Consistency-based Active Learning for Object Detection

Weiping Yu[1], Sijie Zhu[2], Taojiannan Yang[2], Chen Chen[2]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Center for Research in Computer Vision, University of Central Florida, USA

weiping001@e.ntu.edu.sg

{sizhu,taoyang1122}@knights.ucf.edu; chen.chen@crcv.ucf.edu

Abstract

Active learning aims to improve the performance of the task model by selecting the most informative samples with a limited budget. Unlike most recent works that focus on applying active learning for image classification, we propose an effective Consistency-based Active Learning method for object Detection (CALD), which fully explores the consistency between the original and augmented data. CALD has three appealing benefits. (i) CALD is systematically designed by investigating the weaknesses of existing active learning methods, which do not take the unique challenges of object detection into account. (ii) CALD unifies box regression and classification with a single metric, which is not concerned with active learning methods for classification. CALD also focuses on the most informative local region rather than the whole image, which is beneficial for object detection. (iii) CALD not only gauges individual information for sample selection but also leverages mutual information to encourage a balanced data distribution. Extensive experiments show that CALD significantly outperforms existing state-of-the-art task-agnostic and detection-specific active learning methods on general object detection datasets. Based on the Faster R-CNN detector, CALD consistently surpasses the baseline method (random selection) by 2.9/2.8/0.8 mAP on average on PASCAL VOC 2007, PASCAL VOC 2012, and MS COCO. Code is available at https://github.com/we1pingyu/CALD

1. Introduction

One of the biggest bottlenecks of learning-based computer vision algorithms is the scale of annotated data. Recently, various learning methods, e.g. semi-supervised learning [2, 30, 41] and unsupervised learning [5, 6, 15], exploit information from unlabeled data to avoid the expensive cost of annotating data. Given a fixed labeled data pool, however, semi-supervised learning and unsupervised learning still cannot achieve the upper-bound per-

Methods	Task	1st cycle	2nd cycle
LL4AL [44]	classification	+6.5%	+10.1%
	detection	+2.6%	+2.7%
VAAL [36]	classification	+6.2%	+6.0%
	detection	+2.3%	+0.0%
CALD (Ours)	detection	+8.4%	+5.8%

Table 1. Performance of two classification-based active learning methods (LL4AL [44] and VAAL [36]) on classification and detection compared with CALD (Ours) on detection. The percentage refers to the improvement compared with random selection. **The results show that the improvement drops significantly when classification-based active learning methods are transferred from classification to detection**.

formance of fully-supervised learning [30]. Active learning [1, 12, 13, 20, 24, 33, 35, 36, 43, 44, 46] provides a new perspective for reducing the demand of labeled data by selecting the most informative data (i.e. task models can gain higher performance by training on these samples) to be annotated. It has been proved successful on basic vision tasks, e.g. image classification.

As a fundamental and challenging task in computer vision, object detection [4, 10, 25, 45, 47] also suffers from intensive labor and time for data annotation, as it requires both bounding box and class label. Previous works generally follow the spirit of semi-supervised [19, 37, 38] and unsupervised learning [6, 15, 40] to better leverage the unlabeled data, *while not enough effort has been made to improve the efficiency of the annotation process*.

Although active learning methods [13, 32, 36, 44, 46] are popular for image classification, **directly applying classification-based active learning methods to object detection does not lead to satisfactory improvement** (see Table 1), due to **three challenges** of this problem. 1) Classification-based methods only consider the predicted class distribution, while the bounding box prediction could be equally important for selecting informative samples in object detection. 2) Informative objects often exist in lo-
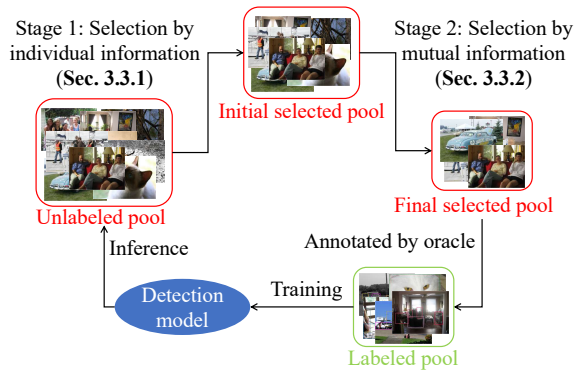
Figure 1. A high-level overview of CALD.

cal regions of images along with other uninformative objects. Simply applying a **global metric**, e.g. the loss of the model, may ignore some informative objects in an image where most objects are uninformative. 3) There is only one class assigned to each sample for classification, while for object detection, one image could have multiple objects with different classes which may have an unbalanced distribution. As a result, classification-based selection may lead to a highly unbalanced class distribution.

