Continuation Multiple Instance Learning for Weakly and Fully Supervised Object Detection

Qixiang Ye, *Senior Member, IEEE*, Fang Wan, *Member, IEEE*, Chang Liu, *Student Member, IEEE*, Qingming Huang, *Fellow, IEEE*, and Xiangyang Ji, *Member, IEEE*

***Abstract*— Weakly supervised object detection (WSOD) is a challenging task that requires simultaneously learning object detectors and estimating object locations under the supervision of image category labels. Many WSOD methods that adopt multiple instance learning (MIL) have nonconvex objective functions and, therefore, are prone to get stuck in local minima (falsely localize object parts) while missing full object extent during training. In this article, we introduce classical continuation optimization into MIL, thereby creating continuation MIL (C-MIL) with the aim to alleviate the nonconvexity problem in a systematic way. To fulfill this purpose, we partition instances into class-related and spatially related subsets and approximate MIL's objective function with a series of smoothed objective functions defined within the subsets. We further propose a parametric strategy to implement continuation smooth functions, which enables C-MIL to be applied to instance selection tasks in a uniform manner. Optimizing smoothed loss functions prevents the training procedure from falling prematurely into local minima and facilities learning full object extent. Extensive experiments demonstrate the superiority of CMIL over conventional MIL methods. As a general instance selection method, C-MIL is also applied to supervised object detection to optimize anchors/features, improving the detection performance with a significant margin.**

***Index Terms*— Continuation optimization, multiple instance learning (MIL), object detection, weakly supervised detection.**

I. Introduction

WEAKLY supervised object detection (WSOD), which requires solely binary annotations indicating whether a class of objects exists in images or not during training, has attracted increasing attention [1]–[6]. Compared with fully supervised object detection that requires labor-intensive bounding-box annotations, WSOD significantly reduces the workload of data annotation. With WSOD, people can leverage rich images with tags on the internet to learn object-level models and, thereby, convert human-supervised object detection to Webly supervised object modeling.

Nevertheless, WSOD remains an open problem when considering its performance is still far behind that of supervised detection methods (~20% on the PASCAL VOC detection benchmark) [7]–[10]. The challenge lies in that both detection models and object locations are latent and require to be estimated at the same time.

Multiple instance learning (MIL) has been a major WSOD method [11], [12], which treats labeled images as bags and estimates latent object locations (instances) when learning object detectors. However, it is observed that various MIL models [7], [13] are prone to learn object parts while missing full object extent, particularly in the early training epochs. The nature behind the phenomenon is the nonconvexity of MIL's objective function, which results in selecting discriminative object parts for image classification while missing full object extent for localization [see Fig. 1(a)]. Researchers have tried solving this problem by introducing regularization terms, e.g., spatial priori [8], [11], [14], [15], min-entropy [7], object-specific pixel gradient [16], reinforcement region searching [17], and online instance refinement [3], [12], [18]. Despite the progress, the local minimum problem remains not explored from the perspective of optimization.

In this article, we introduce classical continuation optimization [19] to MIL and propose continuation MIL (C-MIL)[1] to systematically explore the nonconvexity problem. In C-MIL, the object proposals in an image are regarded as instances, while images are regarded as bags of instances. Different from conventional MIL methods that select the most discriminative instance during training, C-MIL selects the discriminative instance subsets where instances are class-related and spatially related, i.e., having similar object class scores and overlapping with each other [see Fig. 1(b)].

We accordingly define a parametric strategy to implement continuation optimization, which enables C-MIL to be applied to instance selection tasks in a uniform manner. Specifically, we implement C-MIL by introducing a continuation parameter that increases from 0 to 1 during the training procedure. Based on the continuation parameter, instances in an image are partitioned into subsets. When the continuation parameter equals 0,

Manuscript received April 30, 2020; revised January 21, 2021; accepted March 30, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002202, in part by NSFC under Grant 61620106005, Grant 61836012, and Grant 62006216; and in part by the Post-Doctoral Innovative Talent Support Program under Grant 119103S304. *(Corresponding author: Fang Wan.)*

Qixiang Ye and Chang Liu are with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qxye@ucas.ac.cn; liuchang615@mails.ucas.ac.cn).

Fang Wan and Qingming Huang are with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: wanfang@ucas.ac.cn; qmhuang@ucas.ac.cn).

Xiangyang Ji is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: xyji@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3070801.

Digital Object Identifier 10.1109/TNNLS.2021.3070801

[1] Code for C-MIL is available at github.com/WanFang13/C-MIL

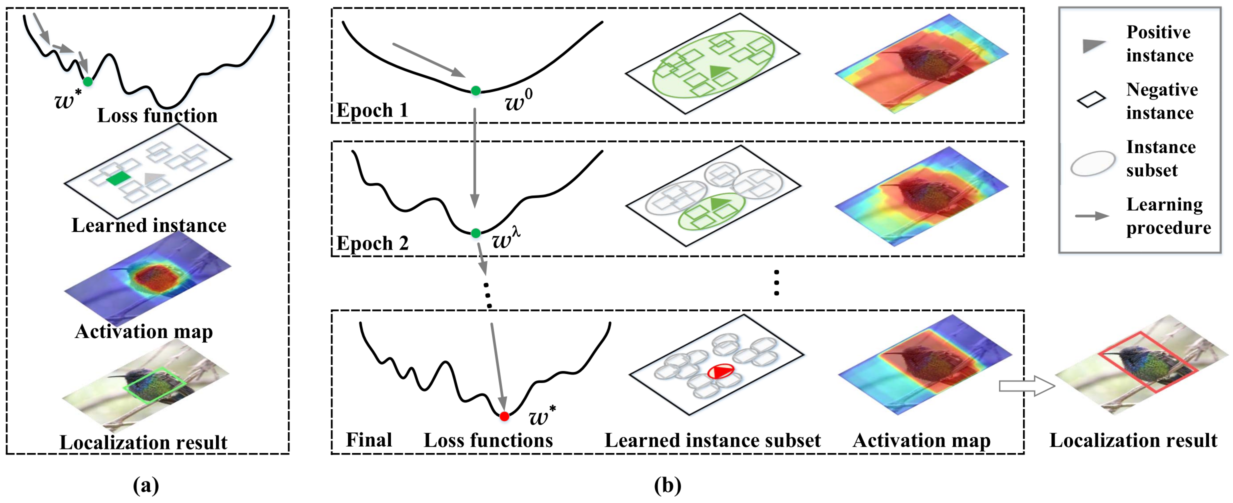


Fig. 1. Comparison of optimization procedures of (a) MIL and (b) C-MIL. MIL falls into local minima and falsely localizes an object part, which is caused by the nonconvex loss function. By constructing a series of functions, which can approximate the original objective/loss function but are easier to optimize, C-MIL alleviates the nonconvexity problem, pursues a stronger global minimum, and localizes the full object extent. (Best viewed in color.)

all instances in the image are partitioned into a single subset. At this moment, the objective function of C-MIL is equivalent to that of the image classification method [20] and is convex. When the continuation parameter equals 1, each instance subset contains a single instance. At this moment, the objective function of C-MIL degenerates to that of MIL. By continually increasing the continuation parameter, the instance subsets gradually decrease in size from containing all instances to a single instance.

Within the instance subsets, we construct a series of functions that can approximate the original objective function but are easier to optimize, Fig. 1(b). Training such objective functions prevents the model from falling prematurely into the local optimum and, therefore, approaches the global optimum. Consequently, instances in most discriminative subsets are selected, and the ones in less discriminative subsets are suppressed. The selected instance subsets are capable of collecting various object parts, which are combined to discover stable semantic extremal regions (SSERs) indicating full object regions, Fig. 1(b).

C-MIL was first proposed in our CVPR oral article [21] and is promoted to a general learning method by defining general objective functions in this full version. C-MIL is applied to not only instance selection under weakly supervised settings but also anchor/feature optimization under fully supervised settings. The contributions of this work include the following.

1) We propose a C-MIL method, which, by defining a series of smoothed objective functions to approximate MIL's objective function, systematically alleviates the nonconvexity problem of MIL.
2) We propose a parametric strategy to connect smoothed objective functions with instance subsets, which enables C-MIL to be applied to instance selection tasks in a uniform manner.
3) We apply C-MIL to WSOD, with the aim to prevent the training procedure from falling prematurely into local minima and facilitate discovering full object extent. We also apply C-MIL to supervised object detection to select optimal anchors/features and improve the adaptability of detectors.
4) We achieve significant performance gains for WSOD and supervised object detection on commonly used benchmarks, including PASCAL VOC and MS-COCO.

II. Related Work

In this section, we first review major approaches for weakly supervised methods. We then review continuation and smoothing methods for nonconvex optimization.

A. Weakly Supervised Method

Considering the unavailability of object locations, WSOD approaches leveraged latent variable learning or MIL to estimate object locations. Recently, the MIL network, which integrates MIL with deep feature learning, has attracted increased attention.

1) Latent Variable Learning: Latent variable learning was based on a hypothesis that a class of instances shapes a single compact cluster, while the negative ones form multiple diffuse clusters. Under such a hypothesis, Wang *et al.* [4], [22] calculated cluster labels of object proposals using probabilistic Latent Semantic Analysis (pLSA) and proposed a statistics-based approach to determine positive categories. Bilen and Song [2], [13] proposed leveraging latent variable clustering to discover object regions, object-part configurations, and subcategories, which were further used to learn detection models. Ye *et al.* [3] proposed a progressive latent model to discover object locations and learn detectors with progressive optimization.

Various latent ariable methods require solving the nonconvexity problem and, therefore, suffering from local minimum, which implies that they could falsely localize object parts or multiple objects. To solve this problem, object symmetry and class mutual exclusion information [1], convex clustering [13], and model smoothing [5] were introduced to the optimization functions as regularization.

2) MIL: As a major line of the WSOD method, MIL first decomposes each training image to a set of region proposals (instances). During training, each set of region proposals is treated as a "bag," and the true object locations are learned by iteratively performing instance selection and detector estimation. MIL works in a way similar to the expectation-maximization algorithm, which simultaneously labels instances and estimates models. Nevertheless, with a nonconvex objective function, MIL is often puzzled by the problem of local minima, particularly when facing a large solution space [7], [13].

