This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Plug and Play Active Learning for Object Detection

Chenhongyi Yang[1*] Lichao Huang[2] Elliot J. Crowley[1]
[1]School of Engineering, University of Edinburgh
[2]Horizon Robotics

Abstract

Annotating datasets for object detection is an expensive and time-consuming endeavor. To minimize this burden, active learning (AL) techniques are employed to select the most informative samples for annotation within a constrained "annotation budget". Traditional AL strategies typically rely on model uncertainty or sample diversity for query sampling, while more advanced methods have focused on developing AL-specific object detector architectures to enhance performance. However, these specialized approaches are not readily adaptable to different object detectors due to the significant engineering effort required for integration. To overcome this challenge, we introduce Plug and Play Active Learning (PPAL), a simple and effective AL strategy for object detection. PPAL is a two-stage method comprising uncertainty-based and diversity-based sampling phases. In the first stage, our Difficulty Calibrated Uncertainty Sampling leverage a category-wise difficulty coefficient that combines both classification and localisation difficulties to re-weight instance uncertainties, from which we sample a candidate pool for the subsequent diversity-based sampling. In the second stage, we propose Category Conditioned Matching Similarity to better compute the similarities of multi-instance images as ensembles of their instance similarities, which is used by the k-Means++ algorithm to sample the final AL queries. PPAL makes no change to model architectures or detector training pipelines; hence it can be easily generalized to different object detectors. We benchmark PPAL on the MS-COCO and Pascal VOC datasets using different detector architectures and show that our method outperforms prior work by a large margin. Code is available at https://github.com/ChenhongyiYang/PPAL

1. Introduction

Object detectors typically need a huge amount of training data [21, 26] annotated with both object category labels and bounding box locations. This annotation is expensive and tedious. If we are required to do some annotation, it would make sense to annotate the images that will be of the greatest benefit when used for training. But how do we know which ones to choose? The goal of active learning is to tell us. Given a large unlabelled data pool, active learning (AL) aims to sample data that would maximally improve a model's performance if that data was annotated and used for training. There are typically two main streams of active learning: (1) Uncertainty-based AL methods [20, 22, 23, 36, 40] select samples that maximise a measure of model uncertainty e.g. those with the least mutual information with the current set of labelled data; (2) Diversity-based AL approaches [1, 3, 13, 16, 29, 31, 39, 42, 47] instead select samples that are representative of the whole distribution of unlabelled data; this can be achieved by minimising the similarities between the features [39] or posterior probabilities vectors [1] inside this subset.

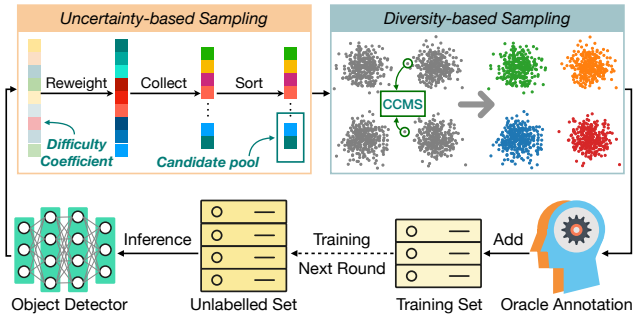


Figure 1. An overview of our two-stage PPAL. In the first Difficulty Calibrated Uncertainty Sampling stage, the objects' uncertainties are re-weighted with the difficulty coefficients that take both classification and localisation into account, and a *candidate pool* of images, which the model is mostly uncertain on, are sampled. In the second diversity-based stage, we run a modified kmeans++ algorithm using the proposed Category Conditioned Matching Similarity (CCMS) to select a set of representative images as active learning queries for the next round of annotation.

