# Multiple Instance Differentiation Learning for Active Object Detection

Fang Wan, *Member, IEEE*, Qixiang Ye, *Senior Member, IEEE*, Tianning Yuan, *Student Member, IEEE*, Songcen Xu, Jianzhuang Liu, Xiangyang Ji, *Member, IEEE*, Qingming Huang, *Fellow, IEEE*

.

**Abstract**—Despite the substantial progress of active learning for image recognition, there lacks a systematic investigation of instance-level active learning for object detection. In this paper, we propose to unify instance uncertainty calculation with image uncertainty estimation for informative image selection, creating a multiple instance differentiation learning (MIDL) method for instance-level active learning. MIDL consists of a classifier prediction differentiation module and a multiple instance differentiation module. The former leverages two adversarial instance classifiers trained on the labeled and unlabeled sets to estimate instance uncertainty of the unlabeled set. The latter treats unlabeled images as instance bags and re-estimates image-instance uncertainty using the instance classification model in a multiple instance learning fashion. Through weighting the instance uncertainty using instance class probability and instance objectness probability under the total probability formula, MIDL unifies the image uncertainty with instance uncertainty in the Bayesian theory framework. Extensive experiments validate that MIDL sets a solid baseline for instance-level active learning. On commonly used object detection datasets, it outperforms other state-of-the-art methods by significant margins, particularly when the labeled sets are small. The code is available at https://github.com/WanFang13/MIDL.

**Index Terms**—Active Learning, Object Detection, Multiple Instance Learning, Instance Differentiation.

## 1 Introduction

With the rise of deep learning, unprecedented progress has been made in the computer vision area. Nevertheless, deep learning models are typically built upon fully supervised methods trained using large-scale datasets, which require intensive human effort for data annotation [1], [2]. Active learning, which tentatively selects a small proportion of informative data from the unlabeled dataset for training, is able to achieve comparable performance with fully supervised methods [3]. Despite the rapid progress of learning methods with less annotations, *e.g.*, weakly supervised learning [4], [5] and semi-supervised learning [6], [7], active learning remains the cornerstone for practical applications thanks to its simplicity and higher performance upper bound.

In the computer vision area, active learning methods are usually specified for image classification tasks. The goal is to select informative images from unlabeled image sets by estimating the information of each unlabeled image [8]–[16]. These methods [9], [16], [17], referred to as image-level active learning, can be categorized to uncertainty-based, representativeness-based, and combinations of them. Uncertainty-based methods [8], [15] involve various image selection metrics according to the uncertainty/informativeness estimated through classifying unlabeled images. Representativeness-based methods [9] try to find out images which can support the distribution of the unlabeled set.

Despite the substantial progress of image-level active learning, there still lacks an instance-level active learning method specified for object detection. Recently, naive approaches that simply aggregate instance-level uncertainty as image-level uncertainty [17]–[19] have been explored. These approaches, however, unfortunately ignore differentiating the informative instances from noisy instances, which hinders selecting informative images, particularly when there is a large number of noisy instances from backgrounds (Fig. 1(a)). The problem about how to accurately estimate image-level uncertainty by observing instance-level uncertainty remains unsolved.

In this paper, we propose an instance-level active learning method, termed multiple instance differentiation learning (MIDL), (Fig. 1(b)), with the aim to bridge the gap between image-level uncertainty and instance-level uncertainty within the Bayesian framework. In this framework, image-level uncertainty is conditionally related to instance uncertainty, instance class probability and instance objectness probability. The instance uncertainty is estimated by the classifier prediction differentiation module. The instance class probability and instance objectness probability are estimated by the multiple instance differentiation module. These two modules are plugged atop the convolutional neural network (CNN) and alternatingly trained in an end-to-end fashion (as shown in Fig. 2).

The classifier prediction differentiation module estimates

- *F. Wan and Q. Huang are with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 100049, China. T. Yuan and Q. Ye are with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing, 100049, China. Q. Ye and Q. Huang are also with the Peng Cheng Laboratory, Shenzhen, China. E-mail: {wanfang,qmhuang,qxye}@ucas.ac.cn, yuantianning19@mail.ucas.ac.cn. Q. Ye is the corresponding author.*
- *Songcen Xu and Jianzhuang Liu are with the Noah's Ark Lab, Huawei Technologies, Shenzhen, 518129, China, Email: {xusongcen,liu.jianzhuang}@huawei.com.*
- *X. Ji is with the Department of Automation, Tsinghua University, Beijing, 100084, China. E-mail: {xyji@tsinghua.edu.cn}.*

instance uncertainty by training two adversarial instance classifiers which differentiate informative instances while aligning the distributions of labeled and unlabeled instances. During training, minimizing the prediction differentiation of classifiers drives learning CNN features to align the distribution of unlabeled instances. Maximizing classifiers' prediction differentiation upon fixed features finds out the informative (hard) instances. Iterative maximizing and minimizing the classifiers' differentiation quantify the distribution overlap and bias of instances which indicates the uncertainty of each instance (as shown in Fig. 2).

The multiple instance differentiation module re-estimates the instance uncertainty by introducing a multiple instance learning (MIL) procedure. During training, each image is considered to be a bag of instances. The instance uncertainty is associated with the instance class probability to guarantee the semantic consistency between the instances and the image. To suppress the noisy instances and highlight representative ones, the instance uncertainty is further weighted by an instance objectness probability. Both of the instance class probability and instance objectness probability are learned by an MIL loss defined upon pseudo image class labels. Through associating the instance uncertainty with the instance class probability and the instance objectness probability, MIDL unifies image uncertainty with instance uncertainty, instance class probability and instance objectness probability in the total probability formula. It thereby can select the most informative images for detector training from the perspective of Bayesian theory.

MIDL evolves from our multiple instance active learning method [20] and is promoted by introducing the multiple instance differentiation module and formulating the Bayesian theory framework. MIDL is also extended from image object detection to video object detection where the informative instances are sparser and more difficult to be identified. The contributions of this work are summarized as follows:

- We propose an instance-level active learning method, multiple instance differentiation learning (MIDL), which bridges the gap between image-level uncertainty and instance-level uncertainty for informative image selection.
- We formulate the relationships among image-level uncertainty with instance uncertainty, instance class probability, and instance objectness probability within the Bayesian theory framework. We further reveal that the methods simply averaging instance-level uncertainty values are special cases of MIDL, supposing that the instance class probability and (or) instance objectness probability follow(s) the uniform distributions.
- We combine MIDL with the deep learning framework and achieve significant performance improvements for active object detection in both images and videos, setting the first solid baseline for instance-level active learning.

## 2 Related Work

Various taxonomies can be used to categorize the large amount of active learning methods, *e.g.*, uncertainty-based [21], [22] *vs.* distribution-based [23]–[25], hand-crafted metrics [26], [27] *vs.* learning loss methods [17],

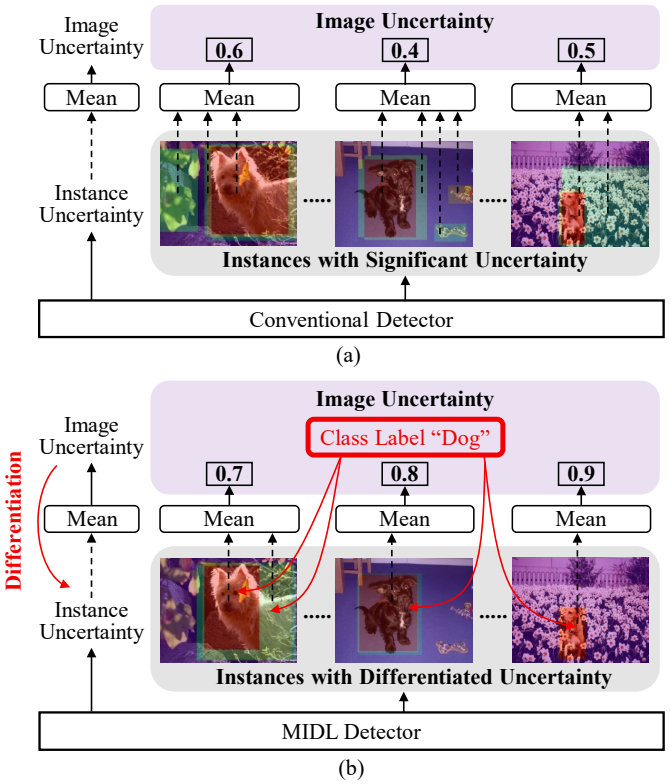



Fig. 1. Comparison of active object detection methods. (a) Conventional methods compute image uncertainty by simply averaging instance uncertainties, ignoring interference from a large number of background instances. (b) The proposed MIDL leverages uncertainty weighting via multiple instance learning to filter out interfering instances while bridging the gap between instance uncertainty and image uncertainty. (Best viewed in color)

and adversarial learning [16], [18], [28] *vs.* self-supervised learning [29]. In this paper, related works are roughly categorized to image-level active learning or instance-level active learning.

### 2.1 Image-Level Active Learning

Active learning, as one of the most important research topics in machine learning, has attracted intensive attention in the past few years. In the computer vision area, active learning was mainly explored for image recognition and the methods can be roughly categorized into uncertainty-based and distribution-based. The goal of the uncertainty-based methods is to define plausible metrics for finding out informative samples, while that of the distribution-based methods try to discover representative samples, which are supposed to have large diversity/representativeness in the feature space. While the classical approach has provided a systematic way [30] to combine uncertainty and diversity of unlabeled instances, we review these two kinds of methods in what follows.

**Uncertainty-based Methods.** Uncertainty is the most popular metric to select informative samples for active learning [3], which can be defined as the posterior probability of a predicted class [31], [32], or the gap between the posterior probabilities of the first and the second predicted classes [21], [22]. It can also be defined upon entropy [21],

[26], [27], which measures the variance of unlabeled samples. The expected model change methods [33], [34] utilized the present model to estimate the expected network gradient or expected prediction changes [35], [36]. The samples which have the potential to cause larger network gradient or expected prediction changes are defined as more informative.

There was a long history to find "plausible" uncertainty metrics for informative sample selection. However, without considering the distributions of labeled and unlabeled sets, it is impracticable to define a unified uncertainty metric for various datasets. Combined with deep learning, an improved uncertainty approach [8] used Monte Carlo Dropout and multiple forward passes to estimate uncertainty. The motivation is to use the large number of parameters in the deep network to "learn" a unified uncertainty model. However, the effectiveness for unseen data distributions or open sets [13] remains questionable. Furthermore the efficiency of sample selection is significantly reduced for the usage of dense dropout layers which hinder the network convergence, particularly when there are large numbers of instances to be considered.

**Distribution-based Methods.** This line of methods select diverse and informative samples by estimating the distribution of unlabeled samples. Clustering algorithms [37] were applied to build the unlabeled sample distribution. Discrete optimization algorithms [23]–[25] were employed to perform sample selection. By considering sample distances to their surrounding samples in the feature space, the context-aware methods [38], [39] selected samples that can represent the global sample distribution. Core-set [9] converted the problem of active learning to a core-set identification problem, $i.e.$, choosing a set of points to support the distribution of the unlabeled set while capturing the diversity of unlabeled samples.

In the deep learning era, active learning methods were combined with representation learning but remain falling into the uncertainty-based or distribution-based routines [10]–[12]. Sophisticated methods have extended active learning to open sets [13] and target domains [40], where the sample distributions are more difficult to be predicted. The self-paced active learning approach simultaneously considered the potential value and easiness of an instance, and try to train the model at low cost by querying the most informative samples in each round [14].

The learning loss approach [17] introduced a network structure to predict the "loss" of the unlabeled samples. It estimated sample uncertainty and selected samples with large "loss" in the way like hard negative mining. The sequential graph convolutional network [41] was proposed to select informative samples, which is able to be applied in the uncertainty-based fashion (UncertainGCN) or distribution-based fashion (CoreGCN). Task-aware variational adversarial active learning (TA-VAAL) [42] modified task-agnostic VAAL [16], which considered data distributions of both label and unlabeled sets by introducing a ranking conditional generative adversarial network to embed the ranking loss to VAAL. VaB-AL [43] trained a variational auto-encoder to handle the data imbalance problem. NCE-Net [44] proposed to reduce the risk of over-estimating unlabeled samples while improving the opportunity to query informative samples by replacing the softmax classifier of the deep neural network with a nearest neighbor classifier. ADS [28] unified the distribution alignment with sample selection by introducing adversarial classifiers to the deep learning framework. It operates in a way like domain adaptation, where the labeled set is regarded as the source domain and the unlabeled set the target domain.

Despite the substantial progress, existing methods were typically designed for image classification and experienced difficulty to bridge the gap between instance-level observation and informative image selection. Early methods [34], [45]–[47] introduced multiple instance learning to select informative images by discovering representative instances. This study aims to fill this gap by proposing the multiple instance differentiation learning, which selects informative instances by building the relationship between image-level uncertainty and instance-level uncertainty.

### 2.2 Instance-Level Active Learning

Instance-level active learning, $e.g.$, active object detection, is far more challenging than active image recognition as it requires to handle the large amount of negative instances and the complex instance distributions.

One solution is to directly extend the learning loss method [17] specified for image recognition to object detection, by simply sorting the loss predictions of instances to evaluate the image uncertainty. This line of approaches, however, could mislead the model towards hard negative mining but fail to reflect the true distribution of unlabeled instances. Another solution is simply aggregating the uncertainty of instances or pixels as the image-level uncertainty [19]. This line of approaches, however, could be seriously deteriorated by the large amount of negative instances/pixels from the backgrounds. The images of more complex backgrounds and higher averaged uncertainty tend to be falsely identified as informative ones. To alleviate the impact of backgrounds, the CDAL approach [48] introduced spatial context to active detection and selected diverse samples according to their distances to the labeled set. Nevertheless, how to define proper spatial context for images of various backgrounds is challenging.

In this study, we propose the MIDL method, with the aim to define a systematic and theoretical framework to estimate image-level uncertainty using instance-level models while considering the distribution of unlabeled instances. Based on the class-level differentiation and instance-level differentiation, MIDL has the advantages to progressively align the distributions of labeled and unlabeled instances while depressing the large number of negative instances.

## 3 Multiple Instance Differentiation Learning

We first overview the proposed MIDL method in Section 3.1. We then detail the MIDL modules within the unified Bayesian theory framework, where the classifier prediction differentiation module for instance uncertainty estimation, the multiple instance differentiation module for image uncertainty estimation, and the joint instance-image uncertainty learning for informative image selection are described in Sections 3.2–3.4 respectively. Finally, the proposed

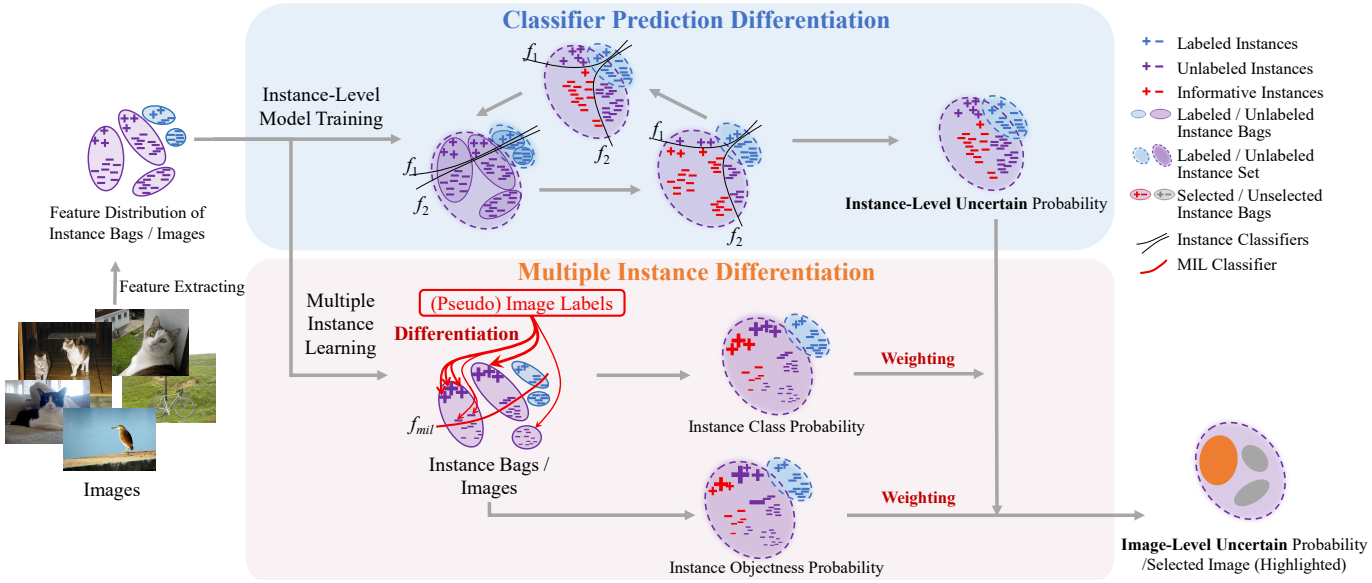



Fig. 2. Overview of the proposed multiple instance differentiation learning (MIDL). In the Bayesian framework, image-level uncertainty is conditionally related to instance uncertainty, instance class probability and instance objectness probability, which are respectively estimated by the classifier prediction differentiation module and the multiple instance differentiation module.

TABLE 1
Terms and symbols.

| Method | Formulation of $p(I\|x,\Theta)$ |
|---|---|
| $x_i$, $x$, $\mathcal{X}$ | instance, image and image set |
| $y_i$, $y$ , $\mathcal{Y}$ | instance label, image label and image label set |
| $f$, $g$, $\theta_f$, $\theta_g$, $\Theta$ | instance classifier/regressor, feature extractor, parameters of each module, parameters of MIDL |
| $p(I\|x,\Theta)$ | image uncertain probability |
| $p(I\|x_i,y_i,x,\Theta)$ | instance uncertain probability when given class label |
| $p(y_i\|x_i,x,\Theta)$ | instance class probability |
| $p(x_i\|x,\Theta)$ | instance objectness probability |

method is analyzed in Section 3.5. To better understanding the contents in the following sections, we summarize the main terms and symbols in Tab. 1.

### 3.1 Overview

Active object detection is defined as a learning task where a small set of images $\mathcal{X}_L^0$ (the labeled set) have instance labels $\mathcal{Y}_L^0$ and a large set of images $\mathcal{X}_U^0$ (the unlabeled set) have no instance label. Each instance label consists of a bounding box $y^{loc}$ and a class label $y^{cls}$. A detection model $M_0$ is initially trained upon the labeled set $\{\mathcal{X}_L^0, \mathcal{Y}_L^0\}$. Given the initial model $M_0$, active object detection iteratively selects a set of images $\mathcal{X}_S^0$ from $\mathcal{X}_U^0$ to label. The newly labeled images are merged with $\mathcal{X}_L^0$ to update the labeled set $\mathcal{X}_L^1$, *i.e.*, $\mathcal{X}_L^1 = \mathcal{X}_L^0 \cup \mathcal{X}_S^0$. The selected image set $\mathcal{X}_S^0$ is expected to be the most informative, *i.e.*, improving the detection performance as much as possible. Based on the updated labeled set $\mathcal{X}_L^1$, the detection model is retrained and updated to $M_1$. The model training and sample selection repeat until the size of labeled set reaches the annotation budget.

Considering the large number[1] of candidate instances in each image, there are three problems that need to be solved for active object detection: (1) how to evaluate the uncertainty of the unlabeled instances using detection models trained on the labeled set; (2) how to precisely estimate the image uncertainty given noisy and redundant instances; (3) how to jointly learn instance and image uncertainty for active image selection.