To alleviate this problem, some approaches [5], [13], [22] used clustering as preprocessing or used bag splitting [23] to reduce the solution space. In multifold MIL [24], [25], the training set was split to "multifolders," where cross validation was carried out to alleviate the nonconvexity problem.

3) MIL Network: MIL networks refer to the deep neural networks (DNNs) fused with MIL, which selects instances and estimates detection models when learning feature representations [11]. While taking advantage of integrating feature learning with detector estimation, MIL networks inherited the nonconvexity drawback of MIL methods. To solve this problem, spatial regularization [11], context information [8], [14], min-entropy regularization [7], [26], object pixel gradient [16], and segmentation collaboration [15], [27], [28] were introduced to MIL networks.

In [15], semantic segmentation was introduced as a network branch within cascaded convolutional networks, which optimized instance selection and semantic segmentation in a two-stage iteration manner. In [8] and [14], context information was incorporated into networks to identify instances that are supported by and standing out from surrounding regions. In [7], [26], the clique-based min-entropy model was proposed as a regularization term to alleviate localization randomness when selecting instances. In [7] and [26], object-aware instance labeling was explored for accurate object localization by considering the completeness of instances.

In MIL Networks, classifier refinement strategies [7], [12], [15], [18], [30] were often used to produce high-quality instances that were treated as pseudo-objects to refine the instance classifier. The MELM method [7] used a recurrent learning algorithm to integrate image classification with object detection and then progressively optimize the classifiers and detectors. Online instance classifier refinement (OICR) [31] and proposal clusters (PCL) [31], [32] further improved object localization based on the observation that iterative generation of proposal clusters [31] could prevent networks from concentrating on object parts.

Despite noticeable progress, existing approaches remain not alleviating the local minimum problem from a perspective of optimization, which hinders the theoretical advance of weakly supervised learning. In this article, we introduce classical continuation optimization into MIL, thereby creating C-MIL, with the intention of alleviating the nonconvexity problem in a systematic way. Our research starts from improving object localization in WSOD, while it could also be applied to general latent variable learning problems.

B. Nonconvex Optimization

1) Continuation Methods: Continuation methods [33], [34] were proposed to deal with the complex optimization problem. In these methods, the objective function of the target task was often smoothed or approximated to multiple easier objective functions. These procedures were performed by introducing a continuation parameter. During model optimization, the continuation parameters were monotonically increased or decreased, and the optimization problem was accordingly converted to a sequence of subproblems, which finally converged to the problem of interest. These methods have achieved great success when facing optimization problems involving nonconvex objective functions with multiple local minima. Curriculum learning [35] shared a similar principle by defining a series of easy-to-difficult subtasks (or subdistributions), which finally converged to the task of interest.

2) Smoothing: Smoothing was a commonly used method in optimization [36] and has been integrated with deep neural networks. In [37], [38], a smoothing method that improved the nonsmooth ReLU activation for better optimization was proposed. In [39], "mollifiers" were introduced to smooth the loss function by gradually increasing its nonlinearity during model optimization, which can converge to stronger global minima [40]. In [41], entropy was introduced to reduce the randomness of object localization, while the essence is to smooth the loss function with an entropy function.

In this research, we propose a parametric strategy to continuously smooth objective functions over class-related and spatially related instance subsets. The proposed strategy enables C-MIL to be a general learning method, which, by specifying smoothed functions, can be applied to general instance selection tasks in image domains.

III. C-MIL

In this section, we first revisit the classical MIL method and clarify the procedure about simultaneously estimating instance labels and learning classifiers under the supervision of instance bag labels [42]. We then analyze the nonconvexity of MIL's objective function. Based on the analysis, we introduce continuation optimization to create C-MIL, with the aim to relax the nonconvex objective function and improve instance selection and detection training.

A. MIL Revisit

Instead of receiving a set of instances (region proposals) that are individually labeled, an MIL learner receives a set of labeled bags (images), each containing multiple instances. In the simple case of multiple-instance binary classification, a bag is labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if at least one instance in it is positive. From a collection of labeled bags, MIL tries to estimate a classifier that can discriminate bags and correctly select positive instances.

Let $x \in \mathcal{X}$ denote a bag of instances and $\mathcal{X}$ a set of bags for training. w denotes model parameters. $y \in \{+1, -1\}$ denotes the label of bag x. $y = +1$ indicates a positive bag that contains at least one positive instance, while $y = -1$

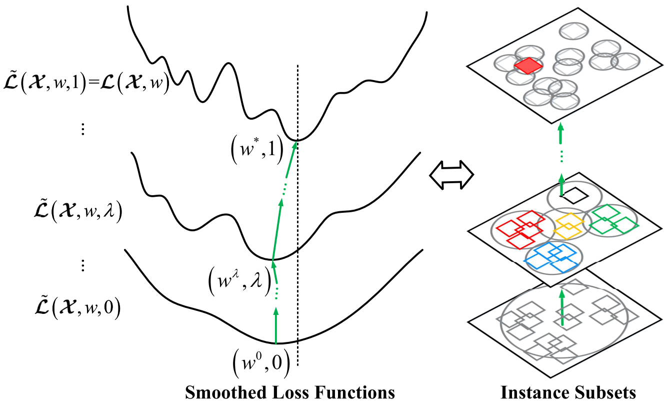

$\tilde{\mathcal{L}}(\mathcal{X},w,1)=\mathcal{L}(\mathcal{X},w)$
$(w^*,1)$
$\tilde{\mathcal{L}}(\mathcal{X},w,\lambda)$
(w^λ,λ)
$\tilde{\mathcal{L}}(\mathcal{X},w,0)$
$(w^0,0)$
Smoothed Loss Functions
Instance Subsets

Fig. 2. Continuation optimization of C-MIL. MIL's nonconvex objective function is approximated by a series of smoothed functions that are easier to be optimized. The smoothed functions are defined within instance subsets, which are partitioned with a parametric strategy. During training, C-MIL traces the series of smoothed functions from a starting point $(w^0, 0)$ to a solution point $(w^*, 1)$, which facilities pursuing a global optimal solution. (Best viewed in color.)

indicates a negative bag where all instances are negative. x_i and y_i (requires to be estimated), respectively, denote instances and instance labels in bag x, where $i \in \{1, 2, \ldots, N\}$ and N denotes the number of instances.

With MIL, an instance classifier is learned to select the top-scored instance x_{i^*} from each instance bag as

$$x_{i^*} = \arg\max_{x_i} f(x_i, w) \tag{1}$$

where $f(\cdot)$ is the instance classifier parameterized with w, which computes the classification score about each positive class. Accordingly, the objective function of MIL is formulated as

$$\begin{aligned}\{w^*, x_{i^*}\} &= \arg\min_{w,x_i} \mathcal{L}(\mathcal{X}, w)\\ &= \arg\min_{w} \sum_{x\in\mathcal{X}} \max(0, 1 - y \max_{x_i} f(x_i, w)) \quad (2)\end{aligned}$$

which defines a hinge-loss function for instance selection and classifier learning. Combined with a deep network, the instance classifier $f(\cdot)$ selects top-scored instances to optimize the classification of bags, as well as updating the instance classifier (network parameter w) using an SGD algorithm.

B. Convexity Analysis

It is notable that the maximum of multiple convex functions is convex. The summation term in (2) is convex when $y = -1$. However, when $y = +1$, (2) can be described as the maximum of multiple concave functions and, therefore, is nonconvex. Theoretically, such a nonconvex function has objective local minima, as shown in the first curve of Fig. 2. Once the instance classifier learned with a nonconvex loss function selects false positives, the learning procedure will be misled by them, particularly in early training epochs.

Based on the analysis, we require elaborating the following two problems: 1) how to plausibly relax the nonconvex function so that we can approach the globally optimal solution and 2) how to reasonably perform instance selection so that multiple instances can be collected to optimize the classifier.

Algorithm 1 Parametric Instance Subset Partition

```
Input: Instance bag $x$ and continuation parameter $\lambda$.
Output: Instance subsets $x_{I(\lambda)}$.
while $x \neq \emptyset$ do
    $x_{\hat{i}} = \arg\max_i f(x_{I(\lambda)}, w)$ ;
    $x_{I(\lambda)} = \{x_{\hat{i}}\}$;
    for $k = 1, \ldots, |x|$ do
        if $IoU(x_k, x_{\hat{i}}) > \lambda$ then
            $x_{I(\lambda)} = x_{I(\lambda)} \cup x_k,\ x = x \setminus x_k$
        end if
    end for
    $I \leftarrow I + 1$
end while
```

C. Continuation Optimization

1) Formulation: We propose the continuation optimization method and target at solving above problems in a systematic way. Specifically, we introduce a series of smoothed objective functions $\tilde{\mathcal{L}}(\mathcal{X}, w, \lambda)$, which can approximate MIL's objective function $\mathcal{L}(\mathcal{X}, w)$ [see (2)] but are easier to optimize, Fig. 2. The smoothed objective functions are parameterized by λ. During learning, we trace the smoothed functions from a starting point $(w^0, 0)$ to a solution point $(w^*, 1)$, where w^0 is the solution of $\tilde{\mathcal{L}}(\mathcal{X}, w, 0)$ and w^* the solution of $\tilde{\mathcal{L}}(\mathcal{X}, w, 1)$. This procedure, termed continuation optimization, is formulated as

$$\begin{aligned}\{w^0, \ldots, w^\lambda \ldots, w^*\} = \Big\{&\arg\min_{w^0} \tilde{\mathcal{L}}(\mathcal{X}, w, 0), \ldots,\\ &\arg\min_{w^\lambda} \tilde{\mathcal{L}}(\mathcal{X}, w, \lambda), \ldots,\\ &\arg\min_{w^*} \tilde{\mathcal{L}}(\mathcal{X}, w, 1)\Big\}. \quad (3)\end{aligned}$$

In (3), the first objective function is convex when $\lambda = 0$ and returns to the original objective function [see (2)] when $\lambda = 1$.