With a neural network, uncertainty-based and diversity-based AL can be straightforwardly applied to the image classification task. For example, uncertainty-based sampling can be implemented by selecting the unlabelled sam-

*Corresponding Author. Email: chenhongyi.yang@ed.ac.uk

ples with the maximum classification entropy, and the diversity-based can be achieved by minimising the selected samples' mutual similarities computed using their averaged feature maps. However, designing an effective AL strategy for object detection is more challenging. This is because detection consists of both object localisation and classification. It is difficult to quantify *uncertainty* jointly across both tasks e.g. we may have an object that is easy to localise, but hard to classify. Also, it is hard to measure *similarity* when images contain multiple objects with different features. Several works have tackled AL for detection [4, 10, 50] but rely on modifying the architecture of an object detector as well as the training pipeline. This means they cannot be easily integrated into other object detector frameworks without significant engineering effort.

In this work we propose Plug and Play Active Learning (PPAL), a plug-and-play AL algorithm for object detection. It is state-of-the-art and easy to use; it requires no modifications to architectures or training pipelines and works across a wide range of object detectors. PPAL is a two-stage algorithm that combines uncertainty- and diversity-based sampling: in the first stage it selects a *candidate pool* of images with high uncertainty scores from a large unlabelled dataset, then in the second stage it select a highly diverse AL query set for annotation. In more detail, we propose Difficulty Calibrated Uncertainty Sampling (DCUS) for the first stage, in which the category-wise difficulty coefficients are computed and updated during training and then used to re-weight the instance uncertainties when selecting samples. The difficulty coefficients take both classification and localisation difficulties into account, so the two sub-tasks of object detection are balanced in the uncertainty-based sampling stage. DCUS also allows the uncertainty sampling to favour objects in challenging categories, hence benefiting the overall average precision (AP). For the second stage, we propose Category Conditioned Matching Similarity (CCMS): a new method for measuring similarities for multi-instance images, in which every object is matched to its most similar counterpart in the other image to compute instance-wise similarity, and the image-wise similarity is computed by ensembling the instance-wise similarities. Then, CCMS is used off-the-shelf by a modified kmeans++ algorithm to select a diverse and representative subset as the active learning queries. Notably, unlike the recently proposed [4, 10, 50], PPAL does not modify the model architecture or training pipeline of object detectors, thus it is highly generalisable to different types of detectors.

To summaries, our contributions are: **(1)** We propose PPAL, a two-stage active learning algorithm for object detection that combines uncertainty-based and diversity-based sampling. It is plug-and-play, requiring no architectural modifications or any change to training pipelines. **(2)** We show that PPAL outperforms previous object detection AL algorithms across multiple object detectors on the COCO and Pascal VOC datasets. We also show that our method can be easily generalised to different object detectors.

2. Related Work

Active Learning. Active learning approaches can be broadly separated into uncertainty-based and diversity-based methods. Uncertainty-based methods [20, 22, 23, 36, 40] aim to select samples for annotation that maximise some uncertainty measure; common measures include entropy [20, 40] and the margin between largest two predicted class posterior probabilities [11, 20, 36]. In [15] the model uncertainty is estimated by performing multiple forward passes with Monte Carlo Dropout. Learn Loss [48] uses a prediction of the loss as a measure of uncertainty. Diversity-based methods [1, 3, 13, 16, 29, 31, 39, 42, 44, 47] for AL aim to select representative samples so that a small subset of data can describe the whole dataset. Core-set [39] uses a greedy k-centroid algorithm to select a small core set from the unlabelled pool and use Mixed Integer Programming to iteratively improve sample diversity. CDAL [1] utilises predicted probabilities to improve context diversity. Recently, several works have been proposed that combine uncertainty-based and diversity-based AL. In [32, 53], the uncertainty and diversity are balanced by running the k-means algorithm on image features weighted by model uncertainty. In BADGE [2] the uncertainty and diversity are balanced by a k-means++ algorithm seeding on the gradients of the model's last layer. HAC [11] first clusters the unlabelled samples and queries the most uncertain samples in each cluster in a round-robin way.