MIDL handles these three problems by introducing three modules. For the first problem, MIDL uses the classifier prediction differentiation module to highlight informative instances within the unlabeled images as well as aligning the distributions of the labeled and unlabeled instances (shown in the upper part of Fig. 2). This is a procedure to generalize the model trained on the labeled set to the unlabeled set in a way like transfer learning, which fills the distribution gap between the two sets. For the second problem, MIDL weights the instance uncertainty by introducing the instance objectness probability. For the third problem, MIDL introduces MIL to both the labeled and unlabeled sets to estimate the image uncertainty by weighting the instance uncertainty. This is done by treating each image as an instance bag while weighting the instance uncertainty under the supervision of the image classification loss. Optimizing the image classification loss facilitates highlighting truly

1. For example, the RetinaNet detector [49] produces ∼100k of anchors (instances) for an image.

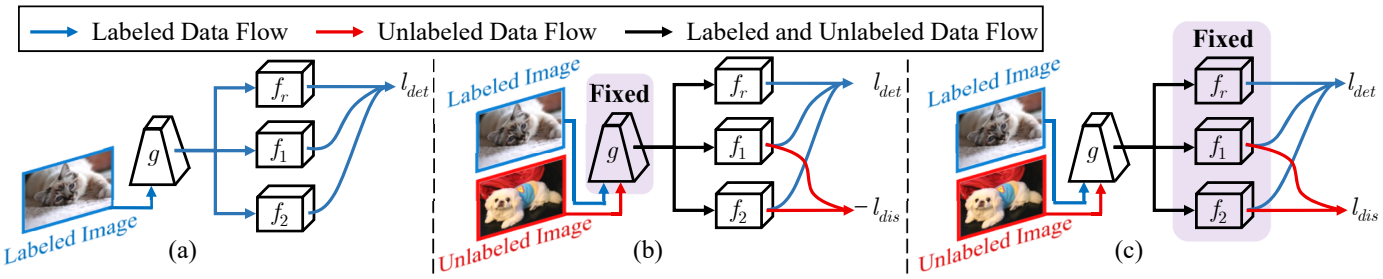



Fig. 3. Network architecture for instance uncertainty estimation by classifier prediction differentiation. (a) Training with the labeled set. (b) Maximizing instance uncertainty by maximizing classifier prediction discrepancy. (c) Minimizing instance uncertainty by minimizing classifier prediction discrepancy.

representative instances belonging to the same object classes while suppressing noisy ones (shown in the lower part of Fig. 2).

### 3.2 Classifier Prediction Differentiation for Instance Uncertainty Estimation

To identify informative instances, we introduce two adversarial classifiers to the detector head (shown in Fig. 3(b)). These two adversarial classifiers are trained on the labeled and unlabeled sets but have maximum prediction differentiation. As shown in Fig. 2(upper), the unlabeled instances which have larger prediction discrepancy by the adversarial classifiers have larger uncertainty. Such instances typically are far away from the labeled set and close to the classifier boundaries. The details are described bellow.

**Training on the Labeled Set.** The detection model, which consists of a feature extractor $g$ parameterized by $\theta_g$ and two instance classifiers $f_1$ and $f_2$ parameterized by $\theta_{f_1}$ and $\theta_{f_2}$, is trained on labeled instances. In the detection model, a bounding-box regressor $f_r$ parameterized by $\theta_{f_r}$ is trained to perform object localization. For object detection, each image $x$ from the labeled set $\mathcal{X}_L$ can be represented by multiple instances $\{x_i, i = 1, ..., N\}$, where each instance corresponds to a feature anchor on the convolutional feature map [49] and $N$ is the instance number in $x$. Let $\{y_i, i = 1, ..., N\}$ denote the set of the instance labels in image $x$. The detection model is trained by optimizing the following detection loss, as

$$\begin{aligned}\arg\min_{\Theta} \mathcal{L}_l &= \sum_{x\in\mathcal{X}_L} l_{det}(x,\Theta) \\ &= \sum_{x\in\mathcal{X}_L}\sum_{i}\Big(\mathrm{FL}(\hat{y}_i^{f_1}, y_i^{cls}) + \mathrm{FL}(\hat{y}_i^{f_2}, y_i^{cls}) \\ &\quad + \mathrm{SmoothL1}(\hat{y}_i^{f_r}, y_i^{loc})\Big),\end{aligned} \tag{1}$$

where $\Theta = \{\theta_{f_1}, \theta_{f_2}, \theta_{f_r}, \theta_g\}$. $\mathrm{FL}(\cdot)$ is the focal loss [49] for dense instance classification and SmoothL1 is the smooth L1 loss for bounding-box regression. $\hat{y}_i^{f_1} = f_1(g(x_i))$ and $\hat{y}_i^{f_2} = f_2(g(x_i))$ denote the classification results and $\hat{y}_i^{f_r} = f_r(g(x_i))$ the localization results. $y_i^{cls}$ and $y_i^{loc}$ respectively denote the ground-truth class label and bounding box label.

**Classifier Differentiation on the Unlabeled Set.** Given the detection model trained on the labeled set, we propose a classifier differentiation process to identify the informative instances by first maximizing the classifiers' prediction discrepancy and then minimizing this discrepancy, as shown in Fig. 2(upper). The unlabeled instances which are far way (biased) from the labeled set (distribution) are regarded as informative. Adding these biased instances to the labeled set facilities aligning the distributions of the labeled and unlabeled sets.

*(1) Maximizing Classifier Discrepancy.* Before the labeled set can precisely represent the unlabeled set, there exists a distribution bias between them, especially when the labeled set is small. The informative instances are in the biased distribution area. To find them out, $f_1$ and $f_2$ are designed as the adversarial instance classifiers with larger prediction discrepancy on the instances close to the classification boundary (shown in Fig. 2(upper)). The instance uncertainty is defined as the prediction discrepancy of $f_1$ and $f_2$.

To find out informative instances, it requires to fine-tune the network and maximize the prediction discrepancy of the adversarial classifiers (shown in Fig. 3(b)). In this procedure, $\theta_g$ is fixed so that the distributions of both the labeled and unlabeled instances are fixed. $\theta_{f_1}$ and $\theta_{f_2}$ are fine-tuned on the unlabeled set to maximize the prediction discrepancies for all instances. At the same time, it requires to preserve the detection performance on the labeled set. This is fulfilled by optimizing the following loss function, as

$$\arg\min_{\Theta\setminus\theta_g} \mathcal{L}_{max} = \sum_{x\in\mathcal{X}_L} l_{det}(x,\Theta) - \lambda \sum_{x\in\mathcal{X}_U} l_{dif}(x,\Theta), \tag{2}$$

where

$$l_{dif}(x,\Theta) = \frac{1}{N\times C}\sum_i\sum_c (\hat{y}_{i,c}^{f_1} - \hat{y}_{i,c}^{f_2})^2 \tag{3}$$

denotes the prediction discrepancy loss. $N$ is the number of instances in image $x$ and $C$ is the number of object classes in the dataset. $\hat{y}_i^{f_1}, \hat{y}_i^{f_2} \in \mathbb{R}^C$ are the instance classification predictions of the two classifiers for the $i$-th instance in image $x$, where $\hat{y}_{i,c}$ is the prediction score of class $c$ for instance $x_i$. $\lambda$ is a regularization hyper-parameter determined by experiment. As shown in Fig. 2(upper), the informative instances with different predictions by the adversarial classifiers tend to have larger prediction differentiation, which means larger uncertainty.

*(2) Minimizing Classifier Discrepancy.* After maximizing the prediction discrepancy, we further propose to minimize the prediction discrepancy to align the distributions of the labeled and unlabeled instances (shown in Fig. 3(c)). In this procedure, the classifier parameters $\theta_{f_1}$ and $\theta_{f_2}$ are fixed, while the parameters $\theta_g$ of the feature extractor are

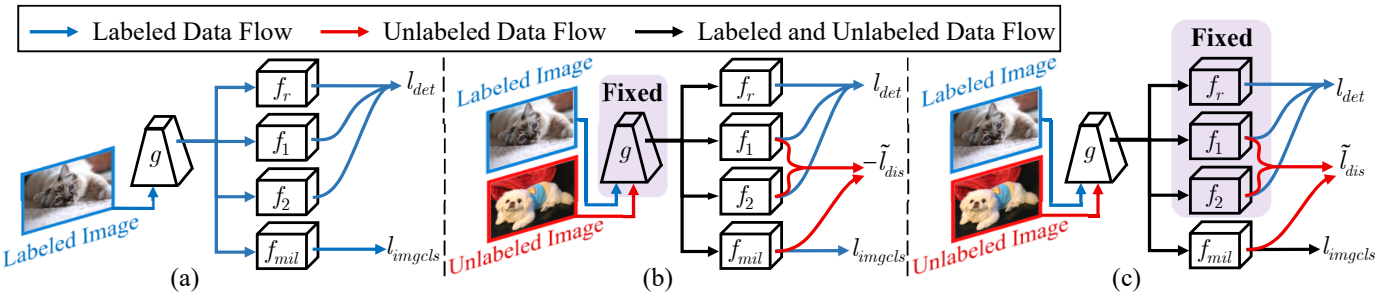



Fig. 4. Network architecture for instance uncertainty estimation by multiple instance differentiation. (a) Training with label set. (b) Instance differentiation and maximizing instance uncertainty. (c) Instance differentiation while minimizing instance uncertainty.

optimized by minimizing the prediction discrepancy loss, as

$$\arg\min_{\theta_g} \mathcal{L}_{min} = \sum_{x \in \mathcal{X}_L} l_{det}(x, \Theta) + \lambda \sum_{x \in \mathcal{X}_U} l_{dif}(x, \Theta). \quad (4)$$

By minimizing the prediction discrepancy, the distribution bias between the labeled set and the unlabeled set is minimized and their features are aligned, as much as possible.

**Iterative Training.** In each active learning cycle, the classifier differentiation procedures repeat so that the instance uncertainty is learned and the instance distributions of the labeled and unlabeled sets are progressively aligned. This actually defines an unsupervised learning procedure, which leverages the information (*i.e.*, prediction discrepancy) of the unlabeled set to improve the detection model.

After iterative training with the labeled set (Eq. 1) and classifier differentiation learning with both the labeled and unlabeled sets (Eq. 2 and Eq. 4 ), the instance-level model is learned and the informative instances are identified by the classifiers' prediction differentiation. The remaining problem of selecting informative images becomes how to compute the image-level uncertainty based upon the instance uncertainty.

### 3.3 Multiple Instance Differentiation for Image Uncertainty Estimation

By performing classifier prediction differentiation, the instance-level uncertainty is estimated by the differentiation of the two classifiers. However, the problem of how to precisely estimate the image uncertainty with noisy and clustered instances remains. MIDL aims to solve this problem in a systemically manner by modeling the image-level uncertainty and instance-level uncertainty in a unified Bayesian probability framework.