2) Parametric Instance Partition: To materialize the smoothed objective functions, we further propose a parametric strategy to partition each instance bag into instance subsets and define smoothed functions within such subsets. Specifically, a bag $x = \{x_1, x_2, \ldots, x_N\}$ is first partitioned into subsets $x = \cup_\lambda x_{I(\lambda)}$ (see Fig. 2). $I(\lambda)$ is the index set for an instance subset. The partition of subsets is controlled by the continuation parameter λ according to the following conditions:

$$\begin{cases}\cup_\lambda x_{I(\lambda)} = x\\ x_{I(\lambda)} \cap x_{I'(\lambda)} = \emptyset \text{ for } \forall I(\lambda) \neq I'(\lambda).\end{cases} \tag{4}$$

According to (4), all subsets for a bag x are minimum sufficient cover to the bag. The subsets should have the properties that, when $\lambda = 0$, a bag x is partitioned into a single subset that includes all instances; when $\lambda = 1$, the bag x is partitioned into multiple subsets, each of which contains a single instance; when λ continuously increases from 0 to 1, the subset gradually dwindles from the instance bag to a single instance.

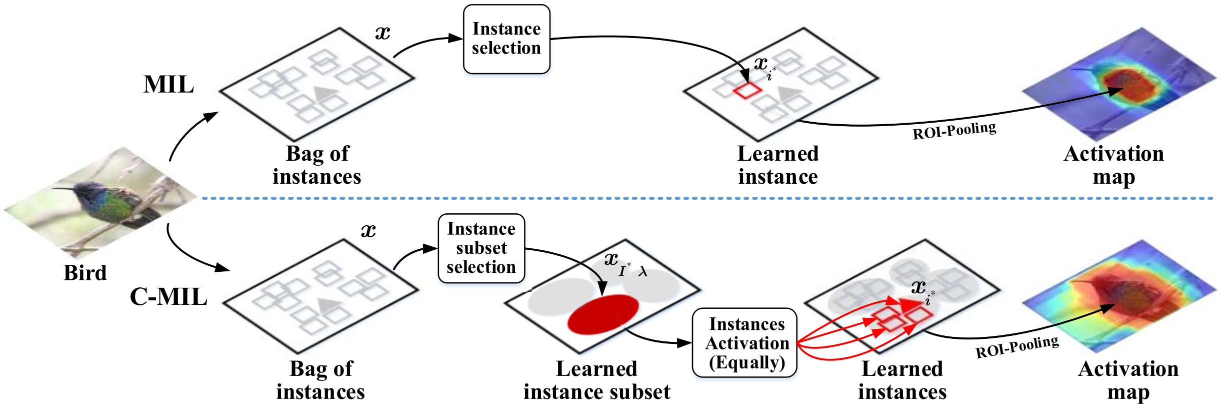


Fig. 3. Comparison of selected instances by C-MIL and MIL. With MIL, the discriminative instance is selected, and thereby, an object part is activated. With C-MIL, the discriminative instance subset is selected, and the full object extent is activated. (Best viewed in color.)

3) Smoothed Objective Function: Based on subset partition, an instance subset selection function, $\mathcal{C}(x_i, w, \lambda)$, is defined to replace the second $\max(\cdot)$ in (2). The objective function defined in (2) is accordingly updated to a smoothed one as

$$\begin{aligned}\{w^*, x_{I^*(\lambda)}\} &= \arg\min_{w, x_{I(\lambda)}} \tilde{\mathcal{L}}(\mathcal{X}, w, \lambda)\\ &= \arg\min_{w} \sum_{x} \max(0, 1 - y\mathcal{C}(x, w, \lambda)) \quad (5)\end{aligned}$$

where $\mathcal{C}(x, w, \lambda)$ is defined as

$$\mathcal{C}(x, w, \lambda) = \mathcal{S}\big(\{f(x_{I(\lambda)}, w) | x_{I(\lambda)} \subseteq x\}\big) \quad (6)$$

where $f(x_{I(\lambda)}, w)$ calculates the scores of all instances in a subset $x_{I(\lambda)}$ and average them to a subset score. $S(\cdot)$ is a pooling operation, *e.g.*, average pooling, which converts instance subset scores to image classification scores. The pooling operation $S(\cdot)$ works in a way like median filtering, with which the function defined by (5) is smoother than that by (2).

4) Model Optimization: Equation (3) defines a series of smoothed objective functions for continuation optimization. In the deep learning framework, each function is differential and can be optimized with the SGD algorithm. When optimizing the model parameter w, the continuation parameter λ changes from 0 to 1, and the smoothed objective functions are optimized one by one. When $\lambda = 0$, all instances in the bag are partitioned into a single subset, and $\mathcal{C}(\cdot)$ calculates the average scores of all instances in a bag. Accordingly, (5) is the maximum of two linear terms and, therefore, is convex and most smoothed. When $\lambda = 1$, each subset contains a single instance, and $\mathcal{C}(\cdot)$ outputs the score of the instance. In this case, (5) is not smoothed, i.e., it deteriorates to MIL's objective function (see Fig. 2).

IV. Weakly Supervised Object Detection

WSOD defines a task to learn object detector, while, solely, image category labels are available. During the training procedure, images are treated as bags, and region proposals generated by Selective Search [43] are treated as instances. The instances selected by C-MIL are used as pseudo-ground truths to learn an object detector.

A. Instance Selection With C-MIL

All the instances (object proposals) are partitioned into subsets according to the parametric partition strategy. To fulfill this purpose, we first sort the instances in each bag using their classification scores $f(x_i, w)$. The following two steps (detailed in Algorithm 1) are iteratively performed to define instance subsets: 1) from the instance set where instances have not been partitioned into any subset, select the top-scored instance x_{i^*} and construct an instance subset with it and 2) from the instance set where instances have not been partitioned into any subset, select the instances whose similarity with x_{i^*} is not less than the continuation parameter λ and include it into the subset constructed in step 1).

Given a continuation parameter $0 \leq \lambda \leq 1$, the smoothed function $\mathcal{C}(x, w, \lambda)$ for instance selection is specified as

$$\mathcal{C}(x, w, \lambda) = \max_{I(\lambda)} f(x_{I(\lambda)}, w). \quad (7)$$

Accordingly, the continuation loss function defined in (5) is specified as

$$\begin{aligned}\{w^*, x_{I^*(\lambda)}\} &= \arg\min_{w, x_{I(\lambda)}} \tilde{\mathcal{L}}(\mathcal{X}, w, \lambda)\\ &= \arg\min_{w} \sum_{x} \max(0, 1 - y_i \max_{I(\lambda)} f(x_{I(\lambda)}, w)).\end{aligned} \quad (8)$$

To optimize the objective/loss function, C-MIL activates instances in the subset $x_{I(\lambda)}$ equally to learn the model parameters. As mentioned in Section III-C2 that the instances in a subset are spatially close to each other, C-MIL is able to collect object parts to activate the full object regions (see Fig. 3). The continuation parameter λ increases from 0 to 1 during optimization. When it equals to 0, each bag x contains only one subset that includes all instances. The term $\max_i f(x_i, w)$ of (2) then can be simplified to $\sum_i f(x_i, w)$. Accordingly, this term becomes convex, and therefore, (2) is convex. When λ equals to 1, the number of subsets equals to that of instances in the bag, i.e., each subset includes only one instance. Therefore, (5) returns back to the loss function of MIL [see (2)]. When $0 < \lambda < 1$, each subset contains multiple instances, and

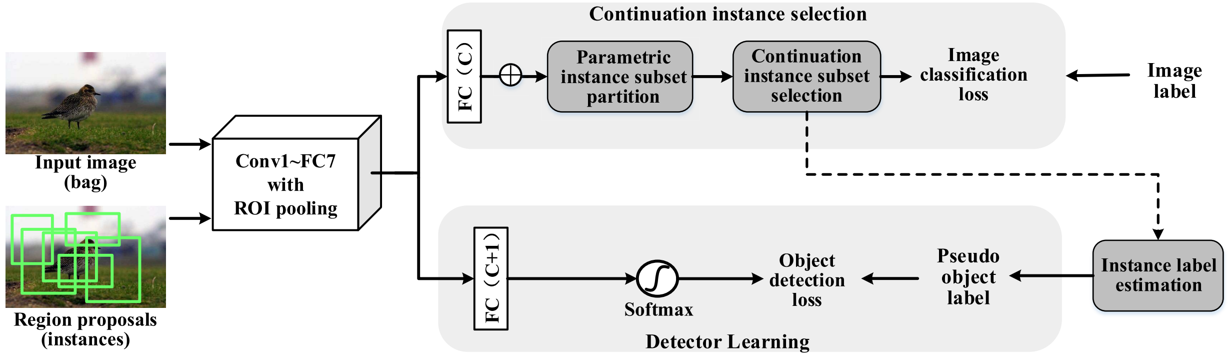


Fig. 4. Implementation of WSOD with C-MIL, where continuation instance selection is implemented atop a deep network. In the forward procedure, C-MIL selects positive instances from subsets and uses them as pseudo-objects for detector estimation. In the backpropagation procedure, the instance selector and object detectors are jointly optimized. C is the number of object categories.

the loss function falls between the smoothed function and the original loss function (see Fig. 2).

B. Detector Training

WSOD is implemented with a two-branch network structure (see Fig. 4), where Selective Search [43] is employed to extract region proposals (instances) for each image. Following Fast RCNN [10], convolutional layers are used to extract image features, while an ROI-Pooling layer and two fully connected layers are added atop the convolutional layers to extracted features for instances. During detector training, continuation instance selection is performed by C-MIL implemented atop the last convolutional layer. Selected positive/negative instances are used to learn an object detector.

For WSOD, we propose a continuation strategy to predict reliable instances while learning detectors. Specifically, the instances of each bag are partitioned into positives and negatives according to the continuation parameter λ. For the selected instance subset $x_{i(\lambda)^*}$ and the top scored instance x_{i^*} in it, instances in the bag are labeled as positives or negatives based on their spatial distances, as $y_i = +1$ if $IoU(x_i, x_{i^*}) \geq 1 - \lambda/2$ or $y_i = -1$ if $IoU(x_i, x_{i^*}) \leq \lambda/2$, where IoU computes the Intersection of Union over two instances.

During detector training, when λ changes from 0 to 1, the threshold $1 - \lambda/2$ decreases from 1 to 0.5, while the threshold $\lambda/2$ increases from 0 to 0.5.

With the change of λ, more and more instances are labeled as positives/negatives by the detector $g_z(x_i, w_g)$, which gradually learned with the cross-entropy loss. In the training procedure, the instance selector and object detector are optimized with network propagation, along with network parameters updated iteratively (see Fig. 4).

During inference, the learned detector is used to predict the scores for each instance, and nonmaximum suppression (NMS) is performed to remove overlapping instances.

