

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065

## Abstract

Despite the substantial progress of active learning for image recognition, there still lacks an instance-level active learning method specified for object detection. In this paper, we propose Multiple Instance Active Learning (MIAL), to select the most informative images for detector training by observing instance-level uncertainty. MIAL defines an instance uncertainty learning module, which leverages the discrepancy of two adversarial instance classifiers trained on the labeled set to predict instance uncertainty of the unlabeled set. MIAL treats unlabeled images as instance bags and feature anchors in images as instances, and estimates the image uncertainty by re-weighting instances in a multiple instance learning (MIL) fashion. Iterative instance uncertainty learning and re-weighting facilitate suppressing noisy instances, toward bridging the gap between instance uncertainty and image-level uncertainty. Experiments validate that MIAL sets a solid baseline for instance-level active learning. On commonly used object detection datasets, MIAL outperforms state-of-the-art methods with significant margins, particularly when the labeled sets are small. Code is available at <https://github.com/CVPR2021-ID453/MIAL>.

## 1. Introduction

The key idea behind active learning is that a machine learning algorithm can achieve better performance with fewer training samples if it is allowed to select the data it wants to learn from. Despite the rapid progress of learning methods with less supervision, e.g., weakly supervised learning and semi-supervised learning, active learning remains the cornerstone of many practical applications for its simplicity and higher performance bound.

In the computer vision area, active learning has been widely explored for image classification (active image classification) by empirically generalizing the model trained on the labeled sets to the unlabeled sets [9, 28, 18, 35, 4, 22, 34, 23, 32]. Uncertainty-based methods define various metrics for selecting informative images and adapting trained models to the unlabeled set [9]. Distribution-based approaches

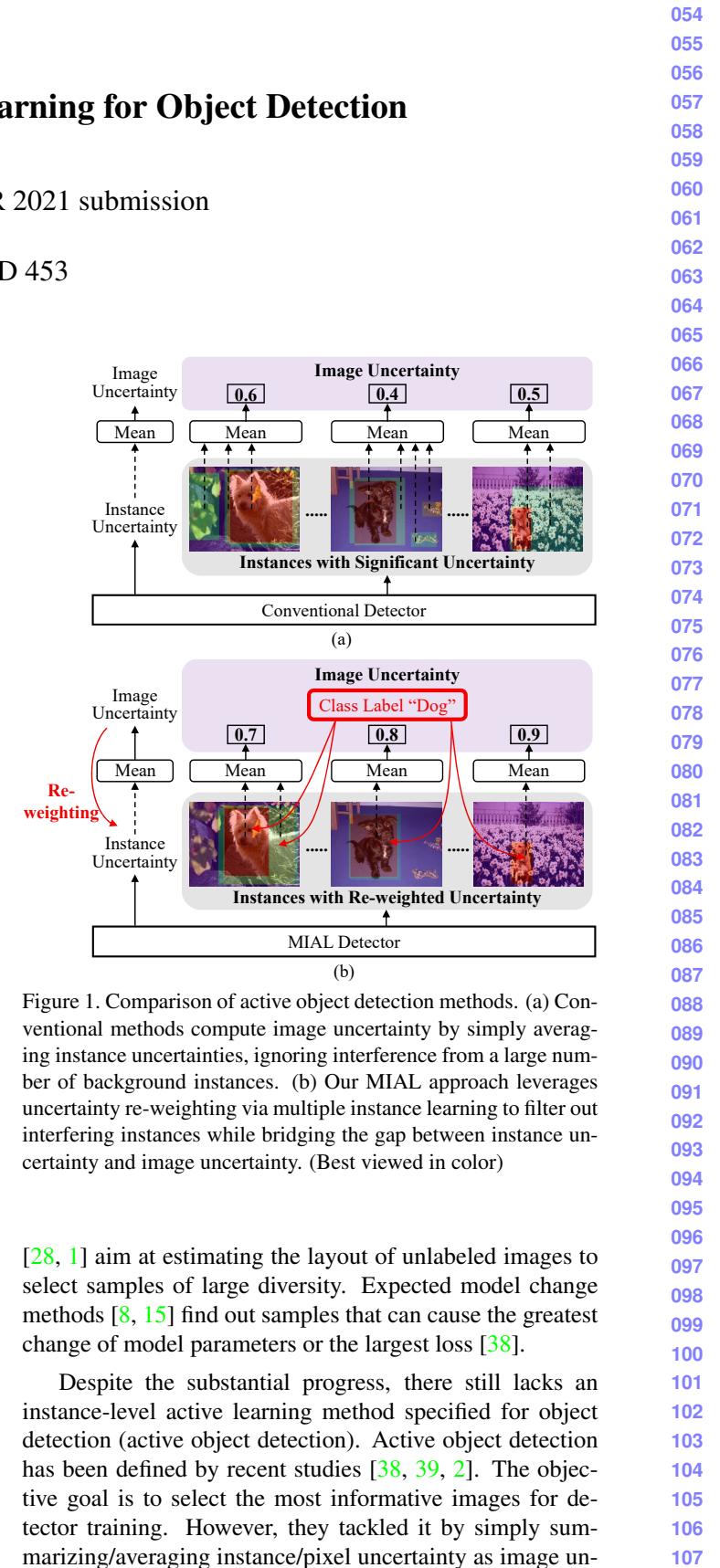


Figure 1. Comparison of active object detection methods. (a) Conventional methods compute image uncertainty by simply averaging instance uncertainties, ignoring interference from a large number of background instances. (b) Our MIAL approach leverages uncertainty re-weighting via multiple instance learning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

[28, 1] aim at estimating the layout of unlabeled images to select samples of large diversity. Expected model change methods [8, 15] find out samples that can cause the greatest change of model parameters or the largest loss [38].

Despite the substantial progress, there still lacks an instance-level active learning method specified for object detection (active object detection). Active object detection has been defined by recent studies [38, 39, 2]. The objective goal is to select the most informative images for detector training. However, they tackled it by simply summarizing/averaging instance/pixel uncertainty as image un-

certainty and unfortunately ignored the large imbalance of negative instances in object detection, which causes significant noisy instances in the background and interferes with the learning of image uncertainty, Fig. 1(a).