In this paper, we propose a Consistency-based Active Learning framework for object Detection (CALD). As shown in Fig. 1, in the first stage, we apply data augmentations to each unlabeled image and feed the original and augmented images to the initial detector (trained on randomly selected images before the process of active learning). We then calculate individual information, which unifies box regression and classification and focuses on local regions, based on the consistency between predictions of the original and augmented images to select informative samples to the initial selected pool. In the second stage, we further filter samples in the initial selected pool via mutual information (i.e. distance of class distributions of the selected pool and labeled pool) to **alleviate unbalanced class distribution**, leading to the final selected pool that meets the annotation budget. The main contributions are threefold:

- We discover the gap between active learning for image classification and object detection, which leads to the performances drop when existing classification-based active learning methods are transferred to detection. Our analysis inspires three principled guidelines of how to design an effective detection-specific active learning approach.
- We propose a novel detection-specific active learning method, CALD, considering the consistency of both bounding box and predicted class distribution when augmentation is applied to overcome the challenges brought by inconsistencies between classification and detection.
- Extensive experiments on PASCAL VOC [11], MS COCO [28] with Faster R-CNN [34] and RetinaNet [27] show that CALD outperforms state-of-the-art task-

agnostic and detection-specific active learning methods.

2. Related Work

Label-efficient object detection. Currently, the most popular and successful object detection frameworks are Convolutional Neural Network (CNN)-based one-stage or two-stage detectors such as Faster R-CNN (FRCNN) [34] and RetinaNet [27]. Compared with image classification [16, 23], object detectors need to implement both regression and classification tasks on local regions [18]. From the perspective of annotation, object detection requires not only classification but also bounding box. Various label-efficient methods are proposed to better leverage the information from unlabeled data. Most of them follow a paradigm of unsupervised or semi-supervised learning.

One popular class of semi-supervised learning methods of object detection [19, 29, 37, 38] are based on augmentation [2, 7, 17, 41, 42] and regularization. The key idea is to first generate pseudo labels for unlabeled images then use them to fine-tune the detector with augmentations. Another prevailing learning paradigm is unsupervised learning [5, 6, 15, 40] which aims to learn a better representation with unlabeled data. Then the model can be deployed as the backbone for downstream tasks such as object detection. *All the mentioned methods focus on taking advantage of unlabeled data, while the annotation procedure for supervised training is ignored.*

Classification-based active learning. A flurry of active learning methods [1, 8, 12, 13, 22, 31–33, 36, 44, 46] have been proposed for image classification. The most popular methods are based on pool-based selective sampling [1, 12, 35, 36, 44]. Pool-based methods continuously select the most informative samples from the unlabeled samples (i.e. unlabeled pool) as selected samples (selected pool) for labeling, and add them to the labeled samples (labeled pool) with a limit of budget. Learning Loss for Active Learning (LL4AL) [44] predicts target losses of unlabeled samples. Higher loss indicates the sample has higher uncertainty under the task model.

Another representative *task-agnostic* active learning method, called Variational Adversarial Active Learning (VAAL) [36], learns a latent space from a VAE and trains an adversarial network to discriminate samples between unlabeled and labeled data.

Detection-specific active learning. Unfortunately, there are limited works using pure active learning for object detection. Most related works [3, 9, 14] focus on classification, ignoring box regression or relying on the assistance of semi-supervised learning. The work closely following the standard active learning is [21], which introduces two methods: Localization Tightness with the classification information (LT/C) and Localization Stability with the classification information (LS+C). The former is based on the overlapping

ratio between the region proposals and the final prediction. Therefore, it can only be applied to two-stage detectors. The latter is based on the variation of predicted object locations when input images are corrupted by noise, which ignores the difference of classification.

A part of Self-supervised Sample Mining (SSM) [39] can be classified as active learning. SSM takes two steps to select samples: the first step is based on classification and the second step uses copy-paste strategy to cross validate the uncertainty of images. This method can easily lead to a distribution of samples with little diversity.

3. Method

3.1. Problem Overview

Given a fixed annotation budget, the active learning paradigm follows an iterative setting with C cycles and each cycle has $1/C$ of the total budget. Each cycle consists of **metric calculation** (which indicates how much information is in the samples – the core of active learning for selecting informative samples), data sampling and model training. In the i-th cycle, we have unlabeled images x_U in the unlabeled pool X_U^i, meanwhile there is also a labeled pool with ground truth (X_L^i, Y_L^i). Active learning aims to select the most informative samples from X_U^i to the final selected pool X_F^i, annotate them by oracle Ω, and add them to (X_L^i, Y_L^i). The process can be formulated as: $(X_L^{i+1}, Y_L^{i+1}) = (X_L^i, Y_L^i) \cup \Omega(X_F^i)$. Since our method has two stages (Fig. 1), we use X_I^i to denote the initial selected pool of the first stage. We will omit the superscript i since the operation is exactly the same in all the C cycles.

3.2. Guidelines of A Good Metric – Motivation

Before diving into the details of our proposed method, we first shed light on the following important question as the motivation of our approach. *The core of active learning lies in finding a good metric that selects the most informative samples, but what is a good metric for object detection?*

The information represented by a reasonable metric should follow two principles: 1) The information of samples (both labeled pool and selected pool) should generally decrease as learning cycle progresses, because the more knowledge the model learns, the less **new** information that samples can provide. 2) The information of the selected pool using the metric should be higher than that of the labeled pool, because the detectors have already learned most of the information in labeled pool and thus look for samples with more information to improve performance.