V. Supervised Object Detection

Beyond WSOD, C-MIL can also be applied to improve the detection performance of supervised object detectors. In what follows, we first revisit RetinaNet [44], a representative one-stage detection method. We then apply C-MIL to ReinaNet for anchor/feature selection during detector training.

A. RetinaNet Revisit

A RetinaNet detector is made up of a backbone network and two subnets: one for object classification and the other for object localization. The feature pyramid network (FPN) is used as the backbone network for feature extraction. From each feature map in the feature pyramid, a classification subnet predicts category probabilities, while a box regression subnet predicts object locations using anchor boxes as the reference locations. Each anchor corresponds to a feature vector on the convolutional feature map. Considering the extreme imbalance of foreground-background classes, presented as positive–negative anchors after anchor-object matching, the focal loss [44] is adopted to prevent the vast number of easy negatives from overwhelming the detector during training.

For a special class of object, let x_i denote an anchor in a training image. The label $y_i \in \{+1, -1\}$ of an anchor is empirically determined according to its overlap with an object. An anchor is positive, i.e., $y_i = +1$, if its IoU with a ground-truth bounding-box is larger than a threshold (e.g., 0.5). It is a negative anchor, i.e., $y_i = -1$, when the IoU is smaller than a threshold (e.g., 0.4). The remaining anchors are ignored. The labeled anchors are then used to supervise detection network training as

$$\begin{aligned} w^* &= \arg\min_{w, x_i} \mathcal{L}(\mathcal{X}, w) \\ &= \arg\min_{w, x_i} \sum_x \sum_i \mathcal{L}^c(x_i, w) + \delta(y_i)\mathcal{L}^l(x_i, w) \end{aligned} \tag{9}$$

where $\mathcal{L}^c(\cdot)$ and $\mathcal{L}^l(\cdot)$, respectively, denote the classification loss and localization (bounding-box regression) loss. In RetinaNet, $\mathcal{L}^c(\cdot)$ is defined as a focal loss (FL) and $\mathcal{L}^l$ Smooth-L1 loss(SL) [44]. $\delta(y) = 1$ if $y = +1$; otherwise, $\delta(y) = 0$, which means that the second (localization) term of (9) is only valid for positive anchors.

During network training, each anchor independently supervises the learning for object classification and object localization, without considering whether classification and

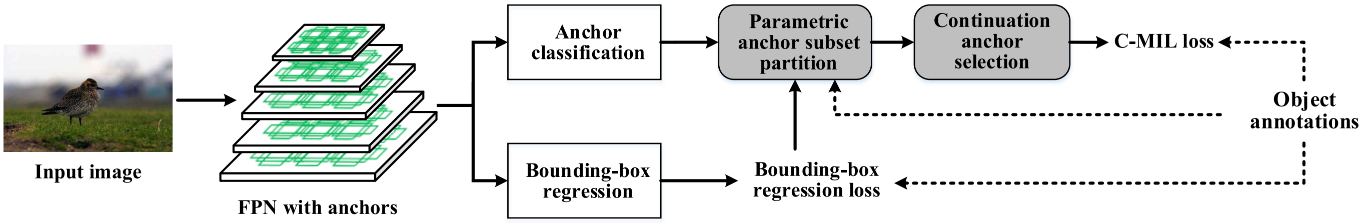


Fig. 5. Implementation of supervised object detection with C-MIL, where anchors are partitioned into subsets and continuation anchor selection is implemented atop a deep network.

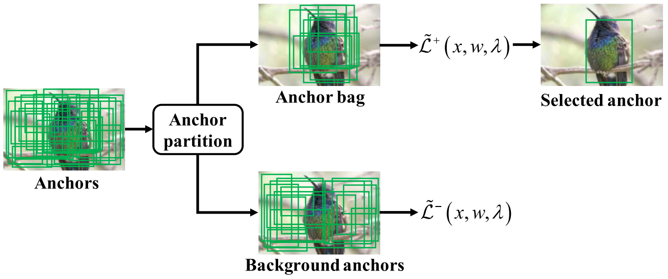

Anchors
Anchor partition
Anchor bag
$\tilde{\mathcal{L}}^{+}(x,w,\lambda)$
Selected anchor
Background anchors
$\tilde{\mathcal{L}}^{-}(x,w,\lambda)$

Fig. 6. Illustration of the anchor bag partition when applying C-MIL to supervised object detection.

localization modules are compatible (simultaneously achieve high scores) on assigned anchors. This could cause failure detections when anchors of accurate localization but lower classification confidence are suppressed by the following NMS procedure.

B. Anchor Selection With C-MIL

To alleviate the drawbacks of independent anchor/feature optimization, we propose using C-MIL to select an optimal anchor for each ground-truth object when training detectors. The improved detector is referred to as RetinaNet-C-MIL, as shown in Fig. 5.

To train RetinaNet-C-MIL, we first construct an anchor bag x for each object. Fig. 6 illustrates the process of anchor bag (instance subset or clique) partition when applying C-MIL to supervised object detection. The anchors are first partitioned into anchor bags or background anchors according to their IoUs with the ground-truth boxes. Specifically, the anchors whose IoUs with the closest ground-truth box b are larger than a threshold (top-k) will be partitioned into the anchor bag of b, and the anchors that do not belong to any anchor bags are partitioned into background anchors. To evaluate the localization confidence, we define the localization loss as $f^{l}\big(x_{I(\lambda)}, w\big) = e^{\mathcal{L}^{l}(x_i, w)}$. During training, we evaluate the classification and localization scores, $f^{c}(x_i, w)$ and $f^{l}(x_i, w)$, of each anchor bag x. The detection score $f(x_i, w)$ for each anchor is computed as $f(x_i, w) = f^{c}(x_i, w) \times f^{l}(x_i, w)$. Such scores are used for anchor selection, as well as guiding the calculation of classification and localization loss (see Fig. 5). According to (6), we specify the following continuation function for anchor selection as

$$\mathcal{C}(x, w, \lambda) = \frac{1}{|x|} \sum_{x_{I(\lambda)} \in x} f\big(x_{I(\lambda)}, w\big) \tag{10}$$

and define the continuation anchor selection procedure as

$$\begin{aligned} \{w^{*}, x_{I^{*}(\lambda)}\} &= \mathop{\arg\min}_{w, x_{I(\lambda)}} \tilde{\mathcal{L}}(\mathcal{X}, w, \lambda) \\ &= \mathop{\arg\min}_{w, x_{I(\lambda)}} \sum_{x} \tilde{\mathcal{L}}^{+}(x, w, \lambda) + \tilde{\mathcal{L}}^{-}(x, w, \lambda) \end{aligned} \tag{11}$$

where

$$\tilde{\mathcal{L}}^{+}(x, w, \lambda) = \max(0, 1 - y \frac{1}{|x|} \sum_{I(\lambda)} f\big(x_{I(\lambda)}, w\big) \tag{12}$$

and

$$\tilde{\mathcal{L}}^{-}(x, w, \lambda) = FL\big(f^{c}(x_i, w), y_i | x_i \in x^{-}\big) \tag{13}$$

where x^{-} contains the anchors that do not belong to any positive bags. $y_i = -1$ for all $x_i \in x^{-}$. As illustrated in Fig. 6, the optimal anchor for object detection will be selected by (12), while the background anchors are suppressed by (13).

C. Detector Training

According to the idea of continuation optimization, we gradually change λ from 0 to 1 during detector training. When $\lambda = 0$, all anchors in an anchor bag are used to classify and localize an object. When $\lambda = 1$, the size of anchor bags reduces to a single anchor, and an optimal anchor is selected for the object. During the feedforward procedure, we calculate the classification and localization scores of each anchor with $f^{c}(\cdot)$ and $f^{l}(\cdot)$. A set of anchors (or a single optimal anchor) is selected to minimize the loss defined in (11). During the backpropagation procedure, network parameters are updated using an SGD algorithm under the supervision of selected anchors.

The testing procedure of RetinaNet-C-MIL is exactly the same as RetinaNet. We use the learned network parameters to predict classification scores and object bounding boxes, which are fed to an NMS procedure for object detection. As C-MIL is only applied in the detector training procedure, RetinaNet-C-MIL has negligible computation cost overhead.

VI. Experiments

As a general method for instance section and model learning, C-MIL was validated on WSOD and supervised object detection. In each task, we first introduced experimental settings and then validated the proposed C-MIL method. We also reported the performance of detectors learned by C-MIL and compared them with state-of-the-art detectors.

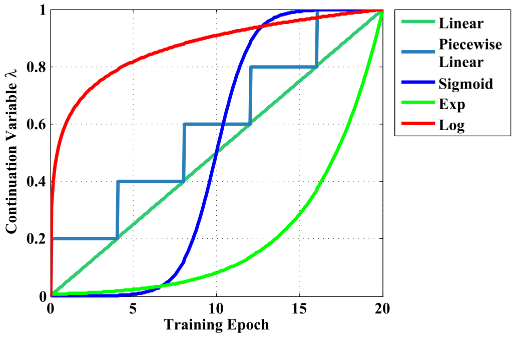


Fig. 7. Five functions defined to calculate continuation parameter.

A. Weakly Supervised Object Detection

For WSOD, we evaluate C-MIL on the PASCAL VOC 2007 and 2012, and MS-COCO 2014 data sets. We use mean average precision (mAP) [45] and correct localization (CorLoc) [46] as the evaluation metrics.

1) Experimental Settings:

a) Data Sets: The PASCAL VOC 2007 data set contains 9963 images of 20 object categories that are split into two sets: 5011 for *trainval* and 4952 for *test*. The PASCAL VOC 2012 data set is made up of 22 531 images of 20 object categories that are split into two subsets: 11 540 for *trainval* and 10 991 for *test*. The MS-COCO 2014 data set has 80 object categories, with challenging aspects including dense and small and occluded objects.

b) CNN models: We implement C-MIL with two CNN models (VGGF and VGG16) that are pretrained on the ILSVRC 2012 data set. VGGF [47] contains five convolutional layers and three fully connected layers, while VGG16 [20] contains 13 convolutional layers and three fully connected layers. For both VGGF and VGG16, we replaced the last max pooling layer with the ROI-pooling layer, as in [10]. We remove the FC8 layer in both CNN models and add the C-MIL module.

c) Object proposals: We used Selective Search [43] to extract about 2000 object proposals from each image. As in Fast RCNN, the fast setting was adopted to generate proposals. Small proposals whose width or height is less than 20 pixels were removed.

d) Learning settings: We resized the input images into five scales {480, 576, 688, 864, 1200} with respect to their larger side (height or width), as in [11], [12], [14], and [15]. For a training image, it was randomly resized to one of the scales, while it was randomly flipped. Accordingly, during the test, each image was augmented into ten images. The output scores of each proposal from the ten augmented images were averaged. For recurrent learning, SGD was used with a momentum of 0.9, a weight decay of 5e-4, and a batch size of 1. The model iterated 20 epochs where the learning rate was 5e-3 for the first 15 epochs and 5e-4 for the last five epochs. The NMS method used in this article follows existing works [10], [48]. The score and IoU thresholds of NMS are respectively set as 0.005 and 0.3.

e) Baseline: The implementation of WSOD is illustrated in Fig. 4, where two network branches, respectively, perform image classification and object detection. Given region proposals as inputs, the classification branch estimates the image classification score and instance confidence using C-MIL defined in Section III. The instance confidence is then transferred to the object detection and mask prediction branches in a feed-forward manner for object detector training (see Section III).

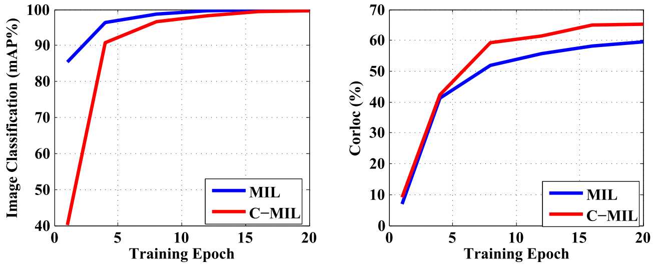


Fig. 8. Change of image classification and object localization performance on the *trainval* set of the PASCAL VOC 2007 data set with VGG16 during training. In the early epochs, MIL achieves higher classification performance. In the later epochs, the classification performance of C-MIL catches up with that of MIL, and localization performance becomes higher than that of MIL.

TABLE I

COMPARISON OF FUNCTIONS TO CALCULATE THE CONTINUATION PARAMETER λ. DETECTION AND LOCALIZATION PERFORMANCE (%) ON THE PASCAL VOC 2007 DATA SET WITH VGGF

Method	Approaches / Continuation Functions	mAP	CorLoc
MIL	ContextNet [14]	36.0	55.0
C-MIL (Ours)	Linear	37.9	58.9
	Piecewise Constant	37.6	57.4
	Sigmoid	38.3	58.4
	Exp	37.1	56.4
	Log	**40.7**	**59.5**

2) Continuation Method: On the VOC 2007 data set, we investigated how to control the parameter λ for continuation optimization on instance selection and detector estimation.

a) Continuation parameter: To implement continuation optimization during training, "Linear," "Piecewise Constant," "Sigmod,""Exp," and "Log" functions (see Fig. 7) were used to generate continuation parameter λ. λ monotonically increases according to either of the defined functions, while the instance subsets gradually dwindle to a single instance according to Algorithm 1. As shown in Table I, with continuation optimization, the detection and localization performance were improved by 1.1%–4.7% and 1.4%–4.5%, respectively, which fully demonstrated the effectiveness of the proposed parametric continuation strategy.

Table I indicates that the "Log" function achieved the best detection mAP and localization CorLoc. When using the "Log," λ increased rapidly in the initial epochs and slowed down in the last epochs (see Fig. 7). This matches the training procedure: in the initial epochs, the instance subsets are large so that various object parts can be collected and fully utilized to fine-tune the network; in the later epochs, the instance subsets become stable, and it required to meticulously select instance for detector estimation.

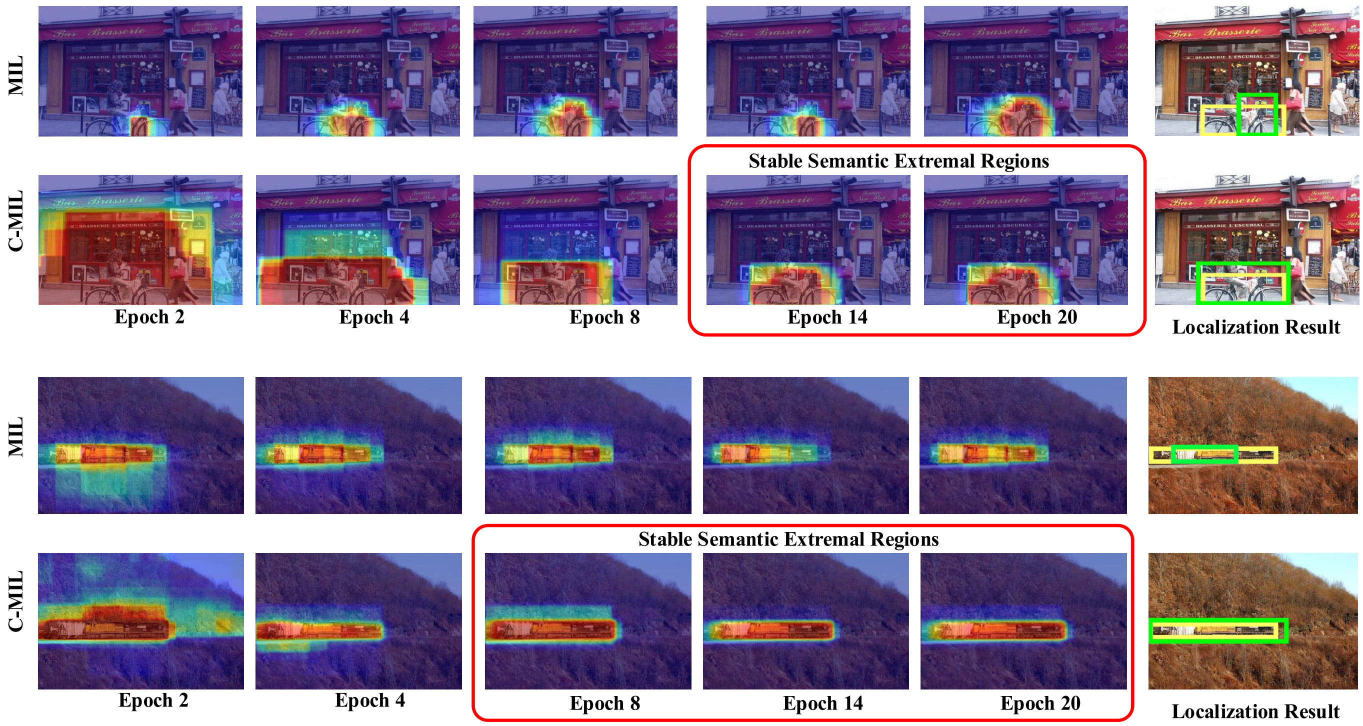


Fig. 9. Illustration of Stable Semantic Extremal Regions (SSERs). MIL activated the discriminative regions but missed the full object extent. In contrast, C-MIL discovered SSERs indicating full object extent. The continuation parameter λ of C-MIL increased from 0 to 1 along with the training procedure (from epoch 0 to epoch 20). In the last column, the yellow and green boxes, respectively, denote ground truths and localization results. The heat maps are calculated by accumulating the scores of region proposals. (Best viewed in color.)

TABLE II

Ablation Studies of C-MIL. Detection Performance (mAP%) on the Pascal VOC 2007 Data Set With VGGF

Method	Instance Selector	Object Detector	mAP
MIL [14]	-	-	36.0
C-MIL (Ours)	✓		39.0
		✓	37.4
	✓	✓	**40.7**

b) Continuation optimization: The ablation experimental results for continuation instance selection and detector estimation are shown in Table II. The usage of the continuation instance selection improved the performance of baseline by 3.0% (from 36.0% to 39.0%); the usage of the continuation detector estimation further improved the performance of baseline by 1.4% (from 36.0% to 37.4%). Combining two modules improved the performance by 4.7% (from 36.0% to 40.7%), which clearly showed the effectiveness of continuation optimization.

Fig. 8 shows the visualization of the evolution of the image classification and object localization performance during training. In the initial training epochs, the performance of MIL is higher than that of C-MIL. In the later epochs, the classification performance gap between MIL and C-MIL gradually decreased, while the localization performance of C-MIL became higher than that of MIL. This is because the main loss of MIL is image classification loss, while it did not optimize for object localization. Consequently, it tended to select object proposals that were discriminative for image classification but missed to localize full objects. In contrast, C-MIL optimized both image classification and object localization by learning instance subsets, where instances are spatially related and class-related. C-MIL was able to avoid the optimization getting stuck in local minima.

c) Training time: We test the training time of the baseline and the proposed C-MIL on VOC 2007 with VGG16 on an NVIDIA GTX 1080Ti GPU. It, respectively, takes the baseline 8.1 h and C-MIL 8.7 h for training. Compared with the baseline method, the computational cost of C-MIL is moderate considering the challenging aspects of WSOD and the significant performance improvement.

3) Stable Semantic Extremal Regions: To analyze and understand the continuation optimization, we visualized the activation of learned subsets during the training procedure (see Fig. 9). It shows that, when λ increases from 0 to 1, the activated region of instance subsets gradually dwindles. In the initial training epochs, subsets were defined as large to collect as many objects/parts as possible. In the later training epochs, the region activated by the subsets stopped dwindling and tended to form stable activation regions around object boundaries. We termed these regions stable semantic extremal regions (SSERs), which, often, turns out to be full object extent.

The emergence of SSERs indicated that C-MIL continuously suppressed backgrounds while activating object regions during learning. Since the semantics inside and outside the object boundary are not continuous, the instance

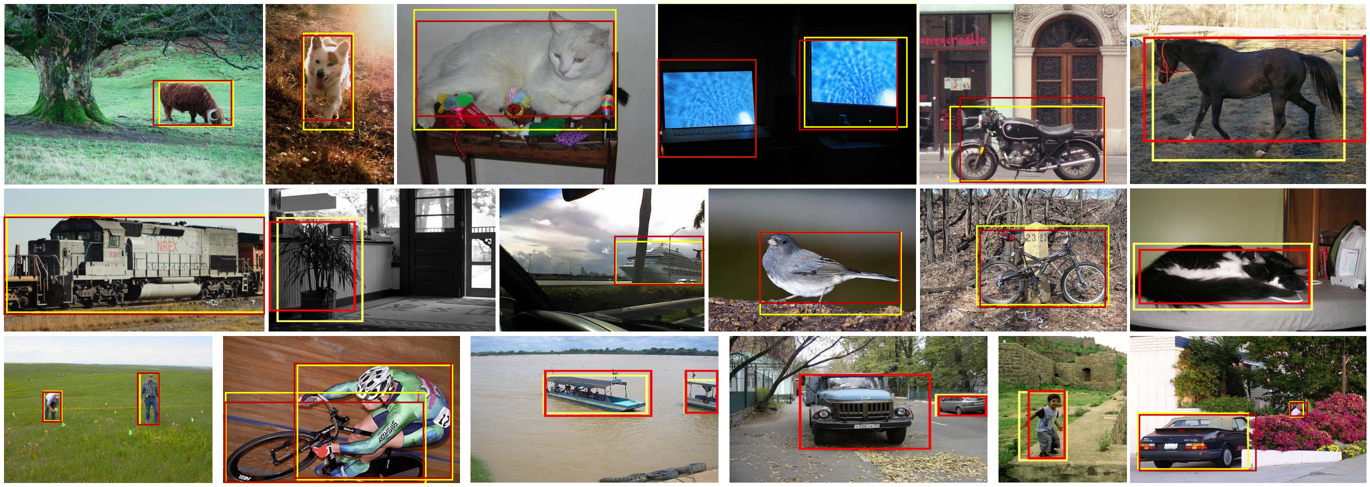


Fig. 10. Object detection examples on the PASCAL VOC 2007 data sets. Yellow bounding boxes denote ground-truth annotations, and red boxes denote detection results. (Best viewed in color.)

TABLE III

DETECTION PERFORMANCE (%) ON THE VOC 2007 TEST SET. COMPARISON OF C-MIL WITH THE STATE OF THE ARTS

Network	Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
VGGF/ AlexNet	MILinear [23]	41.3	39.7	22.1	9.5	3.9	41.0	45.0	19.1	1.0	34.0	16.0	21.3	32.5	43.4	21.9	19.7	21.5	22.3	36.0	18.0	25.4
	MF MIL [25]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
	PDA [18]	49.7	33.6	30.8	19.9	13.0	40.5	54.3	37.4	**14.8**	39.8	9.4	28.8	38.1	49.8	14.5	**24.0**	27.1	12.1	42.3	39.7	31.0
	LCL+Context [22]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
	WSDDN [11]	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.5
	ContextNet [14]	**57.1**	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	**49.2**	**42.0**	47.3	56.6	15.3	12.8	24.8	**48.9**	44.4	47.8	36.3
	WCCN [15]	43.9	**57.6**	**34.9**	**21.3**	14.7	**64.7**	52.8	34.2	6.5	41.2	20.5	33.8	47.6	56.8	12.7	18.8	**39.6**	46.9	52.9	45.1	37.3
	OICR [12]	53.1	57.1	32.4	12.3	15.8	58.2	56.7	39.6	0.9	44.8	39.9	31.0	**54.0**	62.4	4.5	20.6	39.2	38.1	48.9	48.6	37.9
	MELM [7]	56.4	54.7	30.9	21.1	**17.3**	52.8	**60.0**	36.1	3.9	**47.8**	35.5	28.9	30.9	61.0	5.8	22.8	38.8	39.6	42.1	**54.8**	38.4
	C-MIL (Ours)	54.5	55.5	34.4	20.3	16.7	53.4	59.2	**44.6**	8.4	46.0	40.2	40.8	47.7	**63.2**	**22.8**	23.2	39.4	44.3	**53.8**	52.3	**40.7**
VGG16	WSDDN [11]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
	PDA [18]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
	OICR [12]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
	Self-Taught [32]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
	WCCN [15]	49.5	60.6	38.6	29.2	16.2	70.8	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
	TS^{2}C [8]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
	WeakRPN [49]	57.9	**70.5**	37.8	5.7	**21.0**	66.1	**69.2**	59.4	3.4	57.1	**57.3**	35.2	**64.2**	68.6	**32.8**	28.6	50.8	49.5	41.1	30.0	45.3
	PCL [31]	57.1	67.1	40.9	16.9	18.8	65.1	63.7	45.3	17.0	56.7	48.9	33.2	54.4	68.3	16.8	25.7	45.8	52.2	59.1	62.0	45.8
	MELM [7]	55.6	66.9	34.2	29.1	16.4	68.8	68.1	43.0	**25.0**	**65.6**	45.3	53.2	49.6	68.6	2.0	25.4	52.5	56.8	62.1	57.1	47.3
	C-MIL (Ours)	**62.5**	58.4	**49.5**	**32.1**	19.8	70.5	66.1	**63.4**	20.0	60.5	52.9	**53.5**	57.4	**68.9**	8.4	24.6	51.8	**58.7**	**66.7**	**63.5**	**50.5**
FRCNN Re-train	OICR-Ens. [12]	**65.5**	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
	PCL-Ens. [31]	63.2	69.9	47.9	22.6	27.3	71.0	69.1	49.6	12.0	60.1	51.5	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	48.8
	WeakRPN-Ens. [49]	63.0	**69.7**	40.8	11.6	**27.7**	**70.5**	**74.1**	58.5	10.0	**66.7**	**60.6**	34.7	**75.7**	**70.3**	**25.7**	**26.5**	**55.4**	56.4	55.5	54.9	50.4
	C-MIL (Ours)	61.8	60.9	**56.2**	**28.9**	18.9	68.2	69.6	**71.4**	**18.5**	64.3	57.2	**66.9**	65.9	65.7	13.8	22.9	54.1	**61.9**	**68.2**	**66.1**	**53.1**

subsets gradually dwindle to eliminate the backgrounds until it reaches the object boundary. The procedure is related to that of maximally stable extremal regions (MSERs) [50] in a way. The difference is that the MSERs are defined for gray-level stable regions while SSERs for semantic stable regions.

4) Performance:

a) Pascal VOC: Table III shows the performance of C-MIL and a comparison with the SOTA methods on the PASCAL VOC 2007 data set. It shows that C-MIL achieved 40.7% and 50.5% detection performances with the VGGF and VGG16 models, respectively. With VGGF, C-MIL, respectively, outperformed WCCN [15], OICR [12], and MELM [7] by 3.4% (from 37.3% to 40.7%), 2.8% (from 37.9% to 40.7%), and 2.3% (from 38.4% to 40.7%). With VGG16, it outperformed the SOTA WeakRPN [49], PCL [31], and MELM [7] approaches by 6.2% (from 44.3% to 50.5%), 4.7% (from 45.8% to 50.5%), and 3.2% (from 47.3% to 50.5%) respectively, which were significant margins for the challenging WSOD task. Detector examples by C-MIL are shown in Fig. 10.

We further retrained a Fast-RCNN detector using the learned pseudo-objects as the ground truth and achieved 53.1% mAP (see Table III), which outperformed the state-of-the-art OICR-Ens., PCL-Ens., and WeakRPN-Ens. by 6.1% (from 47.0%

TABLE IV

DETECTION AND LOCALIZATION PERFORMANCE (%) ON THE VOC 2012 DATA SET USING VGG16. COMPARISON OF C-MIL WITH THE STATE OF THE ARTS

Method	mAP	CorLoc
WCCN [15]	37.9	-
Self-Taught [32]	38.3	58.8
OICR [12]	37.9	62.1
PCL [31]	40.6	63.2
TS^{2}C [8]	40.0	64.4
WeakRPN [49]	40.8	64.9
MELM [7]	42.4	-
C-MIL (Ours)	**46.7**	**67.4**

TABLE V

LOCALIZATION PERFORMANCE (%) ON THE VOC 2007 *trainval* SET. COMPARISON OF C-MIL WITH THE STATE OF THE ARTS

CNN	Method	mAP
VGG16	WSDDN [11]	53.5
	WCCN [15]	56.7
	OICR [12]	60.6
	TS^{2}C [8]	61.0
	PCL [31]	63.0
	WeakRPN [49]	63.8
	C-MIL (Ours)	**65.0**

to 53.1%), 4.1% (from 48.8% to 53.1%), and 2.7% (from 50.4% to 53.1%). Specifically, the detection performance for "aeroplane" (+3.2%), "bird" (+5.8%), "cat" (+3.5%), and "train" (+4.5%) significantly improved over the WeakRPN-Ens. approach. Table IV is the detection results of C-MIL and SOTA methods on the VOC 2012 data set with VGG16. For object detection, C-MIL, respectively, outperformed the WeakRPN [49], PCL [31], and MELM [7] by 5.9% (from 40.8% to 46.7%), 6.1% (from 40.6% to 46.7%), and 4.3% (from 42.4% to 46.7%).

Tables IV and V show the object localization performance of C-MIL and comparisons with the SOTA methods. It shows that C-MIL, respectively, outperformed the WeakRPN [49] and PCL [31] by 2.5% (from 64.9% to 67.4%) and 3.0% (from 64.4% to 67.4%) on VOC 2012 1.2% (from 63.8% to 65.0%) and 2.0% (from 63.0% to 65.0%) on VOC 2007.

b) MS-COCO: To validate the effectiveness of C-MIL on a large-scale data set, we conducted experiments on MS-COCO 2014 and reported the results in Table VI. It can be seen that C-MIL with a VGG16 network significantly outperformed the MIL-based approach (WSDDN [11]). With C-MIL, we set a solid baseline for weakly supervised object detection on the large-scale MS-COCO data set. On the other hand, we much realize that the detection performance on the MS-COCO data set remains low. The reasons are twofold: 1) the region proposals have a very low recall rate of 57%, which means that more than 40% objects are missed in this step and 2) objects in MS-COCO are much smaller on average than those in VOC 2007, which poses additional challenges to object detectors.

TABLE VI

DETECTION PERFORMANCE (%) ON MS-COCO 2014

Method	CNN	mAP@.5	mAP@[.5,.95]
MIL [11]	VGGF	10.1	3.1
C-MIL (Ours)	VGGF	11.9	4.1
	VGG16	**18.8**	**7.8**

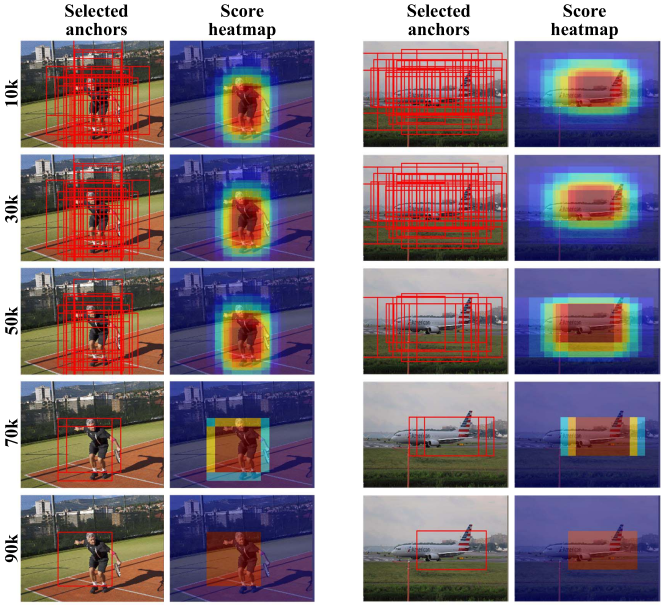


Fig. 11. Evolution of continuation anchor selection when training a supervised object detector. In the first iteration, all anchors/features in a bag have similar scores and are all selected. When training proceeds, some anchors output higher scores than others and are selected. In the last iteration, a single top-scored anchor is selected in each bag. The heat map is calculated by summarizing anchor scores. (Best viewed in color.)

B. Supervised Object Detection

For supervised object detection, the PASCAL VOC 2007 data set was used for the ablation study, while the MS-COCO object detection data set was used for performance comparison. For all the experiments, average precision (AP) [45] was used as the evaluation metric.

1) Experimental Settings: We utilized ResNet-50 and ResNet-101 with FPN as backbone networks. The VOC 2007 *trainval* and VOC 2012 *trainval* sets were used to train detectors and VOC 2007 *test* for evaluation. The detectors were trained in a single GPU with a batch size of 4 and an image size of 500. The initial learning rate was set to 0.005 and decreased by a factor of 10 after 30k and 40k for the 45k setting. For the MS-COCO data set, the *train* set was used to train a detector, and the *test-dev* set is used for testing. We trained the model with eight GPUs. Each GPU contained a minibatch of two images with a size of 800. The initial learning rate was set to 0.01 and decreased by a factor of 0.1 after 60k and 80k for the 90k setting. The score and IoU thresholds of NMS are, respectively, set as 0.05 and 0.5, which are the same as those in RetinaNet [44]. The anchor generation settings were the same as those of RetinaNet [44], i.e., nine anchors with three sizes $\{2^0, 2^{1/3}, 2^{2/3}\}$ and three aspect ratios $\{1:2, 1:1, 2:1\}$ for each pixel on the feature maps. Across

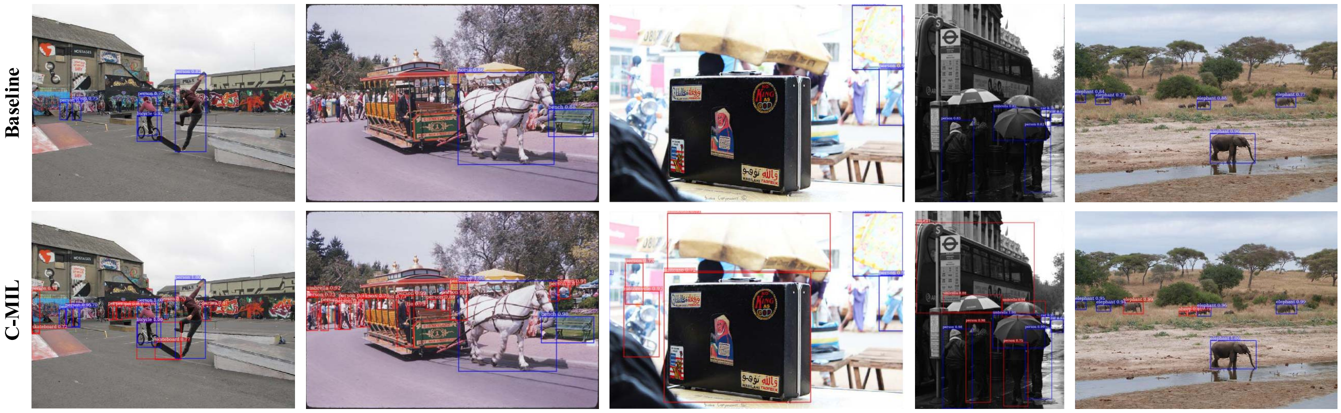


Fig. 12. Detection examples on MS-COCO 2017. Blue boxes denote objects detected by both RetinaNet and RetinaNet-C-MIL. Red boxes denote objects detected by RetinaNet-C-MIL but missed by RetinaNet. RetinaNet-C-MIL detected more slender objects and objects of occlusion. (Best viewed in color.)

TABLE VII
DETECTION PERFORMANCE (%) ON PASCAL 2007 WITH RESNET-50

Detector	AP	AP@0.5	AP@0.75	AP$_S$	AP$_M$	AP$_L$
RetinaNet (baseline) [44]	42.4	70.3	43.3	11.3	27.4	51.5
RetinaNet-MIL (Ours)	44.7	71.3	47.0	11.8	29.3	54.0
RetinaNet-C-MIL* (Ours)	47.6	73.3	51.0	11.1	30.2	57.9
RetinaNet-C-MIL (Ours)	**49.5**	**74.5**	**53.5**	**13.1**	**31.8**	**60.0**

feature levels, the anchors cover the scale range from 32 to 813 pixels with respect to the input image. The synchronized stochastic gradient descent (SGD) was adopted for network optimization. The weight decay of 0.0001 and the momentum of 0.9 are used. A linear warm-up strategy was adopted in the first 500 iterations. We set the regularization factor of positive instances as 0.75 experimentally. To determine anchor number (k) for each anchor bag, we empirically tested $k = 40, 50,$ and 60 and achieved 48.9%, 49.5%, and 49.2% APs, respectively. We, thus, choose 50 anchors in the following experiments.

2) Continuation Anchor Selection: In Table VII, we tested the detection performance by using MIL to select anchors. RetinaNet-MIL improved the AP from 42.4% to 44.7%, which validated that anchor selection can optimize feature representation for detection. When using C-MIL to select anchors, RetinaNet-C-MIL further improved the AP from 44.7% to 49.5%.[2] The large performance gain validated the effectiveness of C-MIL for continuation anchor selection. The nature behind the good performance is that C-MIL defines a learning-to-match mechanism for feature-object correspondence, which pursued optimal features to explain a class of objects in terms of both classification and localization. As shown in Fig. 11, in the initial iterations, all anchors/features in a bag had similar scores and were all selected. When training proceeded, some anchors of high scores were selected. In the last iteration, a single top-scored anchor was selected in each bag.

3) Performance: On VOC 2007, with a ResNet-50 backbone, RetinaNet-C-MIL improved the baseline from 42.4% to 45.9% with 3.5% performance gain (see Table VII). On MS-COCO, with a ResNet-50 backbone, RetinaNet-C-MIL improved the baseline from 35.7% to 38.8% with 3.1% performance gain (see Table VIII), which is a significant margin for the challenging object detection task. Comparison of detection results in Fig. 12 shows that RetinaNet-C-MIL detected more slender objects and objects of occlusion. The most representative features of such objects often bias from their geometric centers, which challenged the IoU-based anchor assignment in RetinaNet but can be well handled by the instance (anchor) selection mechanism defined in C-MIL.

In Table VIII, RetinaNet-C-MIL was compared with the state-of-the-art one- and two-stage detectors on the MS-COCO *test-dev* set. For a fair comparison, we rescaled the images so that their shorter sides are 800 pixels and the longer sides not more than 1333 pixels. For one-stage methods, we compared state-of-the-art detectors, including FoveaBox [51], FSAF [52], and FCOS [53]. With the ResNet-50 backbone, RetinaNet-C-MIL outperformed the state-of-the-art approach FCOS by 1.7%. With the ResNet-101 backbone, RetinaNet-C-MIL achieved 43.2% AP, which, respectively, outperformed RetinaNet by 4.1% and FCOS by 1.5%. As a one-stage detector, RetinaNet-C-MIL outperformed state-of-the-art two-stage detectors, such as IoU-Net [54] and cascade RCNN [55]. It is noteworthy that the performance gains were achieved without any additional computational cost, i.e., it simply modified the training procedure of detectors.

[2] RetinaNet-C-MIL* uses a smoothed function defined with (11), while RetinaNet-C-MIL uses an additional Gaussian weight in (11). The Gaussian weight of an anchor is calculated on IoU between the anchor and the top scored anchor, as $e^{-IoU(x_i,x_i^*)^2/\lambda^2}$.

TABLE VIII

PERFORMANCE (%) COMPARISON WITH THE BASELINE DETECTOR AND STATE-OF-THE-ART DETECTORS ON MS-COCO 2017

Detector	Backbone	AP	AP@0.5	AP@0.75	AP_S	AP_M	AP_L
RetinaNet(baseline) [44]	ResNet-50	35.7	55.0	38.5	18.9	38.9	46.3
FoveaBox [52]	ResNet-50	37.1	57.2	39.5	**21.6**	**41.4**	49.1
FSAF [53]	ResNet-50	37.2	57.2	39.4	21.0	41.2	49.7
FCOS [54]	ResNet-50	37.1	55.9	39.8	21.3	41.0	47.8
RetinaNet-C-MIL(Ours)	ResNet-50	**38.8**	**57.7**	**41.6**	20.4	41.2	**50.2**
RetinaNet(baseline) [44]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
IoU-Net [55]	ResNet-101	40.6	50.9	-	-	-	-
FoveaBox [52]	ResNet-101	40.6	60.1	43.5	23.3	45.2	54.5
FSAF [53]	ResNet-101	40.9	61.5	44.0	24.0	44.2	51.3
FCOS [54]	ResNet-101	41.5	60.7	45.0	24.4	44.8	51.6
Cascade RCNN [56]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
RetinaNet-C-MIL(Ours)	ResNet-101	**43.2**	**62.7**	**46.6**	**24.5**	**46.1**	**55.4**

VII. CONCLUSION

WSOD is an important yet challenging task. A major challenging aspect of WSOD lies in the nonconvexity of the objective function, which makes the learning procedure getting stuck in local minima. In this article, we proposed an elegant method, referred to as C-MIL, and alleviated the nonconvexity problem in a systematic way. C-MIL defined a series of smoothed functions to relax the original objective function based on parametric instance partition. C-MIL can gradually discover stable semantic extremal regions (SSERs) for accurate object localization. C-MIL was also applied to anchor/feature selection in supervised object detection. Significant performance gains over baseline methods and state-of-the-arts validated the effectiveness of C-MIL for general instance selection problems.