In this paper, we propose a Multiple Instance Active Learning (MIAL) approach for object detection, Fig. 1(b), and target at selecting informative images from the unlabeled set by learning and re-weighting instance uncertainty with discrepancy learning and multiple instance learning (MIL). To learn the instance-level uncertainty, MIAL first defines an instance uncertainty learning (IUL) module, which leverages two adversarial instance classifiers plugged atop the detection network (*e.g.*, a feature pyramid network) to learn the uncertainty of unlabeled instances. Maximizing the prediction discrepancy of two instance classifiers predicts instance uncertainty while minimizing classifiers' discrepancy drives learning features to reduce the distribution bias between labeled and unlabeled instances.

To establish the relationship between instance uncertainty and image uncertainty, MIAL incorporates a MIL module, which is in parallel with the instance classifiers. MIL treats each unlabeled image as an instance bag and performs instance uncertainty re-weighting (IUR) by evaluating instance appearance consistency across images. During MIL, the instance uncertainty and image uncertainty are forced to be consistently driven by a classification loss defined on image class labels (or pseudo-labels). Optimizing the image-level classification loss facilitates suppressing the noisy instances while highlighting truly representative ones. Iterative instance uncertainty learning and instance uncertainty re-weighting bridge the gap between instance-level observation and image-level evaluation, towards selecting the most informative images for detector training.

The contributions of this paper include:

(1) We propose Multiple Instance Active Learning (MIAL), establishing a solid baseline to model the relationship between the instance uncertainty and image uncertainty for informative image selection.

(2) We design instance uncertainty learning (IUL) and instance uncertainty re-weighting (IUR) modules, providing effective approaches to highlight informative instances while filtering out noisy ones in object detection.

(3) We apply MIAL to object detection on commonly used datasets, improving state-of-the-art methods with significant margins.

## 2. Related Work

### 2.1. 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 methods are mostly proposed for image classifi-

cation, which can be roughly categorized into uncertainty-based and distribution-based.

**Uncertainty-based Methods.** Uncertainty is the most popular metric to select samples for active learning [29], which can be defined as the posterior probability of a predicted class [17, 16], or the margin between the posterior probabilities of the first and the second predicted class [14, 26]. It can also be defined upon entropy [30, 24, 14], which measures the variance of unlabeled samples. The expected model change methods [27, 31] utilized the present model to estimate the expected network gradient or expected prediction changes [8, 15], which guide the sample selection. MIL-based methods [31, 13, 36, 6] selected informative images by discovering representative instances. However, these methods are designed for image classification and are not applicable to object detection due to the challenging aspect of the crowded and noisy instances.

**Distribution-based Methods.** This line of methods selects diverse and informative samples by estimating the distribution of unlabeled samples. Clustering methods [25] were applied to build the unlabeled sample distribution while discrete optimization methods [11, 7, 37] were employed to perform sample selection. By considering the distances to their surrounding samples, the context-aware methods [12, 3] selected the samples that can represent the global sample distribution. Core-set [28] defined the problem of active learning as core-set selection, *i.e.*, choosing a set of points such that a model learned on the labeled subset captures the diversity of the unlabeled samples.

In the deep learning era, active learning methods remain falling into the uncertainty-based or distribution-based routines [18, 35, 4]. Sophisticated methods have extended active learning to the open sets [22], or combined it with self-paced learning [34]. Nevertheless, it remains questionable whether or not the intermediate feature representation is effective for sample selection. The learning loss approach [38] can be categorized as either uncertainty-based or distribution-based. By introducing a network structure to predict the “loss” of the unlabeled samples, it estimates sample uncertainty and distribution, and selects samples with large “loss” like hard negative mining.

### 2.2. Active Learning for Object Detection

Despite the substantial progress of active learning, few methods were specified for active object detection, which faces complex instance distributions in the same images and thereby is far more challenging than active image classification. By simply sorting the loss predictions of instances to evaluate the image uncertainty, the learning loss method [38] specified for image classification was directly applied to object detection. In [2], the image-level uncertainty was estimated by the uncertainty of a large number of background pixels. The CDAL approach [1] introduced

162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215

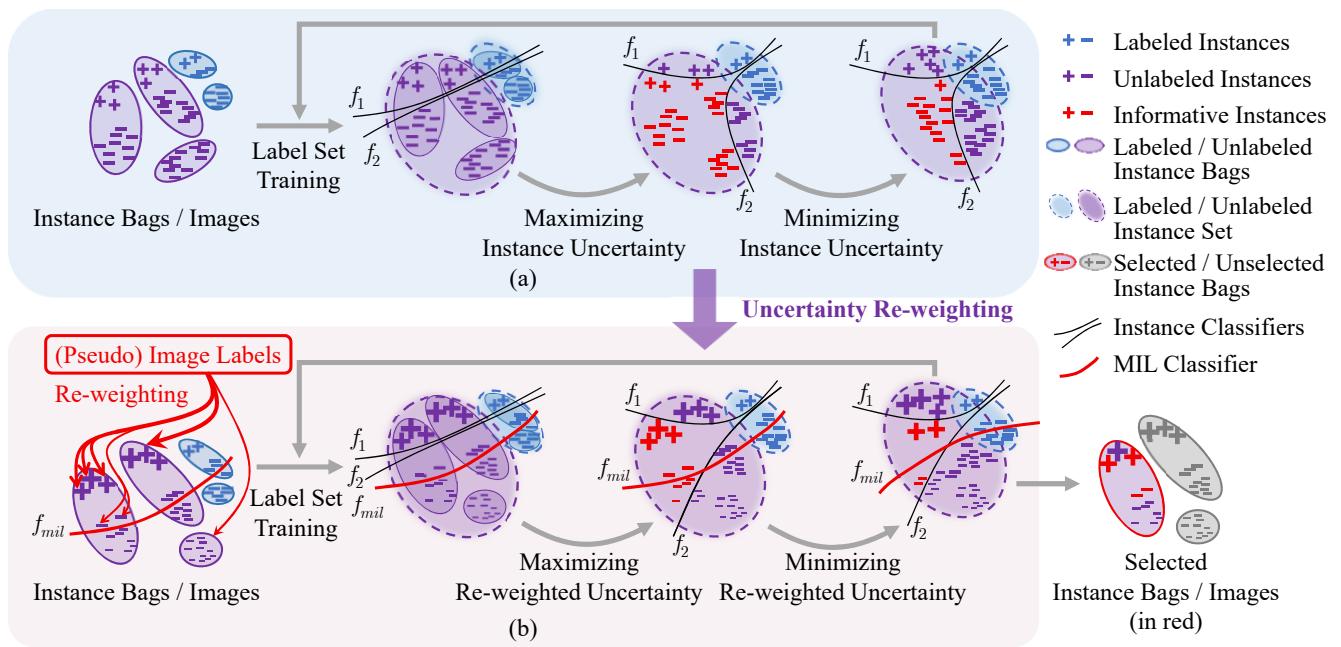


Figure 2. MIAL illustration. (a) Instance uncertainty learning (IUL) utilizing two adversarial classifiers. (b) Instance uncertainty re-weighting (IUR) using multiple instance learning. Bigger symbols (“+” and “-”) indicate larger weights. (Best viewed in color)

spatial context to active detection and selected diverse samples according to their distances to the labeled set. Existing approaches simply used instance/pixel-level observations to represent the image-level uncertainty. There still lacks a systematic method to learn the image uncertainty by leveraging instance-level models (*e.g.*, object detectors).

### 3. The Proposed Approach

#### 3.1. Overview

For active object detection, a small set of images  $\mathcal{X}_L^0$  (the labeled set) with instance labels  $\mathcal{Y}_L^0$  and a large set of images  $\mathcal{X}_U^0$  (the unlabeled set) without labels are given. For each image, the label consists of bounding boxes ( $y_x^{loc}$ ) and categories ( $y_x^{cls}$ ) for objects of interest. A detection model  $M_0$  is firstly initialized by using the labeled set  $\{\mathcal{X}_L^0, \mathcal{Y}_L^0\}$ . With the initialized model  $M_0$ , active learning targets at selecting a set of images  $\mathcal{X}_S^0$  from  $\mathcal{X}_U^0$  to be manually labeled and merging them with  $\mathcal{X}_L^0$  to form the new 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  $\mathcal{X}_S^0$  should be the most informative, *i.e.*, can improve the detection performance as much as possible. Based on the updated labeled set  $\mathcal{X}_L^1$ , the task model is retrained and updated to  $M_1$ . The detection model training and sample selection procedures repeat several cycles until the number of labeled images reaches the annotation budget.

Considering the large number<sup>1</sup> of instances in each im-

<sup>1</sup>For example, the RetinaNet detector [19] produces ~100k of anchors (instances) for an image.

age, there are two key problems for active object detection: (1) how to evaluate the uncertainty of the unlabeled instances using the detector trained on the labeled set; (2) how to precisely estimate the image uncertainty while filtering out noisy instances. MIAL handles these two problems by introducing two learning modules respectively. For the first problem, MIAL incorporates instance uncertainty learning, with the aim of highlighting informative instances in the unlabeled set, as well as aligning the distributions of the labeled and unlabeled set, Fig. 2(a). It is motivated by the fact that most active learning methods remain simply generalizing the models trained on the labeled set to the unlabeled set. This is problematic when there is a distribution bias between the two sets [10]. For the second problem, MIAL introduces MIL to both the labeled and unlabeled set to estimate the image uncertainty by re-weighting the instance uncertainty. This is done by treating each image as an instance bag while re-weighting the instance uncertainty under the supervision of the image classification loss. Optimizing the image classification loss facilitates highlighting truly representative instances belonging to the same object classes while suppressing the noisy ones, Fig. 2(b).

#### 3.2. Instance Uncertainty Learning

**Label Set Training.** Using the RetinaNet as the baseline [19], we construct a detector with two discrepant instance classifiers ( $f_1$  and  $f_2$ ) and a bounding box regressor ( $f_r$ ), Fig. 3(a). We utilize the prediction discrepancy between the two instance classifiers to learn the instance

216		270
217		271
218		272
219		273
220	Labeled Instances	274
221	Unlabeled Instances	275
222	Informative Instances	276
223	Labeled / Unlabeled Instance Bags	277
224	Labeled / Unlabeled Instance Set	278
225	Selected / Unselected Instance Bags	279
226	Instance Classifiers	280
227	MIL Classifier	281
228	(Pseudo) Image Labels	282
229	Re-weighting	283
230	Label Set Training	284
231	$f_1$	285
232	$f_2$	286
233	$f_{mil}$	287
234	Maximizing Re-weighted Uncertainty	288
235	Minimizing Re-weighted Uncertainty	289
236		290
237		291
238		292
239		293
240		294
241		295
242		296
243		297
244		298
245		299
246		300
247		301
248		302
249		303
250		304
251		305
252		306
253		307
254		308
255		309
256		310
257		311
258		312
259		313
260		314
261		315
262		316
263		317
264		318
265		319
266		320
267		321
268		322
269		323

324

325

326

327

328

329

330

331

332

333

334

335

336

uncertainty on the unlabeled set. Let  $g$  denote the feature extractor parameterized by  $\theta_g$ . The discrepant classifiers are parameterized by  $\theta_{f_1}$  and  $\theta_{f_2}$  and the regressor by  $\theta_{f_r}$ .  $\Theta = \{\theta_{f_1}, \theta_{f_2}, \theta_{f_r}, \theta_g\}$  denotes the set of all parameters, where  $\theta_{f_1}$  and  $\theta_{f_2}$  are independently initialized.

In object detection, each image  $x$  from the labeled set  $\mathcal{X}_L$  can be represented by multiple instances  $\{x_i, i = 1, \dots, N\}$  corresponding to the feature anchors on the feature map [19].  $N$  is the number of the instances in image  $x$ .  $\{y_i, i = 1, \dots, N\}$  denote the labels for the instances in the image  $x$ . Given the labeled set, a detection model is trained by optimizing the following detection loss, as

$$\operatorname{argmin}_{\Theta} l_{det}(x) = \sum_i \left( FL(\hat{y}_i^{f_1}, y_i^{cls}) + FL(\hat{y}_i^{f_2}, y_i^{cls}) + SmoothL1(\hat{y}_i^{f_r}, y_i^{loc}) \right), \quad (1)$$

where  $FL(\cdot)$  is the focal loss function for instance classification and  $SmoothL1(\cdot)$  is the smooth L1 loss function for bounding-box regression [19].  $\hat{y}_i^{f_1} = f_1(g(x_i))$ ,  $\hat{y}_i^{f_2} = f_2(g(x_i))$  and  $\hat{y}_i^{f_r} = f_r(g(x_i))$  denote the prediction (classification and localization) results for the instances.  $y_i^{cls}$  and  $y_i^{loc}$  denote the ground-truth class label and bounding box label, respectively.

**Maximizing Instance Uncertainty.** Before the labeled set can precisely represent the unlabeled set, it is common that there exists a distribution bias between the labeled and unlabeled set, especially when the labeled set is small. The informative instances should be localized in the biased distribution region. To find out these instances,  $f_1$  and  $f_2$  are designed as two adversarial instance classifiers which tend to have larger prediction discrepancy upon instances close to the class boundary, Fig. 2(a). The instance uncertainty is defined as the prediction discrepancy of  $f_1$  and  $f_2$  on the unlabeled set.

To find out the most informative instances, it requires to fine-tune the network and maximize the prediction discrepancy of the adversarial classifiers, Fig. 3(b). In the maximizing procedure,  $\theta_g$  is fixed so that the distributions of both the labeled and unlabeled instances are fixed.  $\theta_{f_1}$  and

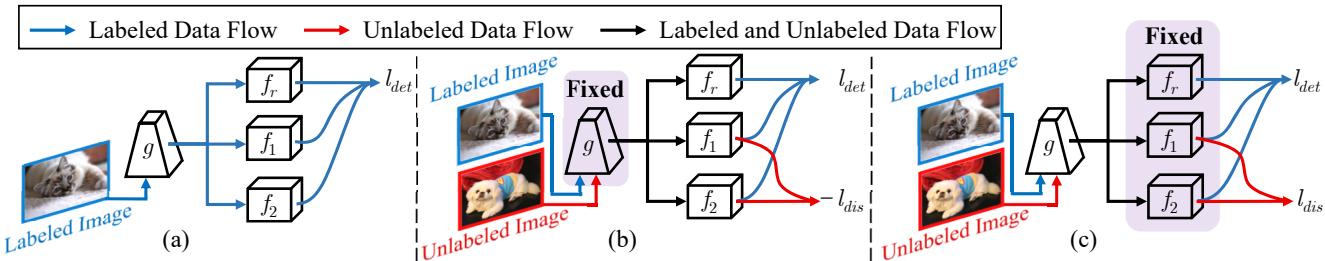


Figure 3. Network architecture for instance uncertainty learning. (a) Label set training. (b) Maximizing instance uncertainty by maximizing classifier prediction discrepancy. (c) Minimizing instance uncertainty by minimizing classifier prediction discrepancy.

$\theta_{f_2}$  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

$$\operatorname{argmin}_{\Theta \setminus \theta_g} \mathcal{L}_{max} = \sum_{x \in \mathcal{X}_L} l_{det}(x) - \sum_{x \in \mathcal{X}_U} \lambda \cdot l_{dis}(x), \quad (2)$$

where

$$l_{dis}(x) = \sum_i |\hat{y}_i^{f_1} - \hat{y}_i^{f_2}| \quad (3)$$

denotes the prediction discrepancy loss defined on the prediction discrepancy.  $\hat{y}_i^{f_1}, \hat{y}_i^{f_2} \in \mathbb{R}^{1 \times C}$  are the instance classification predictions of the two classifiers for the  $i$ -th instance in image  $x$ , where  $C$  is the number of object classes in the dataset.  $\lambda$  is an experimentally determined regularization parameter. As shown in Fig. 2(a), the informative instances with different predictions by the adversarial classifiers tend to have larger prediction discrepancy and are assigned larger uncertainty scores.

**Minimizing Instance Uncertainty.** After maximizing the prediction discrepancy, we further propose to minimize the prediction discrepancy to align the distributions of the labeled and unlabeled instances, 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 optimized by minimizing the prediction discrepancy loss, as

$$\operatorname{argmin}_{\theta_g} \mathcal{L}_{min} = \sum_{x \in \mathcal{X}_L} l_{det}(x) + \sum_{x \in \mathcal{X}_U} \lambda \cdot l_{dis}(x). \quad (4)$$

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.

In each active learning cycle, the max-min prediction discrepancy procedures repeat several times so that the instance uncertainty is learned and the instance distributions of the labeled and unlabeled set 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.

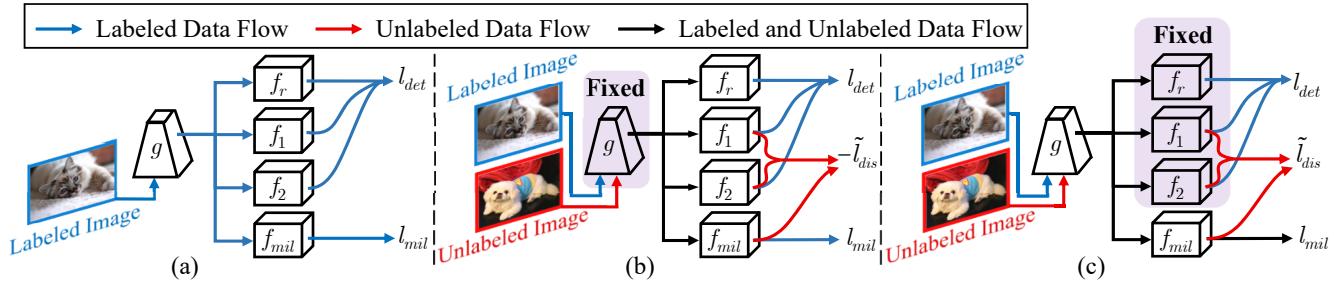


Figure 4. Network architecture for instance uncertainty re-weighting. (a) Label set training. (b) Re-weighting and maximizing instance uncertainty. (c) Re-weighting while minimizing instance uncertainty.

### 3.3. Instance Uncertainty Re-weighting

With instance uncertainty learning, the informative instances are highlighted. However, as there is a large amount ( $\sim 100k$ ) of instances in each image, the instance uncertainty may be not consistent with the image uncertainty. Some instances of high uncertainty are simply background noise or hard negatives for the detector. We thereby introduce an MIL procedure to bridge the gap between the instance-level and image-level uncertainty by filtering out noisy instances.

**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 re-weights the instance uncertainty scores by minimizing the image classification loss. This actually defines an Expectation-Maximization procedure [33, 5] to re-weight instance uncertainty across bags while filtering out noisy instances.