The reason why the metric of *predicted loss* cannot properly denote the information in samples when LL4AL [44] is transferred from classification to detection lies in the inconsistency of the two tasks. There are at least **three inconsistencies** between them. First, detectors perform both box

regression and classification during training and testing, so a single loss cannot represent the two types of uncertainty at the same time. Secondly, detectors focus more on local regions. For example, if LL4AL gives low uncertainty for an image, which means most part of the image is uninformative and some informative patches with small areas may be ignored. Thirdly, since images for the detection task usually contain *multiple objects*, the selected samples may contain some uninformative objects which are more likely to cause unbalanced class distribution than classification.

LS+C [21] partly addresses these issues by computing the average stability of all bounding boxes, while ignoring the classification confidence. It then selects the prediction with the lowest confidence as the most informative patch. The sum of these two metrics is considered as the final metric. However, it still fails to find the most informative combination of box regression and classification.

In contrast to existing methods, our framework follows three guidelines to tackle the inconsistencies: 1) Bounding box and classification predictions are considered together in one metric. 2) Our metric is computed based on local regions instead of the average information of the global image. 3) We apply an extra step to **alleviate the unbalanced class distribution of selected samples**. In the next section, we elaborate the two stages of the proposed CALD.

3.3. Consistency-based Active Learning for Object Detection (CALD)

In each active learning cycle, we have a detector Θ with parameters θ and a series of data augmentations $\mathcal{A}$ (e.g. flipping). For an unlabeled image x_U, its augmented versions are $x_U' = \mathcal{A}(x_U)$. The predictions (i.e. bounding boxes and classification scores) of x_U can be represented as $\Theta(x_U; \theta)$, and they can be mapped onto the augmented images x_U'. Therefore, we denote the transformation of the k-th prediction of x_U as reference prediction including **reference box** b_k, **reference score** s_k, which can be formulated as:

$$\{b_k\}, \{s_k\} = \mathcal{A}(\Theta(x_U; \theta)) \qquad (1)$$

Note that class-wise $s_k = [\varphi_1, \varphi_2, \cdots, \varphi_n, \cdots]^{\mathrm{T}}$, where φ_n denotes the confidence of the n-th class. The way $\mathcal{A}$ transforming predictions is similar to augmentation of images. Take horizontal flipping for instance, we get the box prediction of the augmented image by horizontally flipping the corresponding box from the original image and inheriting the classification prediction.

The predictions of augmented images obtained directly by the detector Θ can be expressed as:

$$\{b_j'\}, \{s_j'\} = \Theta(x_u'; \theta) \qquad (2)$$

$\{b_j'\}, \{s_j'\}$ are the sets of boxes and class-wise scores of predictions on x_U'. j denotes the j-th prediction.
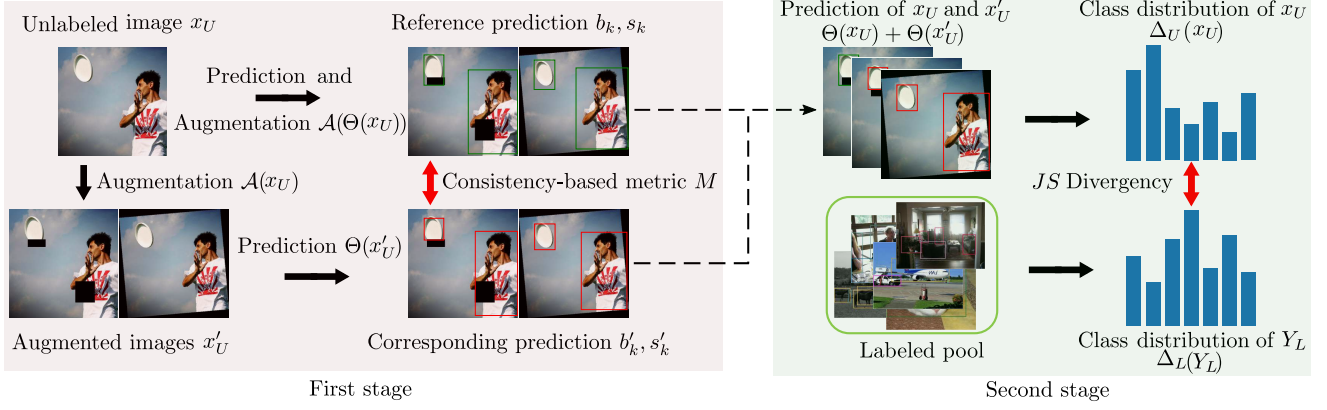
Figure 2. The two stages of the proposed CALD. The first stage selects samples based on individual information while the second stage uses mutual information to further refine the selected samples. Individual information is assessed by the consistency-based metric of reference and corresponding predictions. Mutual information refers to the JS divergence of class distributions of an unlabeled image and the labeled pool.

As shown in Fig. 2, in the first stage we extract individual information from images by consistency-based metric M between reference and matching predictions. According to the rank of M, we form an initial selected pool which is slightly over the budget. In the second stage, we evaluate the mutual information between samples in the initial pool and labeled pool and decide the final selected pool to meet the annotation budget.