Let $p(I|x, \Theta)$ be the probability of image $x$ to be informative under the instance-level model parameterized by $\Theta$. $p(I|y_i, x_i, x, \Theta)$ denotes the uncertain probability of instance $x_i$ given class label $y_i$. As described in Section 3.2, an unlabeled instance with large prediction differentiation is an "outlier" of the labeled set and seen as uncertain one. Based on the instance prediction differentiation, we define the instance uncertain probability as

$$p(I|y_i, x_i, x, \Theta) = \left(\text{Sigmoid}(\hat{y}_{i,c}^{f_1}) - \text{Sigmoid}(\hat{y}_{i,c}^{f_2})\right)^2, \quad (5)$$

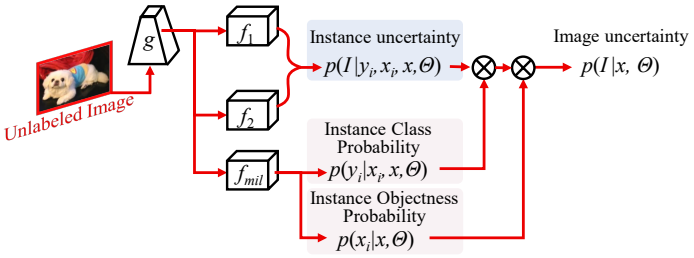



Fig. 5. Multiple instance differentiation architecture for image uncertainty estimation.

and re-formulate Eq. 3 as

$$l_{dif}(x, \Theta) = \sum_i \left( \sum_{y_i} p(I|y_i, x_i, x, \Theta) \frac{1}{C} \right) \frac{1}{N}. \quad (6)$$

In Eq. 6, the uncertain probability of an image is calculated by averaging the uncertain probabilities of instances in the image, *i.e.*, treating each instance with equal importance. Obviously, Eq. 6 ignores the differentiation among instances, where most crowded background instances are less important for informativeness estimation. Furthermore, for a specific object class, the instance uncertainty also varies from image to image. Therefore, when performing Eq. 6, the image uncertainty would be interfered by plenty of noisy instances, which causes the inconsistency between the image uncertainty and instance uncertainty. To differentiate the instances in image $x$ during training, we respectively replace the terms $1/C$ and $1/N$ in Eq. 6 with the instance class probability $p(y_i|x_i, x, \Theta)$ and the instance objectness probability $p(x_i|x, \Theta)$. Then, Eq. 6 is generalized to

$$\begin{aligned} l_{dif}(x, \Theta) &= \sum_i \left( \sum_{y_i} p(I|y_i, x_i, x, \Theta) p(y_i|x_i, x, \Theta) \right) p(x_i|x, \Theta) \\ &= \sum_i p(I|x_i, x, \Theta) p(x_i|x, \Theta) \\ &= p(I|x, \Theta). \end{aligned} \quad (7)$$

Eq. 7 shows that the relationship between the instance uncertainty and the image uncertainty is established upon the instance class probability $p(y_i|x_i, x, \Theta)$ and the instance objectness probability $p(x_i|x, \Theta)$. To estimate these probabilities, we respectively define the MIL procedure on both the labeled and unlabeled sets.

**Multiple Instance Learning.** MIL treats each image

as an instance bag and utilizes the instance classification predictions to estimate the bag labels. In turn, it weights the instance uncertainty scores by minimizing the image classification loss. This actually defines an Expectation-Maximization procedure [50], [51] to weight instance uncertainty across images/bags while filtering out noisy instances. Let $\boldsymbol{y}$ denote the image class label, where $\boldsymbol{y}_c \in [0, 1]$ denotes whether the image contains objects of class $c$. In the labeled set, $\boldsymbol{y}$ can be immediately obtained based on the ground-truth labels $y_i$ of objects in the image. Based on the relationship between image class label $\boldsymbol{y}$ and instance class $y_i$, the image class probability $p(\boldsymbol{y}|x, \Theta)$ can be predicted by the bag of instances as

$$\begin{aligned} p(\boldsymbol{y}|x, \Theta) &= \sum_i p(\boldsymbol{y}|x_i, x, \Theta) p(x_i|x, \Theta) \\ &= \sum_i p(y_i|x_i, x, \Theta) p(x_i|x, \Theta). \end{aligned} \tag{8}$$

The MIL loss is then defined as

$$\begin{aligned} l_{mil}(x, \Theta) = -\sum_{\boldsymbol{y}} \Big( & \boldsymbol{y} \log p(\boldsymbol{y}|x, \Theta) \\ & + (1 - \boldsymbol{y}) \log(1 - p(\boldsymbol{y}|x, \Theta)) \Big). \end{aligned} \tag{9}$$

By combining Eq. 8 and Eq. 9, the instance class probability $p(y_i|x_i, x, \Theta)$ and the instance objectness probability $p(x_i|x, \Theta)$ are learned when Eq. 9 is optimized.

**Instance Class and Objectness Probabilities.** Given the uncertainty $p(I|y_i, x_i, x, \Theta)$ of each instance $x_i \in x$ and Eq. 7, the first step of computing image $x$'s uncertainty is to calculate the instance class probability $p(y_i|x_i, x, \Theta)$ and the instance objectness probability $p(x_i|x, \Theta)$. To achieve this goal, we introduce an MIL branch $f_{mil}$ parameterized by $\theta_{f_{mil}}$, as shown in Fig. 4 and Fig. 5. The network parameters $\Theta$ is then updated to $\Theta = \{\theta_{f_1}, \theta_{f_2}, \theta_{f_r}, \theta_g, \theta_{f_{mil}}\}$. $f_{mil}$ contains two sub-branches which output $\hat{y}_{i,c}^{f_{mil}^c}$ and $\hat{y}_{i,c}^{f_{mil}^o}$ respectively. The instance class probability $p(y_i|x_i, x, \Theta)$ is predicted upon $\hat{y}_{i,c}^{f_{mil}^c}$ as

$$p(y_{i,c}|x_i, x, \Theta) = \begin{cases} \dfrac{\exp(\hat{y}_{i,c}^{f_{mil}^c})}{\sum_c \exp(\hat{y}_{i,c}^{f_{mil}^c})}, & \text{if} \max_c \hat{y}_{i,c}^{f_{mil}^c} > \delta_{fg} \\ \dfrac{1 - \exp(\hat{y}_{i,c}^{f_{mil}^c})}{\sum_c 1 - \exp(\hat{y}_{i,c}^{f_{mil}^c})}, & \text{otherwise}, \end{cases} \tag{10}$$

where $\delta_{fg}$ is an empirical threshold for foreground. In Eq. 10, when $\max_c \hat{y}_{i,c}^{f_{mil}^c} > \delta_{fg}$, the instance is highly confident to be foreground and therefore the instance class probability is directly assigned by the output of the MIL branch. However, when $\max_c \hat{y}_{i,c}^{f_{mil}^c} \leq \delta_{fg}$, the instance is likely to be background. As there are no background images (images does not contain any of the foreground objects) in the dataset, the MIL branch does not predict confidence for background class. Considering that the background instances takes up the largest proportion (>90%) of instances in the object detection task, we use a simple reverse operation to compute the class probability of background instance as the second line of Eq. 10. The instance objectness probability is predicted upon $\hat{y}_{i,c}^{f_{mil}^o}$ as

$$p(x_i|x, \Theta) = \frac{\exp(\max_c \hat{y}_{i,c}^{f_{mil}^o})}{\sum_i \exp(\max_c \hat{y}_{i,c}^{f_{mil}^o})}. \tag{11}$$

When Eq. 9 is optimized, the foreground instances have high instance class probability and instance objectness probability while the background ones have low probabilities. By combining Eqs. 10 and 11 with Eq. 7, the foreground instances with rich information are highlighted, while the redundant and noisy background instances with little information are suppressed. Consequently, the image uncertainty is mainly defined on instances which can most discriminate the image class and the image uncertainty and instance uncertainty are unified.

TABLE 2
Formulations of active learning methods. $C$ and $N$ are numbers of classes and anchors respectively. $\Theta$ denotes the network parameters.

| Method | Formulation of $p(I|x, \Theta)$ |
|---|---|
| CDAL [48] | $\sum_{x_i} \big(p(I|x_i, x, \Theta)\big) \frac{1}{N}$ |
| LL4AL [17] | $\sum_{x_i} \big( \sum_{y_i} p(I|y_i, x_i, x, \Theta) \times \frac{1}{C} \big) \times \frac{1}{N}$ |
| MI-AOD [20] | $\sum_{x_i} \big( \sum_{y_i} p(I|y_i, x_i, x, \Theta) p(y_i|x_i, x, \Theta) \big) \times \frac{1}{N}$ |
| MIDL | $\sum_{x_i} \big( \sum_{y_i} p(I|y_i, x_i, x, \Theta) p(y_i|x_i, x, \Theta) \big) p(x_i|x, \Theta)$ |

### 3.4 Joint Instance-Image Uncertainty Estimation for Active Image Selection

Combining Eqs. 5, 10 and 11, the image uncertainty $p(I|x, \Theta)$ in Eq. 7 is estimated by using the instance class probability and the instance objectness probability, where the informative instances are highlighted to ensure the consistency between the image uncertainty and the instance uncertainty. Finally, the learning loss of MIDL is defined as

$$\begin{aligned} \arg\min_{\Theta} \mathcal{L}_{MIDL} = & \sum_{x \in \mathcal{X}_L} l_{det}(x, \Theta) \\ & + \mathcal{L}_{max}(x, \Theta \setminus \theta_g) + \mathcal{L}_{min}(x, \theta_g) \\ & + \sum_{x \in \mathcal{X}_L} l_{mil}(x, \Theta)). \end{aligned} \tag{12}$$

As shown in Fig. 2, with a network feed-forward procedure, the adversarial instance classifiers output prediction discrepancy to estimate instance uncertainty. The instance class probability and instance objectness probability are predicted by the MIL branch. With a network back-propagation procedure, the gradient of each instance is weighted by these probabilities to highlight the informative instances. After multiple network feed-forward and backward procedures, the image uncertainty is estimated.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3277738



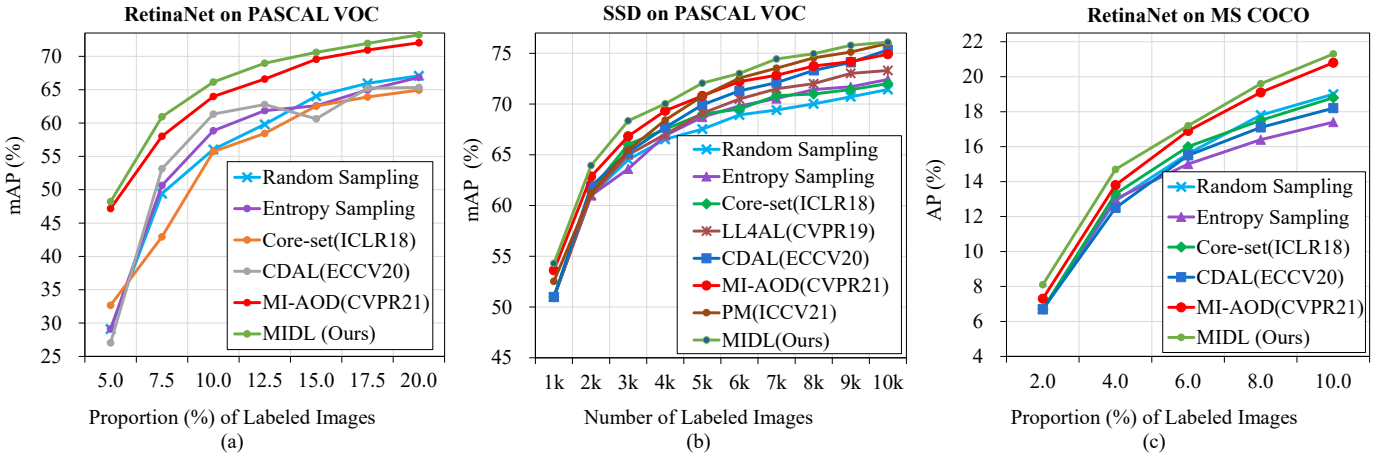



Fig. 6. Performance comparison of active object detection methods. (a) On PASCAL VOC using RetinaNet. (b) On PASCAL VOC using SSD. (c) On MS COCO using RetinaNet.

### 3.5 Discussion

In this section, we analyze the relations between existing instance-level learning methods and the proposed MIDL. It can be seen in Tab. 2 that existing methods that simply average instance-level uncertainty are special cases of MIDL. CDAL [48] uses contextual diversity to estimate uncertainty but ignores the differentiation of instances. LL4AL [17] uses the predicted loss as the instance uncertainty for each class. However, the differentiations of both the class and instance are ignored. MI-AOD [20] weights each instance with the class score, but still ignores the differentiation of crowded instances. Our proposed MIDL considers both the class differentiation and instance differentiation and therefore can precisely estimate image-level uncertainty, solving instance-level active learning in a systematic framework.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets**. For image object detection task, we use the PASCAL VOC and MS COCO datasets. The *trainval* sets of PASCAL VOC 2007 and 2012 datasets which contain 5011 and 11540 images are used for training. The VOC 2007 *test* set is used for evaluation under the metric of mean average precision (mAP). The MS COCO dataset contains 80 object categories with challenging aspects including dense objects and small objects with occlusion. We use the *train* set with 117k images for active learning and the *val* set with 5k images for evaluating AP. For video object detection task, we use the large-scale ImageNet VID dataset [52], which contains 30 object categories. The *train* set of ImageNet VID contains 3,862 videos and the *val* set contains 555 videos. For the large redundancy of video frames, it is important to learn discriminate detectors using few informative instances. Following the settings in [53], we train object detectors on the *train* set and evaluate them on the *val* set. For instance segmentation, we use the augmented PASCAL VOC 2012, which is a combination of the original PASCAL VOC 2012 and the whole SBD [54]. It contains 20 object categories with 10,582 training images and 1,449 *val* images.

**Active Learning Settings.** We use the RetinaNet [49] equipped with ResNet-50 and SSD [55] equipped with VGG-16 as the base detector. For RetinaNet, MIDL uses 5.0% of randomly selected images from the training set to initialize the labeled set on PASCAL VOC. In each active learning cycle, it selects 2.5% of the training images from the rest unlabeled set until the labeled images reach 20.0% of the training set. For the large-scale MS COCO, MIDL uses only 2.0% of randomly selected images from the training set to initialize the labeled set, and then selects 2.0% of the training images from the rest of the unlabeled set in each cycle until reaching 10.0% of the training set. In each cycle, the model is trained for 26 epochs with the mini-batch size 2 and the learning rate 0.001. After 20 epochs, the learning rate decreases to 0.0001. The momentum and the weight decay are set to 0.9 and 0.0001, respectively. For SSD, we follow the settings in LL4AL [17] and CDAL [48], where 1k images in the training set are selected to initialize the labeled set and 1k images are selected in each cycle. The learning rate is 0.001 for the first 240 epochs and reduced to 0.0001 for the last 60 epochs. The mini-batch size is set to 32 which is required by LL4AL.

We compare MIDL with random sampling, entropy sampling, Core-set [9], LL4AL [17], CDAL [48] and our previous work MI-AOD [20]. For entropy sampling, we use the averaged instance entropy as the image uncertainty. We repeat all experiments for 5 times and use the mean performance. MIDL and other methods share the same random seed and initialization for a fair comparison. $\lambda$ defined in Eqs. (2) and (4) is set to 10.

### 4.2 Performance

#### 4.2.1 Image Object Detection

**PASCAL VOC.** In Fig. 6, we run the proposed MIDL on a single TITAN RTX/A100 GPU, report its performance and compare it with state-of-the-art methods. Using either the RetinaNet [49] or SSD [56] detector, MIDL outperforms state-of-the-art methods with large margins. Particularly, it respectively outperforms state-of-the-art methods by 15.54% (Core-set), 7.81% (CDAL), and 4.81% (CDAL) when using 5.0%, 7.5%, and 10.0% samples. With 20.0% samples, MIDL

TABLE 3
Performance comparison on the ImageNet VID dataset.

| Methods | mAP (%) on Proport. (%) of Labeled Imgs. | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 15 | 20 | 25 | 30 | 100.0 |
| Random | 44.21 | 47.71 | 55.60 | 58.84 | 61.77 | 77.83 |
| Entropy | 44.21 | 49.62 | 56.60 | 58.57 | 61.99 | |
| MI-AOD [20] | 44.48 | 52.71 | 57.03 | 59.97 | 63.11 | |
| MIDL | **44.72** | **54.10** | **57.85** | **60.62** | **63.70** | |

TABLE 4
Instance segmentation performance comparison on the PASCAL VOC 2012 dataset.

| Methods | mAP$_5$0 (%) on Proport. (%) of Labeled Imgs. | | | | | |
|---|---|---|---|---|---|---|
| | 5 | 10 | 15 | 20 | 25 | 100.0 |
| Random | **15.23** | 41.51 | 48.42 | 52.29 | 53.81 | 63.60 |
| Entropy | **15.23** | **42.65** | 50.68 | 53.85 | 55.34 | |
| MI-AOD [20] | 15.12 | 42.36 | 51.89 | 54.50 | 56.33 | |
| MIDL | 15.15 | 42.31 | **52.25** | **55.21** | **57.21** | |

achieves 73.23% detection mAP, which significantly outperforms CDAL and MI-AOD by 7.9% and 1.2%. The improvements validate that MIDL can precisely learn instance uncertainty while selecting informative images. When using the SSD detector, MIDL outperforms state-of-the-art methods in almost all cycles, demonstrating the general applicability of MIDL to object detectors.

**MS COCO.** MS COCO is a challenging dataset with more categories, denser objects, and larger scale variation, where MIDL also outperforms the compared methods (Fig. 6). Particularly, it respectively outperforms Core-set and CDAL by 1.3%, 1.2%, and 2.2%, and 1.3%, 2.0%, and 2.8% when using 2.0%, 4.0%, and 10.0% labeled images.

#### 4.2.2 Video Object Detection

In Tab. 3, we report the performance of the proposed MIDL for video object detection on the ImageNet VID dataset [52]. Compared with the baseline method "Random Sampling", the "Entropy Sampling" approach achieves 1.91% (49.62% $vs.$ 47.71%) improvement with 15% labeled videos. When more videos (25%) are selected, "Entropy" becomes slightly worse than "Random". In the last cycle (30% labeled videos), "Entropy Sampling" outperforms "Random Sampling" by 0.22%% (61.99% $vs.$ 61.77%). These results show that "Entropy Sampling" has the chance to discover informative videos but may suffer from less informative videos without differentiating instances and classes. Unlike "Entropy Sampling", the MI-AOD approach consistently outperforms both "Random Sampling" and "Entropy Sampling" in all active learning cycles. MIDL further improves MI-AOD by 0.59% (63.70% $vs.$ 63.11%) and significantly outperforms "Random Sampling" and "Entropy Sampling" by 1.93% and 1.71% respectively at the last training cycle, which indicates the effectiveness of class prediction differentiation and multiple instance differentiation for informative object selection from video clips.

#### 4.2.3 Instance Segmentation

In Tab. 4, we report the performance of the proposed MIDL for instance segmentation on the PASCAL VOC 2012 dataset. The "Entropy Sampling" approach outperforms the baseline method "Random Sampling" by 1.53% (55.34% vs 53.81%) with 25% labeled images. MI-AOD consistently outperforms both "Random Sampling" and "Entropy Sampling" in all active learning cycles. MIDL further improves MI-AOD by 0.88% (57.21% $vs.$ 56.33%) and significantly outperforms "Random Sampling" and "Entropy Sampling" by 1.93% and 1.71% respectively at the last training cycle, which indicates the effectiveness of class prediction differentiation and multiple instance differentiation for informative instance segmentation.

### 4.3 Ablation Analysis

**Classifier Prediction Differentiation.** As shown in Tab. 5, with the classifier prediction differentiation module, the detection performance is improved up to 70.06% in the last cycle, which outperforms the Random method by 2.97% (70.06% $vs.$ 67.09%), demonstrating the effectiveness of the class prediction differentiation module for instance uncertainty estimation.

**Multiple Instance Differentiation.** In Tab. 5, the classifier prediction differentiation module achieves comparable performance with the method using the random image selection strategy in the early cycles. This is because there are significant noisy instances that make the instance uncertainty inconsistent with image uncertainty. After using the multiple instance differentiation module to differentiate the instance uncertainty, the performance at early cycles is improved by 5.04%∼17.09% in the first three cycles (row 5 $vs.$ row 2 in Tab. 5). In the last cycle, the performance is improved by 1.28% (68.48% $vs.$ 67.20%) in comparison with the classifier prediction differentiation module and 1.39% in comparison with the Random method (68.48% $vs.$ 67.09%). Interestingly, when using 100.0% images for training, the detector with the multiple instance differentiation module outperforms the detector without the multiple instance differentiation module by 1.09% (78.37% $vs.$ 77.28%). When further applying the instance objectness probability, the performance are improved in all cycles (row 9 $vs.$ row 8). These results clearly show that the multiple instance differentiation module can suppress the interfering instances while highlighting more representative ones, which can indicate informative images for detector training. Compared with hand-crafted sample selection strategies (Rand., Max Unc. and Mean Unc. in Tab. 5), selecting images by the image uncertainties learned by the joint instance-image uncertainty learning module (Section 3.4) further improves the detection performance.

With the multiple instance differentiation module and joint instance-image uncertainty learning module, MIDL outperforms the SOTA method MI-AOD [20] that only uses the instance class probability in all cycles (row 9 $vs.$ row

TABLE 5
Module ablation on PASCAL VOC. The first line shows the result of the baseline method with random image selection. "Max Unc." denotes that the image uncertainty is represented by the maximum instance uncertainty. "Mean Unc." denotes the averaged instance uncertainty. "CP" and "COP" respectively denote instance class probability and both of instance class and objectness probability. "CPD", "MID" and "JUL" respectively denote classifier prediction differentiation, multiple instance differentiation and joint instance-image uncertainty learning.

| | Training | | Sample Selection | | | | mAP (%) on Proportion (%) of Labeled Images | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CPD | MID | Rand. | Max Unc. | Mean Unc. | JUL | 5.0 | 7.5 | 10.0 | 12.5 | 15.0 | 17.5 | 20.0 | 100.0 |
| 1 | | | ✓ | | | | 28.31 | 49.42 | 56.03 | 59.81 | 64.02 | 65.95 | 67.09 | 77.28 |
| 2 | ✓ | | ✓ | | | | 30.09 | 49.17 | 55.64 | 60.93 | 64.10 | 65.77 | 67.20 | |
| 3 | ✓ | | | ✓ | | | 30.09 | 49.79 | 58.94 | 63.11 | 65.61 | 67.84 | 69.01 | |
| 4 | ✓ | | | | ✓ | | 30.09 | 49.74 | 60.60 | 64.29 | 67.13 | 68.76 | 70.06 | |
| 5 | | CP | ✓ | | | | 47.18 | 57.12 | 60.68 | 63.72 | 66.10 | 67.59 | 68.48 | **78.37** |
| 6 | | CP | | ✓ | | | 47.18 | 57.58 | 61.74 | 64.58 | 66.98 | 68.79 | 70.33 | |
| 7 | | CP | | | ✓ | | 47.18 | 58.03 | 63.98 | 66.58 | 69.57 | 70.96 | 72.03 | |
| 8 | ✓ | | | | | ✓ | 32.58 | 52.90 | 60.97 | 61.39 | 66.19 | 67.47 | 69.39 | |
| 9 | | COP | | | | ✓ | **48.21** | **60.94** | **66.14** | **68.98** | **70.59** | **71.94** | **73.23** | |

TABLE 6
Performance under different hyper-parameters $\lambda$.

| $\lambda$ | mAP (%) on Proportion (%) of Labeled Imgs. | | | | | | |
|---|---|---|---|---|---|---|---|
| | 5.0 | 7.5 | 10.0 | 12.5 | 15.0 | 17.5 | 20.0 |
| 1 | 48.53 | 61.04 | 65.33 | **68.72** | 69.69 | 71.40 | 72.49 |
| 5 | 48.53 | 60.86 | 65.16 | 68.45 | 69.41 | 71.42 | 72.52 |
| 10 | 48.53 | **61.67** | 66.00 | 68.31 | 69.76 | 71.20 | **72.70** |
| 20 | 48.53 | 61.61 | **66.58** | 68.65 | **70.29** | **71.72** | 72.49 |
| 50 | 48.53 | 60.67 | 64.36 | 67.45 | 69.49 | 71.09 | 72.12 |

TABLE 7
Comparison of time costs on PASCAL VOC.

| Method | Time (h) on Proportion (%) of Labeled Imgs. | | | | | | |
|---|---|---|---|---|---|---|---|
| | 5.0 | 7.5 | 10.0 | 12.5 | 15.0 | 17.5 | 20.0 |
| Random | 0.77 | 1.12 | 1.45 | 1.78 | 2.12 | 2.45 | 2.78 |
| CDAL [48] | 1.18 | 1.50 | 1.87 | 2.19 | 2.68 | **2.83** | **2.82** |
| MI-AOD [20] | **1.03** | **1.42** | **1.78** | **2.18** | **2.55** | 2.93 | 3.12 |
| MIDL | 1.06 | 1.46 | 1.83 | 2.24 | 2.62 | 3.01 | 3.21 |

7). Especially, MIDL significantly outperforms MI-AOD by 3.64%, 2.02% and 1.73% respectively in the 2nd, 3rd, 4th cycles. In the last cycle, the detection mAP reaches 72.7%, which outperforms MI-AOD by 0.63% (72.70% $vs$. 72.03%).

**Hyper-parameters and Time Cost.** The effect of the regularization factor $\lambda$ defined in Eqs. (2) and (4) is shown in Tab. 6. MIDL has the best performance when $\lambda$ is set to 10. Tab. 7 shows that MIDL costs less time at early cycles than CDAL.

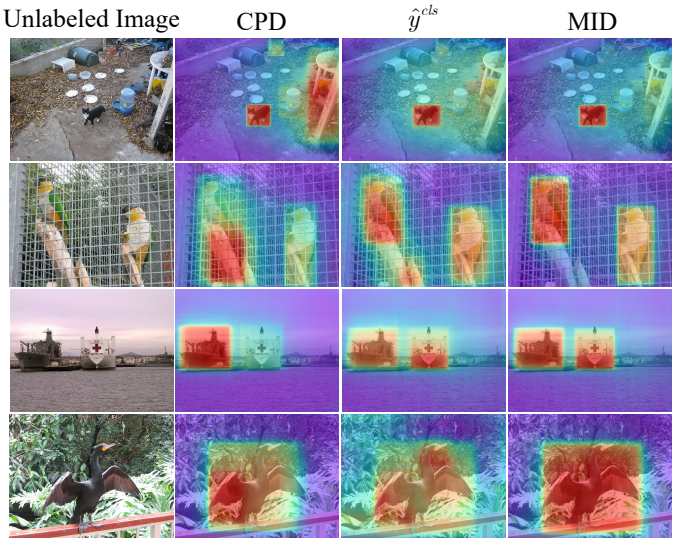



Fig. 7. Visualization of learned and differentiated instance uncertainty and image classification score. "CPD" and "MID" respectively denote classifier prediction differentiation and multiple instance differentiation. (Best viewed in color)

### 4.4 Model Analysis

**Visualization Analysis.** In Fig. 7, we visualize the learned and re-weighted uncertainty and image classification scores of instances. The heatmaps are calculated by summarizing the uncertainty scores of all instances. With only the classifier prediction differentiation module, there exist interference instances from the background (row 1) or around the true positive instance (row 2), and the results tend to miss the true positive instances (row 3) or instance parts (row 4). MIL can assign high image classification scores to the instances of interest while suppressing backgrounds. As a result, MIDL leverages the image classification scores to weight instances towards accurate instance uncertainty prediction. In Fig. 8, we visualize the instance uncertain

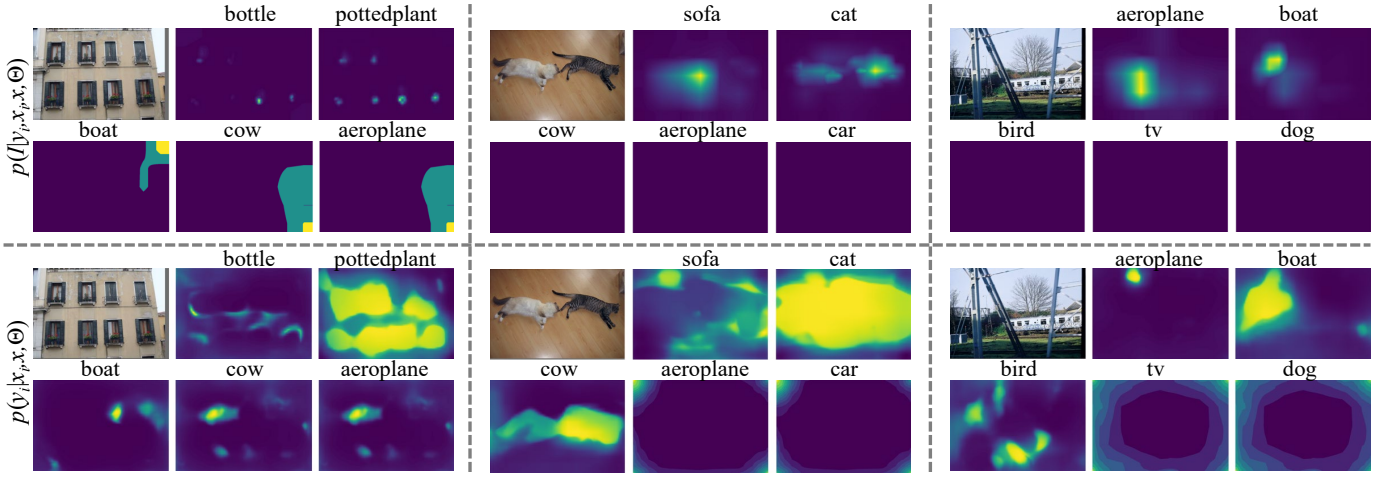



Fig. 8. Visualization of instance uncertain and instance class probability with respect to object categories.

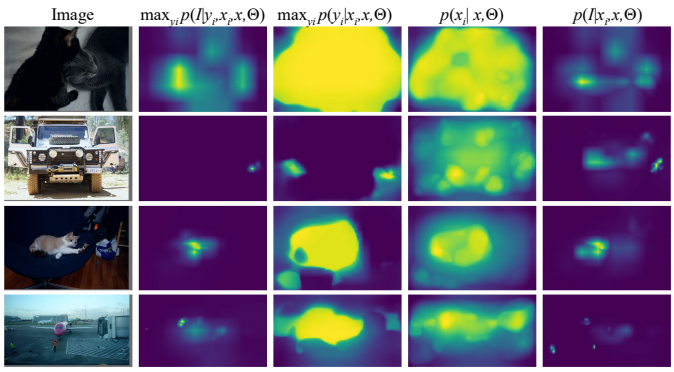



Fig. 9. Visualization of probabilities in MIDL.

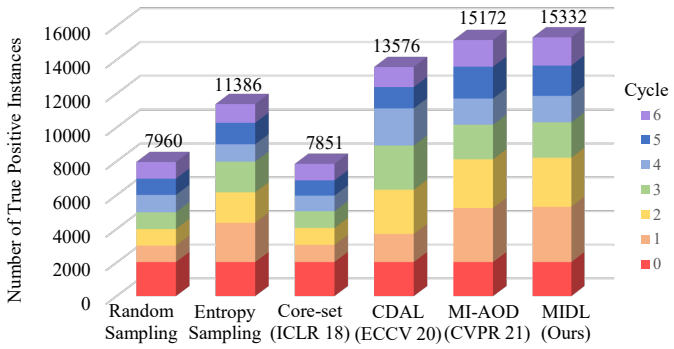



Fig. 10. The number of true positive instances selected in each active learning cycle on PASCAL VOC using RetinaNet.

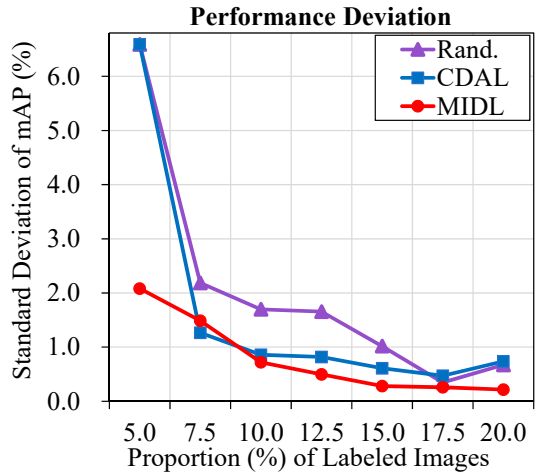



Fig. 11. Performance deviation comparison on the PASCAL VOC datasets. "Rand." denotes random sampling.

probability $p(I|y_i, x_i, x, \Theta)$ (Fig. 8 upper) and instance class probability $p(y_i|x_i, x, \Theta)$ (Fig. 8 lower) with respective to the object categories. The categories in row 1 and row 3 are with highest uncertainty scores ($\text{argmax}_{y_i} \sum_{x_i} p(I|y_i, x_i, x, \Theta)$), while categories in row 2 and row 4 are with lowest uncertainty scores. In Fig. 8 left and middle, it can be seen that $p(I|y_i, x_i, x, \Theta)$ lacks semantic discrimination and effected by the similar classes (Line 1). $p(y_i|x_i, x, \Theta)$ is discriminative to suppress the similar classes but can not indicate the informative areas in images (Line 3). Combining $p(I|y_i, x_i, x, \Theta)$ with $p(y_i|x_i, x, \Theta)$ results in discovering informative area with less noise (as shown in Fig. 9 last column). In Fig. 8 right, the classification is failed and both $p(I|y_i, x_i, x, \Theta)$ and $p(y_i|x_i, x, \Theta)$ focus on background noise. In Fig. 9, we visualize the instance uncertain probability $p(I|y_i, x_i, x, \Theta)$, the instance class probability $p(y_i|x_i, x, \Theta)$, and instance objectness probability $p(x_i|x, \Theta)$. From the last column of Fig. 9, one can be seen that MIDL is able to differentiate the instance uncertain probability $p(I|y_i, x_i, x, \Theta)$ and discovers as much as informative instances of foreground objects with least background noise.

**Statistical Analysis.** In Fig. 10, we calculate the number of true positive instances selected in each active learning cycle. It can be seen that MIDL significantly hits more true positives in all learning cycles. This shows that the proposed MIDL approach can activate true positive objects better while filtering out interfering instances, which facilities selecting informative images for detector training.

### 4.5 Robustness Analysis

**Performance Deviation.** In Fig. 11, we compare the performance deviation of MIDL with those of CDAL and random

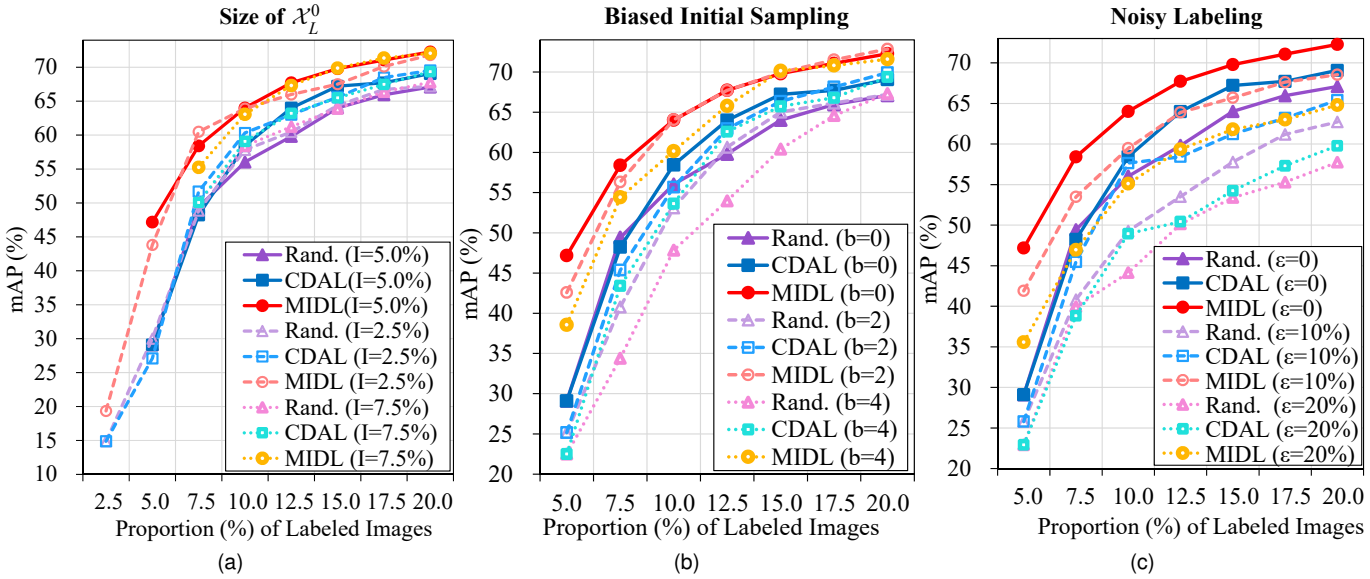



Fig. 12. Performance comparison with respect to the size of $\mathcal{X}_L^0$, biased initial sampling and noisy labeling on PASCAL VOC datasets. "Rand." denotes random sampling.

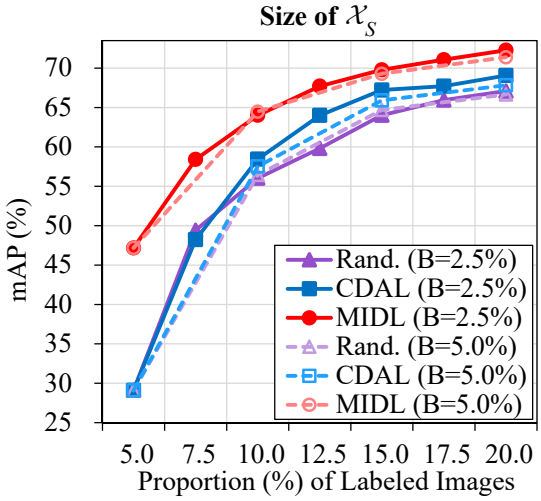



Fig. 13. Performance comparison under different sizes of $\mathcal{X}_S$ on the PASCAL VOC datasets. "Rand." denotes random sampling.

sampling. The performance deviations of CDAL and random sampling are significantly larger than that of MIDL when the labeled set is small. With more labeled images, the performance deviations of all methods decrease, while that of MIDL keeps smaller than other methods at almost all cycles. These results further validate that by introducing the MIL classifier, MIDL can suppress noisy instances and therefore achieves more robust performance than CDAL and random sampling.

**Size of $\mathcal{X}_L^0$.** We conduct experiments under different sizes of $\mathcal{X}_L^0$ to analyze the "cold start" issue [57] (Fig. 12(a)). $I = |\mathcal{X}_L^0|$ denotes the size of $\mathcal{X}_L^0$. In all experiments, the size of $\mathcal{X}_S$ is set to 2.5% of the training images. The starting sizes $I$ for each method are set to 5.0% (solid lines), 2.5% (dashed lines) and 7.5% (dotted lines) respectively. It can be seen that MIDL outperforms CDAL and random sampling with all starting sizes $I$, demonstrating its robustness to the cold start issue. When the size of $\mathcal{X}_L^0$ is set to 2.5% of the training images, the performances of all methods significantly drop, as the initialed labeled set is too small. The performance of MIDL increases largely in the second cycle, validating that MIDL can select more informative samples when the initial labeled sets are small.

**Biased Initial Sampling.** We make an analysis on how the sampling bias of the initially labeled sets affects the detection performance. With biased sampling, the sample distributions of the labeled and the unlabeled sets are not consistent, which imposes challenges to active learners. We model a possible form of bias in the initial labeled sets by not providing images and labels for $b$ chosen classes at random and we compare it to the cases where the initial labeled images are randomly selected from all classes ($i.e.$, $b = 0$). Fig. 12(b) shows the performances for $b = 0$ (solid lines), $b = 2$ (dashed lines) and $b = 4$ (dotted lines). With the biased initial sampling, MIDL outperforms the compared methods at all active learning cycles.

**Noisy Labeling.** We randomly change the image labels to its similar classes, which is thought to be the major annotation noise caused by low quality images and/or non-professional annotators. To simulate images with noisy labels, divide the PASCAL VOC dataset to 4 super-classes ($i.e.$, person, animal, vehicle and indoor super-classes [58]) and 20 sub-classes. Let $\epsilon$ denote the percentage of selected objects in the training set. We set $\epsilon = 0\%$ (solid lines), $\epsilon = 10\%$ (dashed lines), and $\epsilon = 20\%$ (dotted lines). For each selected object, we change its class label to a random wrong class label from the super-class.

Fig. 12(c) shows the effects of noisy labeling of MIDL and the compared methods. As the percentage of noisy labels increases, CDAL's mAP tends to close to that of random sampling, while MIDL's mAP maintains superior to those of CDAL and random sampling when $\epsilon = 10\%$ or $\epsilon = 20\%$. The performance of "MIDL ($\epsilon = 20\%$)" is significantly better

than those of "CDAL ($\epsilon = 20\%$)" and even comparable to that of "CDAL ($\epsilon = 10\%$)". This validates that MIDL improves the robustness to noisy labeling.

**Size of $\mathcal{X}_S$.** In each active learning cycle, a set of images $\mathcal{X}_S$ (defined in Section 3.1) are selected and labeled. We make analysis on different sizes of $\mathcal{X}_S$ which are set to 2.5% (solid lines) and 5.0% (dashed lines) of the training images, respectively in Fig. 13. $B = |\mathcal{X}_S|$ denotes the size of $\mathcal{X}_S$. The experiments are conducted with the same initial labeled set and the same annotation budget. It can be seen that with larger sizes of $\mathcal{X}_S$, MIDL remains outperforming the compared methods in all cycles, validating that MIDL's performance is more robust to the size of $\mathcal{X}_S$.

## 5 Conclusion

In this paper, we formulate the instance-level active learning in the Bayesian framework and propose Multiple Instance Differentiation Learning (MIDL) to select informative images. In the Bayesian framework, we estimate the image uncertain probability by performing the total probability formula to differentiate and aggregate the instance-level uncertain probability with the instance class probability and instance objectness probability. MIDL consists of a class prediction differentiation module and a multiple instance differentiation module. During training, the class prediction differentiation module trains the instance-level model on the labeled and unlabeled images to estimate the instance-level uncertain probability. The multipe instance differentiation module learns the instance class probability and instance objectness probability through a multiple instance learning module. We reveal that existing instance-level active learning methods are special cases of MIDL, where the instance class probability or/and instance objectness probability is/are set to be the uniform distribution. Experiments on commonly used object detection and video object detection datasets show that MIDL outperforms state-of-the-art methods with significant margins, particularly when the labeled sets are small. MIDL has set a solid baseline for instance-level active learning.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 62006216, 61836012, and 62171431, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000.

## References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1106–1114.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[3] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012.

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

[5] ——, "Min-entropy latent model for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2395 – 2409, 2019.

[6] G. Papandreou, L. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *ICCV*, 2015, pp. 1742–1750.

[7] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *NeurIPS*, 2015, pp. 3546–3554.

[8] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *ICML*, vol. 70, 2017, pp. 1183–1192.

[9] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *ICLR*, 2018.

[10] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE TPAMI*, vol. 40, no. 1, pp. 7–19, 2018.

[11] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE TCSVT*, vol. 27, no. 12, pp. 2591–2600, 2017.

[12] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *CVPR*, 2018, pp. 9368–9377.

[13] Z. Liu and S. Huang, "Active sampling for open-set classification without initial annotation," in *AAAI*, 2019, pp. 4416–4423.

[14] Y. Tang and S. Huang, "Self-paced active learning: Query the right thing at the right time," in *AAAI*, 2019, pp. 5117–5124.

[15] Z. Liu, S. Li, S. Chen, Y. Hu, and S. Huang, "Uncertainty aware graph gaussian process for semi-supervised learning," in *AAAI*, 2020, pp. 4957–4964.

[16] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *ICCV*, 2019, pp. 5971–5980.

[17] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *CVPR*, 2019, pp. 93–102.

[18] B. Zhang, L. Li, S. Yang, S. Wang, Z. Zha, and Q. Huang, "State-relabeling adversarial active learning," in *CVPR*, 2020, pp. 8753–8762.

[19] H. H. Aghdam, A. Gonzalez-Garcia, A. M. López, and J. van de Weijer, "Active learning for deep detection neural networks," in *ICCV*, 2019, pp. 3671–3679.

[20] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, "Multiple instance active learning for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[21] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *CVPR*, 2009, pp. 2372–2379.

[22] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *ECML*, vol. 4212, 2006, pp. 413–424.

[23] Y. Guo, "Active instance sampling via matrix partition," in *NeurIPS*, 2010, pp. 802–810.

[24] E. Elhamifar, G. Sapiro, A. Y. Yang, and S. S. Sastry, "A convex optimization framework for active learning," in *ICCV*, 2013, pp. 209–216.

[25] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *IJCV*, vol. 113, no. 2, pp. 113–127, 2015.

[26] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *EMNLP*, 2008, pp. 1070–1079.

[27] W. Luo, A. G. Schwing, and R. Urtasun, "Latent structured active learning," in *NeurIPS*, 2013, pp. 728–736.

[28] T. Yuan, M. Fu, F. Wan, S. XU, and Q. Ye, "Agreement-discrepancy-selection: Active learning with progressive distribution alignment," *arXiv*, 2021.

[29] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," in *CVPR*, 2018, pp. 1605–1613.

[30] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, 2014.

[31] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR*, 1994, pp. 3–12.

[32] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine Learning*, 1994, pp. 148–156.

[33] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *ICML*, 2001, pp. 441–448.

[34] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *NeurIPS*, 2007, pp. 1289–1296.

[35] A. Freytag, E. Rodner, and J. Denzler, "Selecting influential examples: Active learning with expected model output changes," in *ECCV*, vol. 8692, 2014, pp. 562–577.

[36] C. Käding, E. Rodner, A. Freytag, and J. Denzler, "Active and continuous exploration with deep neural networks and expected model output changes," *CoRR*, vol. abs/1612.06129, 2016.

[37] H. T. Nguyen and A. W. M. Smeulders, "Active learning using pre-clustering," in *ICML*, 2004.

[38] M. Hasan and A. K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *ICCV*, 2015, pp. 4543–4551.

[39] O. M. Aodha, N. D. F. Campbell, J. Kautz, and G. J. Brostow, "Hierarchical subquery evaluation for active learning on a graph," in *CVPR*, 2014, pp. 564–571.

[40] B. Fu, Z. Cao, J. Wang, and M. Long, "Transferable query selection for active domain adaptation," *arXiv*, 2021.

[41] R. Caramalau, B. Bhattarai, and T. Kim, "Sequential graph convolutional network for active learning," *arXiv preprint arXiv:2006.10219*, 2020.

[42] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning," *arXiv preprint arXiv:2002.04709*, 2020.

[43] J. Choi, K. M. Yi, J. Kim, J. Choo, B. Kim, J. Chang, Y. Gwon, and H. J. Chang, "Vab-al: Incorporating class imbalance and difficulty with variational bayes for active learning," *arXiv preprint arXiv:2003.11249*, 2020.

[44] F. Wan, T. Yuan, M. Fu, X. Ji, Q. Huang, and Q. Ye, "Nearest neighbor classifier embedded network for active learning," *arXiv*, 2021.

[45] S. Huang, N. Gao, and S. Chen, "Multi-instance multi-label active learning," in *IJCAI*, 2017, pp. 1886–1892.

[46] R. Wang, X. Wang, S. Kwong, and C. Xu, "Incorporating diversity and informativeness in multiple-instance active learning," *IEEE TFS*, vol. 25, no. 6, pp. 1460–1475, 2017.

[47] M. Carbonneau, E. Granger, and G. Gagnon, "Bag-level aggregation for multiple-instance active learning in instance classification problems," *IEEE TNNLS*, vol. 30, no. 5, pp. 1441–1451, 2019.

[48] S. Agarwal, H. Arora, S. Anand, and C. Arora, "Contextual diversity for active learning," *CoRR*, vol. abs/2008.05723, 2020.

[49] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE TPAMI*, vol. 42, no. 2, pp. 318–327, 2020.

[50] A. Stuart, T. Ioannis, and H. Thomas, "Support vector machines for multiple-instance learning," in *NeurIPS*, 2002, pp. 561–568.

[51] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016, pp. 2846–2854.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 337–10 346.

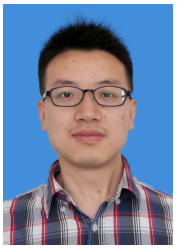
[54] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 International Conference on Computer Vision*, 2011, pp. 991–998.

[55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

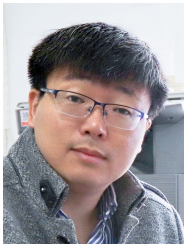
[56] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

[57] M. Gao, Z. Zhang, G. Yu, S. Ö. Arik, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," *CoRR*, vol. abs/1910.07153, 2019.

[58] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

**Fang Wan** received the B.S. degree from Wuhan University, Wuhan, China, in 2013 and a Ph.D degree from University of Chinese Academy of Sciences in 2019. Since 2021, he has been an assistant professor in the School of Computer Sciences and Technology, University of Chinese Academy of Sciences, Beijing. His research interests include computer vision and machine learning, specifically for weakly supervised learning and visual object detection. He has published 20 papers in referred conferences and journals including IEEE CVPR, ICCV, and PAMI.



**Qixiang Ye** (M'10-SM'15) received the B.S. and M.S. degrees in mechanical and electrical engineering from Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences 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. His research interests include image processing, visual object detection and machine learning. He has published more than 100 papers in refereed conferences and journals including IEEE CVPR, ICCV, ECCV, NeurIPS, TNNLS, TIP, and PAMI. He is on the editorial boards of IEEE Transactions on Circuit and Systems on Video Technology and IEEE Transactions on Intelligent Transportation Systems.

**Tianning Yuan** received the B.S. degree from Tsinghua University, Beijing, China, in 2019. Since 2019, he has been a Master student in 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 active learning and self-supervised learning.

**Songcen Xu** received his Bachelor Degree in Wuhan University of Technology in 2008, and the M.Sc. degree in communications engineering and the Ph.D. degree in electronic engineering from the University of York, U.K., in 2011 and 2015, respectively. He is working as a principal research engineer at Noah' Ark Lab, Huawei technologies, in the area of Computer Vision.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3277738



**Jianzhuang Liu** (Senior Member, IEEE) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong in 1997. From 1998 to 2000, he was a Research Fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a Post-Doctoral Fellow, an Assistant Professor, and an Adjunct Associate Professor with The Chinese University of Hong Kong, Hong Kong. In 2011, he joined the Shenzhen Institute of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, as a Professor. He is a Principal Researcher with Huawei Technologies Company Ltd., Shenzhen. He has authored more than 150 papers in the areas of computer vision, image processing, deep learning, and graphics.

**Xiangyang Ji** (M'10) 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. He relieved the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences. He joined Tsinghua University, Beijing, in 2008, where he is a Professor with the Department of Automation, School of Information Science and Technology. He has authored over 100 referred conference and journal papers. His research interests include signal processing, image/video compressing, and intelligent imaging.

**Qingming Huang** (F'18) Qingming Huang is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He received a Bachelor degree in Computer Science in 1988 and a Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia computing, image processing, computer vision and pattern recognition. He has published more than 400 academic papers in prestigious international journals. He has served as general chair, program chair, track chair and TPC member for ACM Multimedia, CVPR, ICCV, ICME, and ICMR. He is a Fellow of IEEE.