Specifically, we add an MIL classifier  $f_{mil}$  parameterized by  $\theta_{f_{mil}}$  in parallel with the instance classifiers, Fig. 4.  $\Theta$  is then updated as  $\tilde{\Theta} = \Theta \cup \{\theta_{f_{mil}}\}$ . The MIL score  $\hat{y}_{i,c}^{f_{mil}}$  for multiple instances in an image is calculated as

$$\hat{y}_{i,c}^{f_{mil}} = \frac{\exp(s_{i,c})}{\sum_c \exp(s_{i,c})} \cdot \frac{\exp((\hat{y}_{i,c}^{f_1} + \hat{y}_{i,c}^{f_2})/2)}{\sum_i \exp((\hat{y}_{i,c}^{f_1} + \hat{y}_{i,c}^{f_2})/2)}, \quad (5)$$

where  $s = f_{mil}(g(x))$  is an  $N \times C$  score matrix, and  $s_{i,c}$  is the element in  $s$  indicating the score of the  $i$ -th instance for class  $c$ . According to Eq. (5), the MIL score  $\hat{y}_{i,c}^{f_{mil}}$  is large only when  $x_i$  belongs to class  $c$  (first term in Eq. (5)) and its instance classification scores  $\hat{y}_{i,c}^{f_1}$  and  $\hat{y}_{i,c}^{f_2}$  are significantly larger than those of others (second term in Eq. (5)).

Considering that the MIL scores of the instances from other classes/backgrounds are small, the MIL loss  $l_{mil}$  is defined by minimizing the image classification loss using the MIL score, as

$$l_{mil}(x) = - \sum_c \left( y_c^{cls} \log \sum_i \hat{y}_{i,c}^{f_{mil}} + (1 - y_c^{cls}) \log(1 - \sum_i \hat{y}_{i,c}^{f_{mil}}) \right), \quad (6)$$

where  $y_c^{cls} \in \{0, 1\}$  denotes the image class label, which can be directly obtained using the instance class label  $y_i^{cls}$  in the labeled set. Optimizing Eq. (6) drives the MIL classifier to activate instances with both large MIL score ( $s_{i,c}$ ) and classification outputs ( $\hat{y}_{i,c}^{f_1} + \hat{y}_{i,c}^{f_2}$ ). The instances with large classification outputs but small MIL scores ( $s_{i,c}$ ) will be suppressed as background. The MIL loss is firstly applied in the label set training procedure to get the initial detector, and then used to re-weight the instance uncertainty in the unlabeled set.

**Uncertainty Re-weighting.** To ensure that the instance uncertainty is consistent with the image uncertainty, we assemble the MIL scores for all classes to a score vector  $w_i$  and re-weight the instance uncertainty as

$$\tilde{l}_{dis}(x) = \sum_i |w_i \cdot (\hat{y}_i^{f_1} - \hat{y}_i^{f_2})|, \quad (7)$$

where  $w_i = \hat{y}_i^{f_{mil}}$ . We then update Eq. (2) to

$$\operatorname{argmin}_{\tilde{\Theta} \setminus \theta_g} \tilde{\mathcal{L}}_{max} = \sum_{x \in \mathcal{X}_L} (l_{det}(x) + l_{mil}(x)) - \sum_{x \in \mathcal{X}_U} \lambda \cdot \tilde{l}_{dis}(x). \quad (8)$$

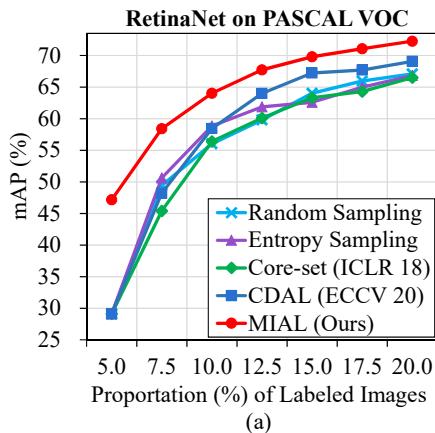
By optimizing Eq. (8), the discrepancies of instances with large MIL scores are preferentially estimated, while those with small MIL scores are suppressed. Similarly, Eq. (4) is updated to

$$\begin{aligned} \operatorname{argmin}_{\theta_g, \theta_{f_{mil}}} \tilde{\mathcal{L}}_{min} &= \sum_{x \in \mathcal{X}_L} (l_{det}(x) + l_{mil}(x)) \\ &+ \sum_{x \in \mathcal{X}_U} (\lambda \cdot \tilde{l}_{dis}(x) + l_{mil}(x)). \end{aligned} \quad (9)$$

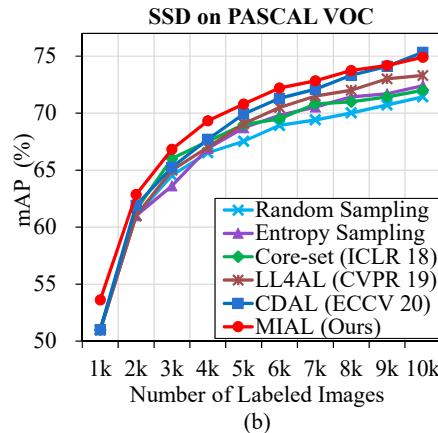
In Eq. (9), the MIL loss is applied to the unlabeled set, where the pseudo image labels are estimated using the outputs of the instance classifiers, as

$$y_c^{pseudo} = \mathbb{1} \left( \max_i \left( \frac{\hat{y}_{i,c}^{f_1} + \hat{y}_{i,c}^{f_2}}{2} \right), 0.5 \right), \quad (10)$$

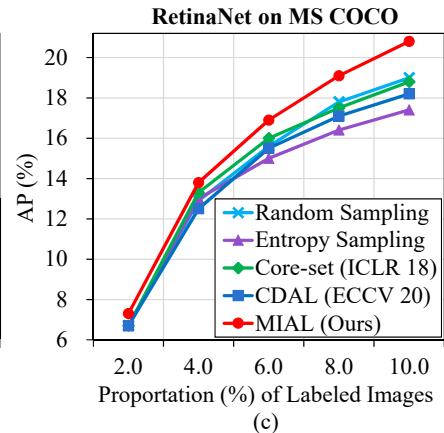
where  $\mathbb{1}(a, b)$  is a binarization function. When  $a > b$ , it returns 1; otherwise 0. Eq. (10) is defined based on that

540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553

(a)



(b)



(c)

Figure 5. Performance comparison of active object detection methods. (a) On PASCAL VOC using RetinaNet backbone. (b) On PASCAL VOC using SSD backbone. (c) On MS COCO using RetinaNet backbone.

instance classifiers can find true instances but are easy to be confused by complex backgrounds. Therefore, we use the maximum instance score to predict pseudo image labels and then leverage MIL to reduce background interference. According to Eqs. (5) and (6), the MIL loss ensures the highlighted instances are representative of the image, *i.e.*, minimizing the image classification loss so that the gap between the instance uncertainty and image uncertainty is minimized. By iteratively optimizing Eqs. (8) and (9), informative object instances with the same class are statistically highlighted, while the background instances are suppressed during the instance uncertainty learning procedure.

### 3.4. Informative Image Selection

In each active learning cycle, after the instance uncertainty learning (IUL) and the instance uncertainty re-weighting (IUR) procedures, we select the most informative images from the unlabeled set by observing the top- $k$  instance uncertainty defined in Eq. (3) for each image, where  $k$  is a hyperparameter. This is based on the fact that the noisy instances have been suppressed and the instance uncertainty has been consistent with the image uncertainty. The selected images are merged into the labeled set for the next learning cycle.

## 4. Experiments

In this section, we firstly introduce the experimental settings. We then report the detection performance of MIAL and compare it with state-of-the-art methods. We finally present ablation study with visualization analysis.

### 4.1. Experimental Settings

**Datasets.** The *trainval* sets of PASCAL VOC 2007 and 2012 datasets are used as the training set, which contain 5011 and 11540 images, respectively. The VOC 2007 *test*

set is used to evaluate the detection performance (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 the detection performance (average precision (AP)).

**Active Learning Settings.** We use the RetinaNet [19] with ResNet-50 and SSD [21] with VGG-16 as the base detector. For RetinaNet, MIAL 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% images from the rest unlabeled set until the labeled images reach 20.0% of the training set. For the large-scale MS COCO, MIAL uses only 2.0% of randomly selected images from the training set to initialize the labeled set. It then selects 2.0% 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 0.0005 respectively. For SSD, we follow the settings in [38] and [1], where 1k images in PASCAL VOC dataset 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 [38].

We compare MIAL with random sampling, entropy sampling, Core-set [28], LL4AL [38] and CDAL [1]. 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. MIAL and other methods share the same random seed and initialization for a fair comparison.  $\lambda$  defined in Eqs. (2), (4), (8) and (9) is set to 0.5 and  $k$  mentioned in Sec. 3.4 is set to 10k.

Training		Sample Selection			mAP (%) on Proportion (%) of Labeled Images							
IUL	IUR	Rand.	Max Unc.	Mean Unc.	5.0	7.5	10.0	12.5	15.0	17.5	20.0	100.0
		✓			28.31	49.42	56.03	59.81	64.02	65.95	67.09	77.28
✓		✓			30.09	49.17	55.64	60.93	64.10	65.77	67.20	
✓			✓		30.09	49.79	58.94	63.11	65.61	67.84	69.01	
✓				✓	30.09	49.74	60.60	64.29	67.13	68.76	70.06	
		✓	✓		47.18	57.12	60.68	63.72	66.10	67.59	68.48	78.37
		✓		✓	47.18	57.58	61.74	64.58	66.98	68.79	70.33	
		✓		✓	<b>47.18</b>	<b>58.03</b>	<b>63.98</b>	<b>66.58</b>	<b>69.57</b>	<b>70.96</b>	<b>72.03</b>	

Table 1. Module ablation on PASCAL VOC. The first line shows the result of the baseline method with random image selection. “Max Unc.” and “Mean Unc.” respectively denote that the image uncertainty is represented by the maximum and averaged instance uncertainty.

Training	mAP (%) on Proportion (%) of Labeled Img.				
IUL	2.0	4.0	6.0	8.0	10.0
	51.01	61.48	69.14	75.14	79.77
✓	<b>58.07</b>	<b>67.75</b>	<b>74.91</b>	<b>78.88</b>	<b>80.96</b>

Table 2. The effect of IUL for active image classification. Experiments are conducted on CIFAR-10 using the ResNet-18 backbone while the images are randomly selected in all cycles.

## 4.2. Performance

**PASCAL VOC.** In Fig. 5, we report the performance of MIAL and compare it with state-of-the-art methods. Using either the RetinaNet [19] or SSD [20] detector, MIAL outperforms state-of-the-art methods with large margins. Particularly, it respectively outperforms state-of-the-art methods by 18.08%, 7.78%, and 5.19% when using 5.0%, 7.5%, and 10.0% samples. In the last cycle, with 20.0% samples, MIAL achieves 72.27% detection mAP, which significantly outperforms CDAL by 3.20%. The improvements validate that MIAL can precisely learn instance uncertainty while selecting informative images. When using the SSD detector, MIAL outperforms state-of-the-art methods in almost all cycles, demonstrating the general applicability of MIAL to object detectors.

**MS COCO.** MS COCO is a challenging dataset for more categories, denser objects, and larger scale variation, where MIAL also outperforms the compared methods, Fig. 5. Particularly, it respectively outperforms Core-set and CDAL by 0.6%, 0.5%, and 2.0%, and 0.6%, 1.3%, and 2.6% when using 2.0%, 4.0%, and 10.0% labeled images.

## 4.3. Ablation Study

**IUL.** As shown in Tab. 1, with IUL, the detection performance is improved up to 70.06% in the last cycle, which

$w_i$	Set	mAP (%) on Proportion (%) of Labeled Img.						
		5.0	7.5	10.0	12.5	15.0	20.0	
1	$\emptyset$	30.09	49.17	55.64	60.93	64.10	65.77	67.20
$\hat{y}_i^{f_1}$	$\emptyset$	31.67	50.67	55.93	60.78	64.17	66.22	67.30
1	$\mathcal{X}_L$	42.52	54.08	57.18	63.43	65.04	66.74	68.32
$\hat{y}_i^{f_{mil}}$	$\mathcal{X}$	<b>47.18</b>	<b>57.12</b>	<b>60.68</b>	<b>63.72</b>	<b>66.10</b>	<b>67.59</b>	<b>68.48</b>

Table 3. Ablation study on IUR. “ $w_i$ ” is the  $w_i$  in Eq. (7). “Set” denotes the sample set for IUR.  $\mathcal{X}$  and  $\mathcal{X}_L$  denote the whole sample set and the labeled set, respectively.

outperforms the Random method by 2.97% (70.06% vs. 67.09%). In Tab. 2, IUL also significantly improves the image classification performance with active learning on CIFAR-10. Particularly when using 2.0% samples, it improves the classification performance by 7.06% (58.07% vs. 51.01%), demonstrating the effectiveness of the discrepancy learning module for instance uncertainty estimation.

**IUR.** In Tab. 1, IUL achieves comparable performance with the Random 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 the image uncertainty. After using IUR to re-weight instance uncertainty, the performance at early cycles is largely improved by 5.04%~17.09% in the first three cycles (row 4 vs. row 1 in Tab. 3). In the last cycle, the performance is improved by 1.28% (68.48% vs. 67.20%) in comparison with IUL and 1.39% in comparison with the Random method (68.48% vs. 67.09%). As shown in Tab. 3, MIL score  $\hat{y}_i^{f_{mil}}$  is the best re-weighting metric (row 4 vs. others). Interestingly, when using 100.0% images for training, the detector with IUR outperforms the detector without IUR by 1.09% (78.37% vs. 77.28%). These results clearly verify that the IUR module can suppress the interfering in-

<small>756</small>	<small>λ</small>	<small>k</small>	mAP (%) on Proportion (%) of Labeled Imgs.							<small>810</small>
<small>757</small>	<small>λ</small>	<small>k</small>	5.0	7.5	10.0	12.5	15.0	17.5	20.0	<small>811</small>
<small>759</small>	2	10k	47.18	56.94	64.44	67.70	69.58	70.67	72.12	<small>812</small>
<small>760</small>	1	10k	47.18	57.30	<b>64.93</b>	67.40	69.63	70.53	71.62	<small>813</small>
<small>761</small>	0.5	10k	47.18	<b>58.41</b>	64.02	<b>67.72</b>	<b>69.79</b>	<b>71.07</b>	<b>72.27</b>	<small>814</small>
<small>762</small>	0.2	10k	47.18	58.02	64.44	67.67	69.42	70.98	72.06	<small>815</small>
<small>763</small>	0.5	<i>N</i>	47.18	58.03	63.98	66.58	69.57	70.96	72.03	<small>816</small>
<small>764</small>	0.5	10k	47.18	58.41	<b>64.02</b>	<b>67.72</b>	<b>69.79</b>	<b>71.07</b>	<b>72.27</b>	<small>817</small>
<small>765</small>	0.5	100	47.18	<b>58.74</b>	63.62	67.03	68.63	70.26	71.47	<small>818</small>
<small>766</small>	0.5	1	47.18	57.58	61.74	64.58	66.98	68.79	70.33	<small>819</small>

Table 4. Performance under different hyper-parameters.

<small>772</small> Method	Time (h) on Proportion (%) of Labeled Imgs.							<small>826</small>
	5.0	7.5	10.0	12.5	15.0	17.5	20.0	<small>827</small>
<small>775</small> Random	0.77	1.12	1.45	1.78	2.12	2.45	2.78	<small>828</small>
<small>776</small> CDAL [1]	1.18	1.50	1.87	2.19	2.68	<b>2.83</b>	<b>2.82</b>	<small>829</small>
<small>777</small> MIAL	<b>1.03</b>	<b>1.42</b>	<b>1.78</b>	<b>2.18</b>	<b>2.55</b>	2.93	3.12	<small>830</small>

Table 5. Comparison of time cost on PASCAL VOC.

stances while highlighting more representative ones, which can indicate informative images for detector training.

**Parameters and Time Cost.** The effects of the regularization factor  $\lambda$  defined in Eqs. (2), (4), (8) and (9) and the valid instance number  $k$  in each image for selection are shown in Tab. 4. One can see that we have the best performance when  $\lambda$  is set to 0.5 and  $k$  is set to 10k (for  $\sim$ 100k instances/anchors in each image). Tab. 5 shows that MIAL costs less time at early cycles and slightly more time at later cycles than CDAL.

#### 4.4. Model Analysis

**Visualization Analysis.** In Fig. 6, we visualize the learned and re-weighted uncertainty and MIL scores of instances. The heatmaps are calculated by summarizing the uncertainty scores of all instances. With only IUL, 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 scores to the instances of interesting while suppressing backgrounds. As a result, IUR leverages the MIL scores to re-weight instances towards accurate instance uncertainty prediction.

**Statistical Analysis.** In Fig. 7, we calculate the number of true positive instances selected in each active learning cycle. It can be seen that MIAL hits significantly more true positives in all learning cycles. This shows that the

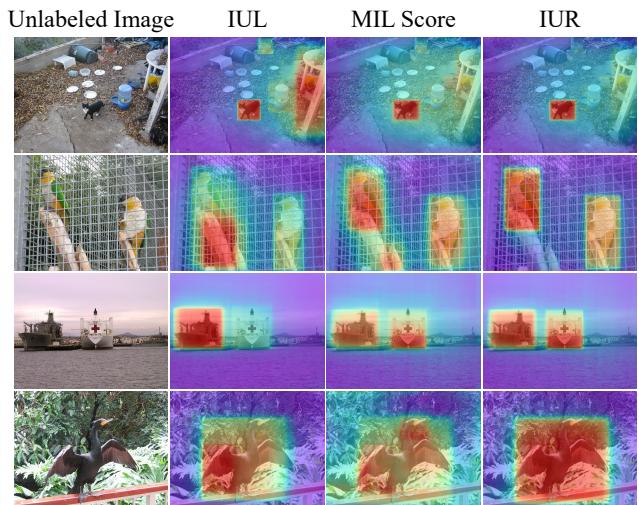


Figure 6. Visualization of learned and re-weighted instance uncertainty and MIL score. (Best viewed in color)

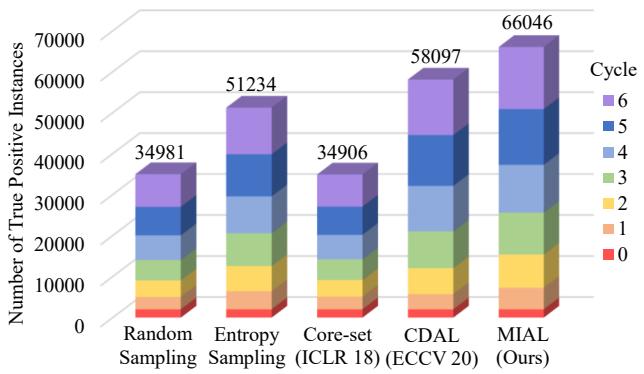


Figure 7. The number of true positive instances selected in each active learning cycle on PASCAL VOC using RetinaNet backbone.

proposed MIAL approach can better activate true positive objects while filtering out interfering instances, which facilitates selecting informative images for detector training.

## 5. Conclusion

We have proposed Multiple Instance Active Learning (MIAL) to select informative images for detector training by observing instance uncertainty. MIAL incorporates a discrepancy learning module, which leverages adversarial instance classifiers to learn the uncertainty of unlabeled instances. MIAL treats the unlabeled images as instance bags and estimates the image uncertainty by re-weighting instances in a multiple instance learning (MIL) fashion. Iterative instance uncertainty learning and instance uncertainty re-weighting facilitate suppressing noisy instances, towards selecting informative images for detector training. Experiments on large-scale datasets have validated the superiority of MIAL, in striking contrast with state-of-the-art methods. MIAL sets a solid baseline for active object detection.

864

**References**

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

- [1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. *CoRR*, abs/2008.05723, 2020. 1, 2, 6, 8
- [2] Hamed Habibi Aghdam, Abel Gonzalez-Garcia, Antonio M. López, and Joost van de Weijer. Active learning for deep detection neural networks. In *ICCV*, pages 3671–3679, 2019. 1, 2
- [3] Oisin Mac Aodha, Neill D. F. Campbell, Jan Kautz, and Gabriel J. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, pages 564–571, 2014. 2
- [4] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *CVPR*, pages 9368–9377, 2018. 1, 2
- [5] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 5
- [6] Marc-Andre Carbonneau, Eric Granger, and Ghyslain Gagnon. Bag-level aggregation for multiple-instance active learning in instance classification problems. *IEEE TNNLS*, 30(5):1441–1451, 2019. 2
- [7] Ehsan Elhamifar, Guillermo Sapiro, Allen Y. Yang, and S. Shankar Sastry. A convex optimization framework for active learning. In *ICCV*, pages 209–216, 2013. 2
- [8] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *ECCV*, volume 8692, pages 562–577, 2014. 1, 2
- [9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, volume 70, pages 1183–1192, 2017. 1
- [10] Denis A. Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, and Sotaro Tsukizawa. Deep active learning for biased datasets via fisher kernel self-supervision. In *CVPR*, pages 9038–9046, 2020. 3
- [11] Yuhong Guo. Active instance sampling via matrix partition. In *NeurIPS*, pages 802–810, 2010. 2
- [12] Mahmudul Hasan and Amit K. Roy-Chowdhury. Context aware active learning of activity recognition models. In *ICCV*, pages 4543–4551, 2015. 2
- [13] Sheng-Jun Huang, Nengneng Gao, and Songcan Chen. Multi-instance multi-label active learning. In *IJCAI*, pages 1886–1892, 2017. 2
- [14] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, pages 2372–2379, 2009. 2
- [15] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Active and continuous exploration with deep neural networks and expected model output changes. *CoRR*, abs/1612.06129, 2016. 1, 2
- [16] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning*, pages 148–156, 1994. 2
- [17] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12, 1994. 2

- [18] Liang Lin, Keze Wang, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Active self-paced learning for cost-effective and progressive face identification. *IEEE TPAMI*, 40(1):7–19, 2018. 1, 2
- [19] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE TPAMI*, 42(2):318–327, 2020. 3, 4, 6, 7
- [20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 7
- [21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, pages 21–37, 2016. 6
- [22] Zhao-Yang Liu and Sheng-Jun Huang. Active sampling for open-set classification without initial annotation. In *AAAI*, pages 4416–4423, 2019. 1, 2
- [23] Zhao-Yang Liu, Shao-Yuan Li, Songcan Chen, Yao Hu, and Sheng-Jun Huang. Uncertainty aware graph gaussian process for semi-supervised learning. In *AAAI*, pages 4957–4964, 2020. 1
- [24] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Latent structured active learning. In *NeurIPS*, pages 728–736, 2013. 2
- [25] Hieu Tat Nguyen and Arnold W. M. Smeulders. Active learning using pre-clustering. In *ICML*, 2004. 2
- [26] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *ECML*, volume 4212, pages 413–424, 2006. 2
- [27] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, pages 441–448, 2001. 2
- [28] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 1, 2, 6
- [29] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2012. 2
- [30] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*, pages 1070–1079, 2008. 2
- [31] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *NeurIPS*, pages 1289–1296, 2007. 2
- [32] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, pages 5971–5980, 2019. 1
- [33] Andrews Stuart, Tsochantaris Ioannis, and Hofmann Thomas. Support vector machines for multiple-instance learning. In *NeurIPS*, pages 561–568, 2002. 5
- [34] Ying-Peng Tang and Sheng-Jun Huang. Self-paced active learning: Query the right thing at the right time. In *AAAI*, pages 5117–5124, 2019. 1, 2
- [35] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE TCSVT*, 27(12):2591–2600, 2017. 1, 2

- 972 [36] Ran Wang, Xi-Zhao Wang, Sam Kwong, and Chen Xu. 1026  
973 Incorporating diversity and informativeness in multiple- 1027  
974 instance active learning. *IEEE TFS*, 25(6):1460–1475, 2017. 1028  
975 2 1029  
976 [37] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and 1030  
977 Alexander G. Hauptmann. Multi-class active learning by 1031  
978 uncertainty sampling with diversity maximization. *IJCV*, 1032  
979 113(2):113–127, 2015. 2 1033  
980 [38] Donggeun Yoo and In So Kweon. Learning loss for active 1034  
981 learning. In *CVPR*, pages 93–102, 2019. 1, 2, 6 1035  
982 [39] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng- 1036  
983 Jun Zha, and Qingming Huang. State-relabeling adversarial 1037  
984 active learning. In *CVPR*, pages 8753–8762, 2020. 1 1038  
985 1039  
986 1040  
987 1041  
988 1042  
989 1043  
990 1044  
991 1045  
992 1046  
993 1047  
994 1048  
995 1049  
996 1050  
997 1051  
998 1052  
999 1053  
1000 1054  
1001 1055  
1002 1056  
1003 1057  
1004 1058  
1005 1059  
1006 1060  
1007 1061  
1008 1062  
1009 1063  
1010 1064  
1011 1065  
1012 1066  
1013 1067  
1014 1068  
1015 1069  
1016 1070  
1017 1071  
1018 1072  
1019 1073  
1020 1074  
1021 1075  
1022 1076  
1023 1077  
1024 1078  
1025 1079