3.3.1 Consistency-based Individual Information (Stage 1)

In order to calculate the consistency of predictions, first we need to match a corresponding prediction (including the **corresponding box** b'_k and **corresponding score** s'_k in $\{b'_j\}, \{s'_j\}$) to each reference prediction b_k. We choose the corresponding box b'_k that has the maximum Intersection over Union (IoU) with b_k. s'_k is the score of b'_k. The matching process $b_k \leftrightarrow b'_k$ can be formulated as:

$$b'_k = \underset{b'_j \in \{b'_j\}}{\operatorname{argmax}} IoU(b'_j, b_k). \quad (3)$$

The next step is to compute the consistency between reference prediction and corresponding prediction. For box regression, we directly use IoU which can clearly indicate the matching degree of two boxes. To measure the distance between two class-wise probabilities, Jensen-Shannon (JS) divergence and Kullback-Leibler (KL) divergence are popular metrics. We specifically take advantage of JS since it has clear upper and lower bounds, allowing us to quantify it in conjunction with IoU. Besides divergence of possibilities, we also adopt the maximum confidence as a weight factor to emphasize the prediction with high confidence. This is because a high-confidence prediction has a greater impact on performance (in case multiple predictions correspond to
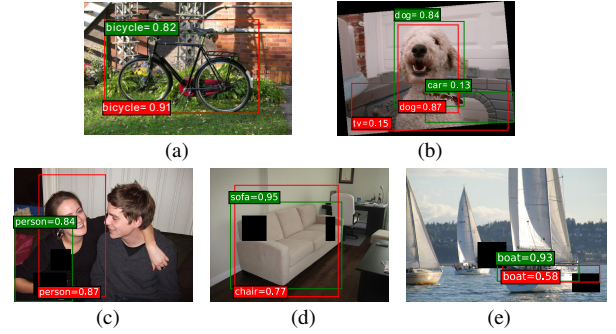

(a)　(b)

(c)　(d)　(e)

Figure 3. Case analysis. Red boxes are reference predictions and green boxes are their corresponding predictions. (a)(b) are uninformative cases while (c)(d)(e) are informative cases.

the same ground truth, only the prediction with the highest confidence is regarded as true positive, while others will be regraded as false positives). Finally, we reverse JS to $1 - JS$ to keep the same trend as IoU. The consistency of the k-th prediction of an image can be computed as the sum of consistencies of boxes C^b_k and scores C^s_k:

$$m_k = C^b_k + C^s_k \quad (4)$$

where

$$C^b_k = IoU(b_k, b'_k)$$
$$C^s_k = \underbrace{\frac{1}{2}[\max_{\varphi_n \in s_k}(\varphi_n) + \max_{\varphi'_n \in s'_k}(\varphi'_n)]}_{\text{weight factor}}(1 - JS(s_k||s'_k)) \quad (5)$$

Obviously, $m_k \in [0, 2]$. Ideally, a smaller consistency m_k indicates a more unstable prediction, which also implies more informative the patch is. *However, directly define m_k as the metric of information does not work well in practice.*

The most important question is when m_k is on the lower bound, is the prediction the most informative one? The answer is not necessarily. For example, as shown in Fig. 3b, for the paired predictions *tv* (i.e. reference prediction) and *car* (i.e. corresponding prediction), C_k^b and C_k^s are both small and m_k is close to its lower bound. However, due to the bad matching and low confidence, this prediction is very unlikely to be the main result. The reason is that the detector may give another accurate prediction of the *dog*, which is also shown in Fig. 3b. In this case, we can observe that the predictions for the main object *dog* are actually very stable and accurate, manifesting an uninformative sample to the detector. But if simply using m_k (the lower the better), this sample is falsely considered as an informative one. In other words, the smallest m_k does not necessarily represent the most informative patch in practice due to the instability or randomness of prediction.

On the other hand, when m_k is close to the upper bound (e.g. Fig. 3a), the detector can handle the augmentation well and give a high-confidence prediction which is likely to be correct due to the high matching degree. Such samples are not informative because the detector can deal with the augmentations well. Based on these observations and analyses, we speculate that m_k of an informative prediction should have two properties: (1) Keeping a certain distance from the lower bound, which means the paired predictions have relative high matching degree and high confidence. If the prediction is wrong, this patch is probably informative, because the prediction is likely the main result of the object (the detector will not give other accurate predictions of the object like the case in Fig. 3b). And this is based on the fact that there cannot be multiple predictions with high confidence in the same area at the same time, according to Softmax and non-maximum suppression (NMS). (2) Being far away from the upper bound, which means the matching degree is worse than when m_k is on the upper bound (such as Fig. 3a). This indicates that the detector cannot cope with common augmentations on the image, and this prediction is likely to be inaccurate (such as Figs. 3c, 3d and 3e).