Object Detection. ConvNet-based object detectors usually follow a two-stage or single-stage design. Two-stage detectors [5, 7, 17, 34] use a region proposal network [34] to extract plausible objects and the RoIAlign operation [17] to extract regional features for further classification and localisation; Single-stage [9, 19, 25, 27, 33, 35, 41, 52] detectors directly predict objects' bounding box and category label on every position of the image feature maps. Recently, a wave of transformer-based detectors [6, 24, 51, 54] have been proposed. They use self-attention to exchange information between query vectors and image features and directly output detected objects without Non-maximum-suppression.

Active Learning for Object Detection. Early attempts at applying AL to object detection [1, 30, 39, 48, 49] involved the direct application of image classification AL algorithms. However, these do not account for the joint classification and localisation tasks that make up detection, or that images can contain multiple different objects. This prompted the design of AL algorithms specifically for detection. MDN [10] modify an object detector to learn a Gaussian mixture model (GMM) for both the classification and regression outputs, and then the uncertainties of both

tasks are derived from the modelled GMMs. In [12], semi-supervised learning is used to deal with false positives and compute model uncertainty. MIAL [50] uses adversarial training to compute model discrepancy, which is used for computing uncertainty. DivProto [43] improves the AL algorithm by replacing the detection scores with model uncertainties in NMS; it also uses a set of diverse prototypes to select the most representative samples. However, a problem of previous works in AL for object detection is that the model architecture or training pipelines are modified to suit the AL purpose, which limits their generalisation ability across different architectures.

3. Method

In this section, we first define the problem of active learning for object detection (Sec. 3.1). Then, the two key innovations of our method are described in detail in Sec. 3.2 and 3.3. Fig. 1 gives a high-level overview of our algorithm.

3.1. Problem Statement

Following [10, 43, 50], we define the problem under the batch active learning setting, i.e. at each round we query a batch of images for oracle annotation instead of a single sample. Suppose there is a training set $X_T = \{x_i\}_{i\in[N_t]}$ with size N_t and a validation set $X_V = \{x_i\}_{i\in[N_v]}$ with size N_v. At round $r \geq 0$, the labelled set is X_L^r and the unlabelled pool is X_U^r where $X_L^r \cap X_U^r = \emptyset$ and $X_L^r \cup X_U^r = X_T$. An object detection model f_θ^r with parameters θ is trained on X_L^r and its performance on X_V is $Z(X_V|f_\theta^r)$, which is usually measured by mean Averaged Precision (mAP). Given budget b, an active learning algorithm selects a query set $X_Q^r \subseteq X_U^r$ with size b for oracle annotation. Then we can get the labelled set $X_L^{r+1} = X_L^i \cup X_Q^r$ and the unlabelled $X_U^{r+1} = X_U^i - X_Q^r$ for the next round $r+1$. Finally, the detection model is trained on X_L^{r+1} to get f_θ^{r+1} with performance $Z(X_V|f_\theta^{r+1})$. After k rounds of active learning, an active learning algorithm is evaluated by the model's improvement over the initial round $\triangle Z^{k,0} = Z(X_V|f_\theta^k) - Z(X_V|f_\theta^0)$, i.e., the AL algorithm that can bring most performance improvement is favoured.

3.2. Difficulty Calibrated Uncertainty Sampling

We propose Difficulty Calibrated Uncertainty Sampling (DCUS) to serve two purposes: (1) it provides a means to score uncertainty for object detection based on both classification and localization; (2) it allows more objects in challenging categories to be sampled. DCUS circumstance the above two challenges by re-weighting the object uncertainties with a category-dependent difficulty coefficient. Intuitively, we aim to raise the importance of categories that the model does not perform well, while down-weight the easy categories. Specifically, suppose there are C classes in the dataset, at the beginning of each AL training round, we define the class-wise difficulties $D^r = \{d_i\}_{i\in[C]}$ and initialise them with all 1. