. The localization is better than [28] (42.1%) and slightly inferior to [20] (45.8%) which are well-designed models based on CNN fine-tuning.

3) Localization Error Analysis: In order to better understand the localization errors, following [6], [23], we summarize the errors to uncover the pros and cons of our localization method. Each predicted bounding box is categorized into the following five cases: 1) correct localization, IoU overlap is greater than 50% with the ground truth. 2) hypothesis completely inside ground truth, 3) ground truth is completely inside the hypothesis, 4) no overlap, IoU equals to zero, and 5) low overlap, none of the above. Fig. 10 shows the error distribution of the proposed method across 20 categories on Pascal VOC 2007 *trainval* set. It can be noted that among the failed modes, the most important failure modality of our method is that an object part is localized instead of the whole object. This is intuitive since in most situations, correct classification only demands catching local discriminative details.

4) Visualizations and Limitations: Fig. 11 shows some localization results on Pascal VOC 2007 *trainval* split. The correct localizations are marked with red bounding boxes, while the failed ones are marked with green. It can be shown that the proposed localization method is able to find objects where there is only one object from the same category, but is short of localizing multiple objects of the same category. Actually, it is the main challenge for weakly supervised localization [6], and is a promising direction for future research.

5) Classification Versus Localization: Comparing classification (Table II) with localization (Table VI), we find that

the least successfully recognized objects are *bottle* (79.1%) and *chair* (78.0%), which are also hard for localization (24.8% and 12.2%). This is because they usually occupy a small fraction of the image, and are within cluttered backgrounds. The exception is *person*, which suffers a low localization accuracy (29.7%), but with a high recognition accuracy (98.7%). This can be explained by the fact that person is easy to be recognized by face, and usually, there exist multiple persons in an image, which offers abundant cues for recognition. In contrast, localization is failed when focusing on the face, and it is hard to tell apart individual person from the crowd.

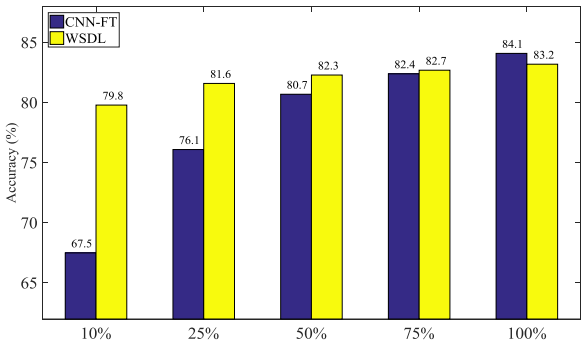


Fig. 12. The classification performance comparisons with different configurations on Pascal VOC 2007 *test* split.

F. Discussions

1) Comparing With RCNN: Our proposed detector learning needs to extract region proposal features and train detectors with SVM classifiers, which shares similar spirit with RCNN. However, these two methods are totally different in terms of experimental setups and application scenarios. RCNN is used for detection, which requires object bounding box annotations at training, while our proposed method is targeted at classification and localization, which only needs image level labels for training. As a result, these two methods differ a lot during detector training and performance evaluation:

• For detector learning, RCNN can easily define the positive and negative samples based on the ground truth annotations, while it is very hard to learn detectors in our scenario. To solve this issue, we propose a weak-to-strong detector learning strategy. For weak detector learning, we propose an exemplar-SVM based spectral clustering technology for pattern mining. While for strong detector learning, we propose a confidence loss sparse Multiple Instance Learning (cls-MIL) strategy. As a result, we train dozens of part detectors for each category, instead of one single object detector in RCNN.

• For performance evaluation, since our method and RCNN are developed to solve different tasks, their performance cannot be compared directly. Therefore, to compare their performance, we train a detection model following RCNN framework with only image-level labels, using the localized objects (Part B, Sec. IV) on the training images as ground truth. This configuration formulates a weakly supervised object detection task. On PASCAL VOC 2007, we obtain a detection accuracy of 32.4% with CaffeNet model, which is comparable with state-of-the-arts [4] (34.5%), [21] (31.0%), [43] (30.9%) using the same model. Comparatively, RCNN achieves a detection accuracy of 58.5%, which is better than ours because RCNN uses the ground truth object annotations during training. This demonstrates that object annotations are crucial in current detection models.

2) Comparing With CNN-Based Fine-Tuning: It is worth noting that our method is different from the CNN-based fine-tuning methods [4], [27], [45]. The main differences are:

• Our method is more powerful when the number of training samples is limited. Due to the large number of parameters, CNN-based fine-tuning requires abundant training samples to avoid overfitting. In contrast, our proposed detector learning is based on SVM classifiers, where each classifier is only determined by a few support vectors. To prove this, we conduct a comparative experiment on PASCAL VOC 2007 by randomly selecting a subset of training images for network fine-tuning and our detector learning. For CNN-based fine-tuning, we choose the end-to-end learning framework WSDDN [4]. For fair comparison, we only retain the top scored 500 edge boxes and do not use data augmentation during fine-tuning. The recognition results are shown in Fig. 12. WSDDN achieves an accuracy of 84.1% when fine-tuning with all the training samples, which is 0.9% better than our detector learning method. However, as the number of training samples decreases, the performance of WSDDN drops fast. For example, the accuracy is 80.7% when using half samples for training, and is only 67.5% when only using 10% samples for training. In contrast, our proposed detector learning is relatively robust to the varying number of training samples. The results are 82.3% with half training samples, and 79.8% with only 10% training images.

• Our method learns detectors by category, which conveniently enables parallel processing to accelerate the learning process. Another advantage is that the proposed method enjoys good transplant characteristics, we only need to learn extra category specific detectors when a new kind of category is added. However, the network has to be fine-tuned from the pretrained network again when adding a new category.

• The complexity of the detector learning is irrespective of the backbone network once the features are extracted, this is beneficial for learning detectors from a deeper network such as VGG-VD. However, the network fine-tuning methods highly dependent on the CNN structures. For example, on PASCAL VOC 2007 dataset, with parallel processing, the strong detector learning can be done in about 20 hours regardless of the networks used. However, it costs about 13 hours (a single GeForce GTX TITAN X is used) to fine-tune the CaffeNet network (the method of [4] is used), and over 60 hours to fine-tune the VDD-VD network.

3) Computational Complexity: The weak to strong detector learning requires iteratively updating detectors, which is the most time consuming module of the proposed method. The good news is that we can learn detectors with parallel processing since they are category independent. In weak detector learning, since the mean feature $\boldsymbol{\mu}_0$ and covariance matrix Σ^{-1} in Eq. (1) are fixed, each iteration only requires updating $\overline{\boldsymbol{x}}_p$.

Note that we only search the new positive samples from the selected discriminative patches, and the process is high-efficiency, *e.g.*, it takes around 2 hours to learn weak detectors on Pascal VOC 2007 dataset. In strong detector learning, the hard negative mining should be processed from all the negative images, which costs most of the time. With parallel processing, it takes about 20 hours for strong detector learning on PASCAL VOC 2007. Note that the whole detector learning is an off-line process, once we have the learned detectors, the extra computation of WSDL is no more than dot products between CNN features and the learned detectors.

VI. Conclusion

In this paper, we propose an effective mid-level image representation approach for visual applications. The proposed framework aims at learning a collection of discriminative part detectors in a weakly supervised paradigm, which only needs the labels of training images, while does not need any object / part annotations. Our approach tackles several key issues in automatic part detector learning. First, we propose an efficient pattern mining technique via spectral clustering of exemplar-SVM detectors. Second, we formulate the detector learning as a confidence loss sparse MIL (cls-MIL) task, which considers the diversity of the positive instances, while avoid drifting away the well localized ones by assigning a confidence value to each positive instance. The proposed method shows notable performance improvements on several recognition benchmarks. Furthermore, we simultaneously considering classification and localization based on the learned detectors, and find that the accumulated responses of part detectors offer satisfactory localization performance, which bridges these two widely studied visual tasks.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Neural Inf. Process. Syst.*, 2002, pp. 561–568.

[2] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 836–849.

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1081–1089.

[4] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2846–2854.

[5] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 105–112.

[6] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2409–2416.

[7] R. Collobert, "Large scale machine learning," Ph.D. dissertation, Univ. Paris VI, Paris, France, 2004.

[8] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.

[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

[10] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. Neural Inf. Process. Syst.*, 2013, pp. 494–502.

[11] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[15] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.

[16] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative decorrelation for clustering and classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 459–472.

[17] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.

[18] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 923–930.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[20] H. Lee, H. Kwon, A. J. Bency, and W. D. Nothwang. (2016). "Fast object localization using a CNN feature map based multi-scale search." [Online]. Available: https://arxiv.org/abs/1604.03517

[21] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3512–3520.

[22] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[23] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 19–34.

[24] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Mining mid-level visual patterns with deep CNN activations," *Int. J. Comput. Vis.*, vol. 121, no. 3, pp. 344–364, Feb. 2017.

[25] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[26] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 89–96.

[27] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?—Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 685–694.

[29] S. N. Parizi, A. Vedaldi, A. Zisserman, and P. Felzenszwalb. (2014). "Automatic discovery and optimization of parts for image classification." [Online]. Available: https://arxiv.org/abs/1412.6598

[30] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "The truth about cats and dogs," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1427–1434.

[31] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[32] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.

[33] W. Ren, K. Huang, D. Tao, and T. Tan, "Weakly supervised large scale object localization with multiple instance learning and bag splitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 405–416, Feb. 2016.

[34] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[36] K. J. Shih, I. Endres, and D. Hoiem, "Learning discriminative collections of part detectors for object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1571–1584, Aug. 2015.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[38] S. Singh, A. Gupta, and A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.
[39] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, "Weakly-supervised discovery of visual pattern configurations," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 1637–1645.
[40] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Dec. 2013, pp. 3400–3407.
[41] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," Comput. Neural Syst., California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[43] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank, "Large-scale weakly supervised object localization via latent category learning," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1371–1385, Apr. 2015.
[44] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.
[45] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.
[46] C. Xiong, L. Liu, X. Zhao, S. Yan, and T. K. Kim, "Convolutional fusion network for face verification in the wild," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 517–528, Mar. 2016.
[47] S. Yan *et al.*, "Generalized hierarchical matching for sub-category aware object classification," in *Proc. Eur. Conf. Comput. Vis. Workshop*. 2012, vol. 5, no. 6.
[48] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1134–1142.
[49] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *Int. J. Compt. Vis.*, vol. 120, no. 1, pp. 1–13, 2016.
[50] L. Zheng, Y. Yang, and Q. Tian. (2016). "SIFT meets CNN: A decade survey of instance retrieval." [Online]. Available: https://arxiv.org/abs/1608.01807
[51] Y. Zhu, J. Jiang, W. Han, Y. Ding, and Q. Tian, "Interpretation of users' feedback via swarmed particles for content-based image retrieval," *Inf. Sci.*, vol. 375, pp. 246–257, Jan. 2017.
[52] Y. Zhu, J. Liang, J. Chen, and Z. Ming, "An improved NSGA-III algorithm for feature selection used in intrusion detection," *Knowl.-Based Syst.*, vol. 116, pp. 74–85, Jan. 2017.
[53] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
[54] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 552–568.

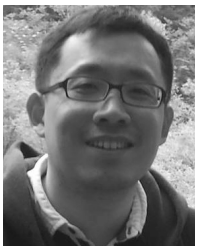
Xiaopeng Zhang received the B.S. degree from Sichuan University, Sichuan, China, in 2011 and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2017. He was a Visiting Researcher with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA, in 2016. He is currently a Post-Doctoral Research Fellow with the National University of Singapore.

His current research interests include object recognition, weakly supervised detection and multimedia signal processing.

Hongkai Xiong (M'01–SM'10) received the Ph.D. degree from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003. He is currently a Distinguished Professor with the Department of Electronic Engineering, SJTU. From 2005 to 2011, he was an Associate Professor and from 2003 to 2005, he was an Assistant Professor. From 2007 to 2008, he was a Research Scholar with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. From 2011 to 2012, he was a Scientist with the Division of Biomedical Informatics at the University of California (UCSD), San Diego, CA, USA. Since 2003, he has been with the Department of Electronic Engineering, SJTU.

His research interests include multimedia signal processing, image and video coding, multimedia communication and networking, computer vision, biomedical informatics, machine learning and published over 200 refereed journal and conference papers. He was the co-author of the Best Paper in the IEEE BMSB 2013, the Best Student Paper in VCIP 2014, the Top 10% Paper Award in VCIP 2016, and the Top 10% Paper Award in MMSP 2011.

Dr. Xiong research projects were funded by NSF, QUALCOMM, Microsoft, and Intel. He received the 2017 the Science and Technology Innovative Leader Talent Award in Ten Thousand Talents Program, the 2016 Yangtze River Scholar Distinguished Professor Award from the Ministry of Education, China, the 2014 National Science Fund for Distinguished Young Scholar Award from Natural Science Foundation of China, the 2017 Shanghai Academic Research Leader Talent Award, the 2014 Shanghai Youth Science and Technology Talent Award, the 2013 Shanghai Shu Guang Scholar Award, the 2017 Baosteel Excellent Faculty Award, the 2009 New Century Excellent Talent Award from the Ministry of Education of China, the 2010 and 2013 SJTU SMC-A Excellent Young Faculty Awards. He has been granted 2017 and 2011 First Prizes of the Shanghai Technology Innovation Award for research achievements on Network-Oriented Video Processing and Dissemination. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.



Weiyao Lin (M'10–SM'16) received the B.E. and M.E. degrees from Shanghai Jiao Tong University, China, in 2003 and 2005, and the Ph.D. degree from University of Washington, Seattle, USA, in 2010. He is currently an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. He has authored or co-authored over 100 technical papers in top journals/conferences. His research interests include image/video processing, video surveillance, and computer vision.

Dr. Lin served as an Associate Editor for *Journal of Visual Communication and Image Representation, Signal Processing: Image Communication, Circuits Systems and Signal Processing*, and IEEE ACCESS.

Qi Tian (M'96–SM'03–F'16) received the B.E. degree in electronic engineering from Tsinghua University in 1992, the M.S. degree in electrical and computer engineering (ECE) from Drexel University in 1996, respectively, and the Ph.D. degree in ECE from University of Illinois at Urbana-Champaign in 2002. He was a tenure-track Assistant Professor from 2002 to 2008 and a tenured Associate Professor from 2008 to 2012. From 2008 to 2009, he took one-year faculty leave at Microsoft Research Asia as a Lead Researcher at the Media Computing Group. He is currently a Full Professor with the Department of Computer Science, University of Texas at San Antonio (UTSA).

He was a co-author of the Best Paper in ACM ICMR 2015, the Best Paper in PCM 2013, the Best Paper in MMM 2013, the Best Paper in ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, the Best Student Paper in ICASSP 2006, and a co-author of the Best Student Paper Candidate in ICME 2015, and a Best Paper Candidate in PCM 2007. He has published over 400 refereed journal and conference papers. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian's research projects were funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He received the 2017 UTSA Presidents Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Awards from the College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is on the Editorial Board of *Journal of Multimedia* and *Journal of Machine Vision and Applications*. He is a Guest Editor of IEEE TRANSACTIONS ON MULTIMEDIA and *Journal of Computer Vision and Image Understanding*. He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Multimedia System Journal*.