To quantify this, the consistency-based metric of an image is defined as

$$M(x_u; \mathcal{A}, \Theta) = \mathbb{E}_{\mathcal{A}}[\min_k |m_k - \beta|], \qquad (6)$$

where β is the base point to represent m_k of the most informative patch. Based on the above analysis, we search the optimal β heuristically: starting from the midpoint of the upper and lower bounds of m_k, the optimal β can be found with several grid searching steps (we only use 5 steps). The optimal value of β searched by this procedure is effective for all datasets and detectors. The reason we adopt minimum value over an image instead of mean value is that we focus on finding the most informative local regions instead of the whole image. Finally, we compute the expectation of

M over multiple augmentations to improve reliability.
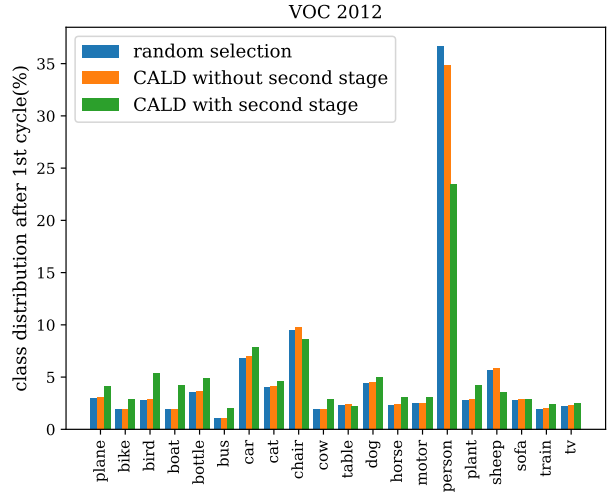

Figure 4. Class distributions of the labeled pool (after the first active learning cycle) with different selection strategies.

3.3.2 Mutual Information (Stage 2)

We find that the class distribution of the labeled pool is unbalanced after random selection. As shown in Fig. 4, the height differences between the blue bars are very large. And this issue cannot be solved by only using individual information (orange bars) for sample selection since there are multiple objects in detection images.

We propose an inter-image metric to alleviate the issue. The idea is to compare the class distribution of each sample in the initial selected pool and that of the whole labeled pool, then select samples with large distances (i.e. having different class distributions from the labeled pool) to form the final selected pool. We use JS divergence to evaluate the distance between two class distributions (i.e. mutual information). The procedures of computing the mutual information are outlined in Algorithm 1. For the labeled pool, we sum all ground truth (Y_L) to represent the class distribution which is computed as

$$\Delta_L(Y_L) = Softmax([\delta_1, \delta_2, \cdots, \delta_m, \cdots]^{\mathrm{T}}),$$
$$\delta_m = \sum_{y_L \in Y_L} \mathcal{I}(y_L = m). \qquad (7)$$

m denotes the m-th category in the dataset and $\mathcal{I}$ is the indicator function. For an unlabeled image x_U in X_I, we only count the highest confidence of predictions in each class due to high certainty. If we follow the notations in Sec. 3.1, denoting the class-wise classification prediction of original and augmented image by s_k and s'_j in which φ_m (φ'_m) is

Algorithm 1 Selection by mutual information in each cycle

Input: Initial selected pool X_I, ground truth of labeled pool Y_L, total budget B, budget per cycle B/C

Functions: Distribution function of labeled pool $\Delta_L(Y_L)$ and single unlabeled image $\Delta_U(x_U)$

    $X_F \leftarrow \{\}$

    while size$(X_F) < B/C$ **do**

        $f = \text{argmax}_{x_U \in X_I}[JS(\Delta_U(x_U)||\Delta_L(Y_L)]$

        $X_F = X_F \cup \{X_I[f]\}$

        $X_I = X_I - \{X_I[f])\}$

    end while

    return X_F

---

the score of m-th class, the process can be formulated as

$$\Delta_U(x_U) = Softmax([\delta_1, \delta_2, \cdots, \delta_m, \cdots]^{\mathrm{T}})$$
$$\delta_m = \max_{s_k \in \{s_k\}} \{\varphi_m | \varphi_m \in s_k\} + \max_{s'_j \in \{s'_j\}} \{\varphi'_m | \varphi'_m \in s'_j\} \quad (8)$$

Return to Fig. 4, we can observe that after selecting by mutual information (green bars), in general, the proportions of the majority categories (such as *person*) have dropped while the proportions of minority categories (such as *bus* and *bike*) have risen, alleviating the unbalanced distribution.

4. Experiments

Datasets. To validate the effectiveness of CALD, we conduct extensive experiments on three benchmarks for object detection: MS COCO [28], Pascal VOC 2007 and VOC 2012 [11]. On VOC 2012 and COCO, we use training set for training and validation set for testing. On VOC 2007, we use trainval set for training and test set for testing. On VOC, we set 500 labeled images as random initialization and 500 as budget per cycle. For RetinaNet, we set 1000 as initialization since 500 images are too few to train a robust model for RetinaNet. On COCO, we set 5000 as initialization and 1000 as budget per cycle by following [21]. Detectors are evaluated with mean Average Precision (mAP) at IoU $= 0.5$ on VOC and with average mAP from IoU $= 0.5$ to IoU $= 0.95$ on COCO, which are standard protocols for these datasets.

Detectors. We employ the popular two-stage detector Faster R-CNN (FRCNN) [34] and single-stage detector RetinaNet [27], both with Resnet50 [16] and FPN [26], as task models. The implementation of the two models follows the default settings of Torchvision. In each cycle we train the models for 20 epochs. *The numbers reported in the results are averages of 3 trials for each method and detector.*

4.1. Comparison with State of the Art

We compare the proposed CALD with random selection (Random), three detection-specific active learning methods

(SSM [39], LS+C and LT/C [21]) and two task-agnostic active learning methods (VAAL [36] and LL4AL [44]), which represent the state-of-the-art (SOTA). As shown in Fig. 5, CALD outperforms the SOTA methods on all three datasets with both FRCNN and RetinaNet detectors.

On VOC 2012 and VOC 2007 with FRCNN, the improvements of CALD over random selection and the second-best method are significant. Specifically, in terms of mAP, CALD is 8.4% and 7.0% higher than random selection, and 5.7% and 3.8% higher than the second-best method LT/C in the first cycle on VOC 2012 and 2007, respectively. This demonstrates the effectiveness of CALD by following the three guidelines: unifying the metric of box regression and classification, focusing on local regions and promoting a balanced class distribution. Also the improvements manifest a consistent trend: in the first cycle, the improvements are the largest and gradually decrease in subsequent cycles (3.1 to 1.2 and 2.3 to 1.4). The reason is that as the number of available unlabeled samples gradually decreases, samples collected by all methods tend to be the same. We observe that the task-agnostic active learning methods LL4AL and VAAL perform bad (5.4% and 5.7% lower than CALD in the first cycle on VOC 2012) on two-stage detector. FRCNN first extracts region proposals and then adopts fine-grained predictions on local patches. Such complexity widens the gap between detection and classification, leading to worse performance of classification-based methods. LT/C, which is specifically designed for two-stage detectors, performs second only to CALD (2.5% lower than CALD on average on VOC 2012). However, its shortcomings are also obvious. First, it cannot be widely used in other detectors (such as one-stage detectors) while CALD can be generalized to any detector. Secondly, it cannot process the classification information finely, since FRCNN dose not give class-wise scores in the first stage. Although LS+C has considered box regression, it does not combine boxes and classification to get a comprehensive metric, so it does not perform well in practice.

As for RetinaNet, the improvement of CALD is also the most significant: 11.8% higher than random selection on average on VOC 2012. Compared with FRCNN, the performance of LL4AL using RetinaNet is slightly closer to CALD (2.5% lower than CLAD on average on VOC 2012). The reason is that RetinaNet has a simpler architecture which directly gets predictions from global features, therefore the global information has a greater impact on the final results than that of FRCNN. However, classification-based methods still cannot take box regression into account.

We also note that CALD yields more improvements in difficult categories. For categories with AP lower than 40% in random selection, we treat them as difficult categories. For the difficult categories (red bars) in Fig. 6, we notice that the improvements are larger than other classes.
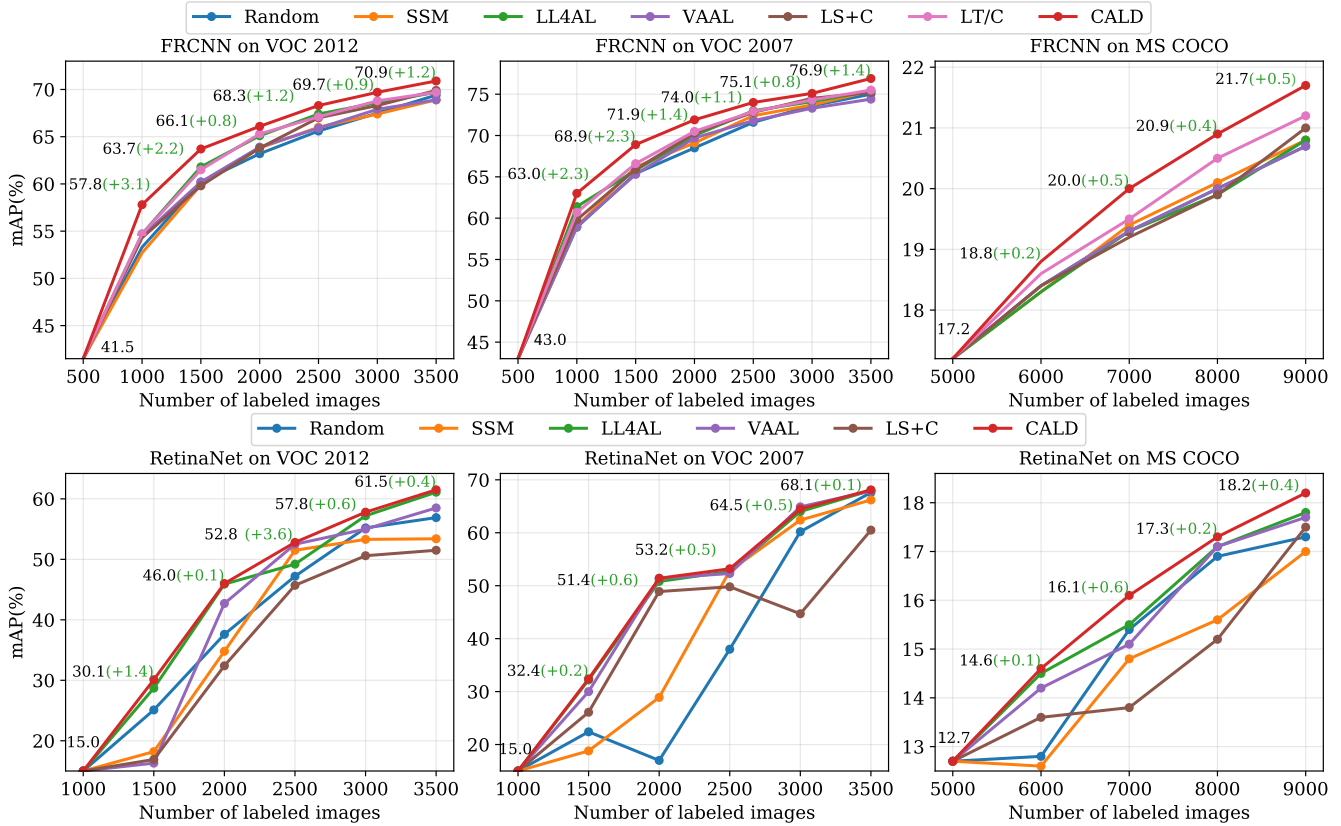
Figure 5. Comparison with SOTA task-agnostic and detection-specific active learning methods (LT/C only applies to two-stage detectors). CALD surpasses all methods comprehensively on three datasets and two detectors. The numbers marked on the points of CALD denote performance and its improvement over the second-best method. In the first row, the second-best methods are all LT/C [21] while the second-best methods are all LL4AL [44] in the second row.
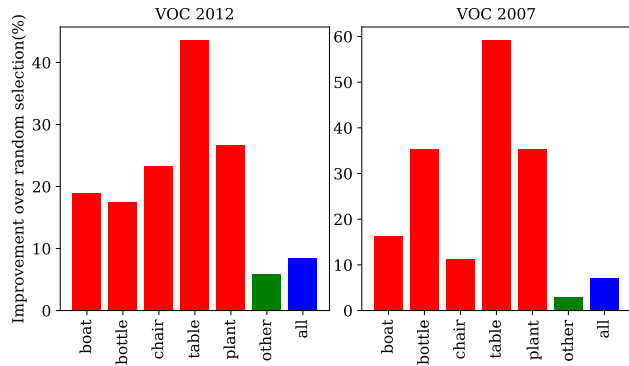

Figure 6. The improvement in difficult classes (red bars) over random selection in the first cycle on VOC. Green and blue bars are improvements for non-difficult classes (others) and all classes.

4.2. Ablation Study

We conduct ablation studies on VOC 2012 with FRCNN.

Consistency-based metric. To validate the consistency-based metric M is reasonable, we conduct ablation study on

different strategies. Firstly, in Eq. 6, we use the minimum $|m_k - \beta|$ of an image. For ablation study, we investigate the performance of using mean $|m_k - \beta|$ for an image, i.e.

$$M(x_u; \mathcal{A}, \Theta) = \mathbb{E}_{\mathcal{A}}[\mathbb{E}_k |m_k - \beta|], \qquad (9)$$

which represents the average global information of the image. As shown in Fig. 7a, the performance (curve of "mean") drops significantly because detectors focus more on local regions rather than global information. Secondly, we remove the base point β, i.e. let β equal to 0. The performance drops since the smallest m_k does not necessarily represent the most informative patch as we discussed in **Sec. 3.3.1** with a case study of Fig. 3b. Finally, if we remove the weight factor in m_k (see Eq. 5), which means we treat all predictions with different confidences equally, the performance also drops slightly. Because prediction with high confidence has a greater impact on performance.

Data augmentations. We compute the detection consistency based on common data augmentations in our method. For simplicity, we use a single uppercase letter to denote one type of augmentation. "F" for horizontal flip, "C" for
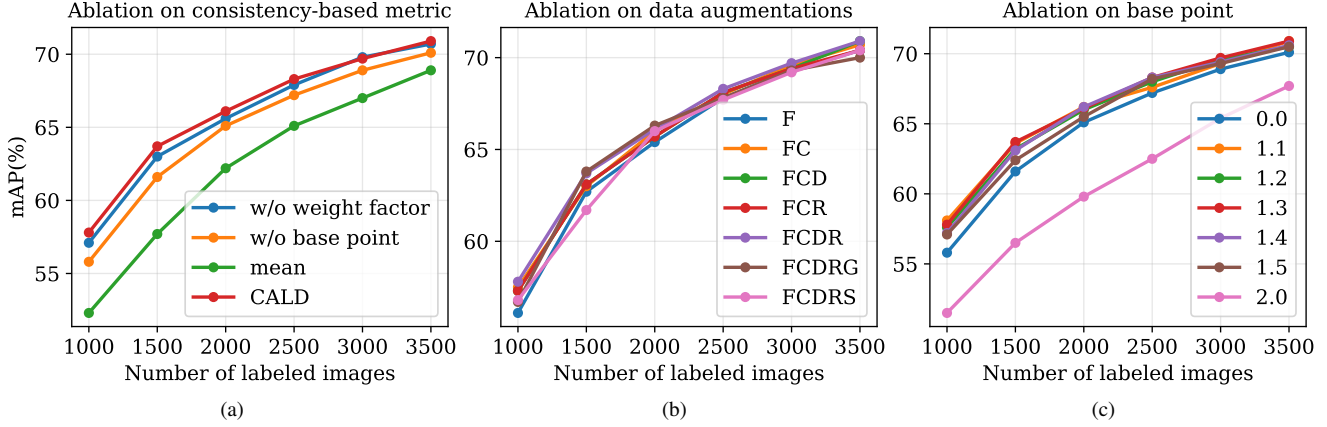
Figure 7. Ablation studies on strategies of consistency-based metric, data augmentations, and base point β.

Expansion ratio	mAP of 1st cycle	mAP of 2nd cycle
0%	56.9	62.8
10%	57.4	63.3
20%	**57.8**	**63.7**
30%	57.3	63.5
40%	57.1	63.7

Table 2. Ablation on the expansion ratio for X_I.

cutout, "D" for downsize, "R" for rotation, "G" for Gaussian noise, and "S" for salt and pepper noise. The combination of letters means we get M by averaging the results of these augmentations. As shown in Fig. 7b, CALD works well with standard augmentations, and does not rely on specific augmentations. The proper combination of augmentations can make the performance more stable. We adopt "FCDR" in CALD.

Base point β. Base point is the parameter β of consistency-based metric in Eq. 6, which denotes the value of m_k of the most informative prediction. From the plots in Fig. 7c, β gets the optimal value around 1.3. Thus $\beta = 1.3$ is used for all datasets and detectors in our experiments. When the value of β goes from the optimal point to the lower bound (0.0) of m_k, the performance of CALD decreases slowly. It can be explained that when m_k is closer to the lower bound, the predictions become unstable which are not necessarily informative. On the contrary, if β is excessively closer to the upper bound, the performance drop quickly. This is because m_k close to the upper bound denotes uninformative predictions (cases like Fig. 3a). When m_k reaches the upper bound (2.0), CALD selects the least informative samples (performance of detector is even worse than random selection), which also indicates that CALD can clearly distinguish whether the sample is informative or not.

Expansion ratio for X_I. As stated in Sec. 3.3, we form the initial selected pool X_I in the first stage by selecting more samples than the budget, so that we can further filter those samples in the second stage to meet the budget for each cy-

cle. Then one question emerges: *how many more samples to use?* Assume the budget of each cycle is 500 images, selecting 600 images for X_I in the first stage means a 20% expansion ratio. We investigate different expansion ratios and report the results in Table 2. Note that 0% in this table indicates our method reduces to **only have the first stage**. We reach two conclusions. (1) Based on the results of 0%, 10% and 20%, there is a clear advantage of leveraging mutual information for sample selection in the second stage. (2) 20% additional budget for X_I yields the best performance, leading to an mAP improvement of 0.9 in both cycles (56.9 *vs.* 57.8; 62.8 *vs.* 63.7). However, keep expanding the budget in the first stage would also cause performance drop (e.g. 30% ratio). This is because more informative samples may be removed by mutual information in the second stage in order to cut back to the fixed budget. Therefore, the experimental results reveal the importance of both individual and mutual information for sample selection.

5. Conclusion

This paper introduces a consistency-based active learning method for object detection, namely CALD. In order to select the most informative samples, it leverages a consistency-based metric to consider the information of box regression and classification simultaneously, which is ignored by previous methods. In addition to sample individual information, CALD also uses mutual information to refine sample selection to encourage a balanced class distribution. Extensive experiments show that CALD with different detectors achieves state-of-the-art performance on several object detection benchmarks under active learning settings.

References

[1] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference*

on *Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 1, 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 1, 2

[3] Alberto Bietti. Active learning for object detection on satellite images. *Technical report, Technical report, Caltech*, 2012. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 2

[8] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995. 2

[9] Sai Vikas Desai, Akshay L Chandra, Wei Guo, Seishi Ninomiya, and Vineeth N Balasubramanian. An adaptive supervision framework for active learning in object detection, 2019. 2

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 1

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 1, 2

[13] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020. 1, 2

[14] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1430–1435, 2020. 2

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 6

[17] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 2

[18] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 2

[19] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 10759–10768. Curran Associates, Inc., 2019. 1, 2

[20] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009. 1

[21] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision*, pages 506–522. Springer, 2018. 2, 3, 6, 7

[22] Vikram Krishnamurthy. Algorithms for optimal scheduling and management of hidden markov model sensors. *IEEE Transactions on Signal Processing*, 50(6):1382–1397, 2002. 2

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2

[24] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994. 1

[25] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 190–191, 2020. 1

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 6

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[29] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection, 2021. 2

[30] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1

[31] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018. 2

[32] Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3071–3079, 2020. 1, 2

[33] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 1, 2

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2, 6

[35] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 1, 2

[36] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019. 1, 2, 6

[37] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 2

[38] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2291–2301, 2021. 1, 2

[39] Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, and Liang Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1613, 2018. 3, 6

[40] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. *arXiv preprint arXiv:2102.04803*, 2021. 1, 2

[41] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 1, 2

[42] Taojiannan Yang, Sijie Zhu, and Chen Chen. Gradaug: A new regularization method for deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[43] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015. 1

[44] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019. 1, 2, 3, 6, 7

[45] Weiping Yu, Taojiannan Yang, and Chen Chen. Towards resolving the challenge of long-tail distribution in uav images for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3258–3267, 2021. 1

[46] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8756–8765, 2020. 1, 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1