APPENDIX
THEORETICAL ANALYSIS OF C-MIL

For simplicity, denote $G(w) = \mathcal{L}(\mathcal{X}, w, 0)$ and $F(w) = \mathcal{L}(\mathcal{X}, w, 1)$. The smooth function can be defined using a convex homotopy as

$$\mathcal{L}(\mathcal{X}, w, \lambda) = (1-\lambda)G(w) + \lambda F(w) \tag{14}$$

which traces an implicitly defined curve $c(s) \in \mathcal{L}^{-1}(0)$ from a starting point $(w_0; 0)$ to a solution point $(\overline{w}; 1)$. G is required to be convex. In what follows, it requires to calculate the critical points of a smooth mapping $f : \mathbf{R}^N \to \mathbf{R}$, where N is the dimension of w. The numerical solution then consists of tracing a smooth curve

$$c(s) = (\lambda(s), w(s)) \in \mathcal{L}^{-1}(0) \tag{15}$$

with starting point $c(0)$ for some given critical point a of G, and starting tangent $\dot{c}(0) = (\dot{\lambda}(0), \dot{w}(0))$, with $\dot{w}(0) > 0$. The aim is to trace the curve c until the homotopy level $\lambda = 1$ is reached and a critical point of F is obtained. According to Sard's theorem [19], if all critical points of F are regular, then it is possible to make a choice of G such that zero is a regular value of $\mathcal{L}$. Accordingly, a theorem [56] is defined as follows.

Theorem 1: Let F and G be smooth functions and $\mathcal{L}$ be the convex homotopy defined in (14), which has zero as a regular value. Let $c(s)$ defined in (15) be a smooth curve obtained by defining the initial value problem as

$$\begin{aligned} \dot{c}(s) &= \sigma t(\mathcal{L}'(c(s))) \\ c(0) &= (0, a) \end{aligned} \tag{16}$$

where $\sigma \in \{+1, -1\}$ is a fixed orientation. Suppose that $\lambda(s)$ is increasing for $s \in [0, \overline{s}]$, $\lambda(\overline{s}) = 1$, and the critical point $b = w(\overline{s})$ of F is regular. The critical points a and b of G and F have the same Morse index.

One issue of Theorem 1 is that it is difficult to guarantee curve c having a monotone λ coordinate and reaching $\lambda = 1$ given a finite arc length. In [58] and [59], it has been observed that it is possible to extract a piecewise smooth curve, which monotonically increases in terms of λ. Considering that deep learning models typically have a large number of parameters, i.e., N is considerably large, it is difficult to directly achieve the numerical solution for Theorem 1. We, therefore, propose to empirically enumerate the possible curves (see Fig. 7) in the context of the monotony of λ and approximate the optimization for Theorem 1 in a small searching space. This makes it possible to implement the continuation methods to optimize nonconvex deep models.

ACKNOWLEDGMENT

The authors would like to thank their sincere appreciation to the Editors and the Reviewers for the constructive comments.

REFERENCES

[1] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2014, pp. 1997–2005.

[2] S. H. Oh, L. Y. Jae, J. Stefanie, and D. Trevor, "Weakly supervised discovery of visual pattern configurations," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1637–1645.

[3] Q. Ye *et al.*, "Self-learning scene-specific pedestrian detectors using a progressive latent model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2057–2066.

[4] W. Chong, R. Weiqiang, H. Kaiqi, and T. Tieniu, "Weakly supervised object localization with latent category learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 431–445.

[5] S. H. Oh, G. Ross, J. Stefanie, M. Julien, H. Zaid, and D. Trevor, "On learning to localize objects with minimal supervision," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1611–1619.

[6] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 343–350.

[7] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, “Min-entropy latent model for weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1297–1306.

[8] Y. Wei *et al.*, “TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 434–450.

[9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[10] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2015, pp. 1440–1448.

[11] H. Bilen and A. Vedaldi, “Weakly supervised deep detection networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2846–2854.

[12] P. Tang, X. Wang, X. Bai, and W. Liu, “Multiple instance detection network with online instance classifier refinement,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3059–3067.

[13] H. Bilen, M. Pedersoli, and T. Tuytelaars, “Weakly supervised object detection with convex clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1081–1089.

[14] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, “ContextLocNet: Context-aware deep network models for weakly supervised localization,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 350–365.

[15] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, “Weakly supervised cascaded convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5131–5139.

[16] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li, “Weakly supervised object detection via object-specific pixel gradient,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5960–5970, Dec. 2018.

[17] D. Zhang, J. Han, L. Zhao, and T. Zhao, “From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5549–5560, Dec. 2020.

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

[19] E. L. Allgower and K. Georg, *Numerical Continuation Methods—An Introduction Springer Series in Computational Mathematics*, vol. 13, Springer, 1990, pp. 1–388.

[20] S. Karen and Z. Andrew, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[21] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, “C-MIL: Continuation multiple instance learning for weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2199–2208.

[22] 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.

[23] 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.

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

[25] R. G. Cinbis, J. Verbeek, and C. Schmid, “Weakly supervised object localization with multi-fold multiple instance learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[26] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, “Min-entropy latent model for weakly supervised object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2395–2409, Oct. 2019.

[27] G. Yan *et al.*, “C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9834–9843.

[28] X. Li, M. Kan, S. Shan, and X. Chen, “Weakly supervised object detection with segmentation collaboration,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9735–9744.

[29] S. Kosugi, T. Yamasaki, and K. Aizawa, “Object-aware instance labeling for weakly supervised object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6064–6072.

[30] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, “WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8292–8300.

[31] P. Tang *et al.*, “PCL: Proposal cluster learning for weakly supervised object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[32] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, “Deep self-taught learning for weakly supervised object localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4294–4302.

[33] S. L. Richter and R. A. Decarlo, “Continuation methods: Theory and applications,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-13, no. 4, pp. 459–464, Aug. 1983.

[34] E. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction*. Berlin, Germany: Springer-Verlag, 1990.

[35] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proc. 26th Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 41–48.

[36] A. Beck and M. Teboulle, “Smoothing and first order methods: A unified framework,” *SIAM J. Optim.*, vol. 22, no. 2, pp. 557–580, Jan. 2012.

[37] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, “Improving deep neural networks using softplus units,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–4.

[38] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289

[39] C. Gulcehre, M. Moczulski, F. Visin, and Y. Bengio, “Mollifying networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–11.

[40] X. Chen, “Smoothing methods for nonsmooth, nonconvex minimization,” *Math. Program.*, vol. 134, no. 1, pp. 71–99, Aug. 2012.

[41] P. Chaudhari *et al.*, “Entropy-SGD: Biasing gradient descent into wide valleys,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–19.

[42] O. Maron and T. Lozano-Pérez, “A framework for multiple-instance learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 570–576.

[43] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[45] 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, Jun. 2010.

[46] 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.

[47] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

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

[49] P. Tang *et al.*, “Weakly supervised region proposal network and object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 352–368.

[50] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, Sep. 2004.

[51] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, “FoveaBox: Beyound anchor-based object detection,” *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[52] C. Zhu, Y. He, and M. Savvides, “Feature selective anchor-free module for single-shot object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.

[53] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

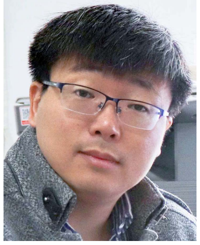
[54] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, “Acquisition of localization confidence for accurate object detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 816–832.

[55] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6154–6162.

[56] E. L. Allgower and K. Georg, *Introduction to Numerical Continuation Methods—Front Matter*. 2003, doi: 10.1137/1.9780898719154.

[57] E. L. Allgower, *Bifurcations Arising in the Calculation of Critical Points Via Homotopy Methods*. Basel, Switzerland: Birkhäuser, 1984, pp. 15–28.

[58] M. E. Henderson, “Complex bifurcation,” Ph.D. dissertation, California Inst. Technol., Pasadena, CA, USA, 1985, pp. 1–112.



Qixiang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

He has been a Professor with the University of Chinese Academy of Sciences, since 2009, and was a Visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, until 2013. He has authored or coauthored more than 100 articles in refereed conferences and journals including IEEE CVPR, ICCV, ECCV, and PAMI. His research interests include visual object detection and machine learning.

Dr. Ye received the Sony Outstanding Paper Award.

Fang Wan (Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2019.

He has been a Post-Doctoral Researcher with the School of Computer Sciences, University of Chinese Academy of Sciences, Beijing. He has authored or coauthored 15 articles in refereed conferences and journals including the IEEE CVPR, ICCV, and PAMI. His research interests include computer vision and machine learning, specifically for weakly supervised learning and visual object detection.

Chang Liu (Student Member, IEEE) received the B.S. degree from Jilin University, Jilin, China, in 2012. He is currently pursuing the Ph.D. degree with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and machine learning, specifically for neural architecture design and visual object detection. He has authored or coauthored ten papers in referred conferences including ECCV, the IEEE ICCV, and the IEEE CVPR.

Qingming Huang (Fellow, IEEE) received the bachelor’s degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of Chinese Academy of Sciences, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 400 academic articles in prestigious international journals. His research areas include multimedia computing, image processing, computer vision, and pattern recognition.

Dr. Huang has served as a General Chair, a Program Chair, a Track Chair, and a TPC member for ACM Multimedia, CVPR, ICCV, ICME, and ICMR.

Xiangyang Ji (Member, IEEE) received the B.S. degree in materials science and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008.

He joined Tsinghua University, Beijing, in 2008, where he is currently a Professor with the Department of Automation, School of Information Science and Technology. He has authored over 100 referred conference and journal articles. His current research interests include signal processing, image/video compressing, and intelligent imaging.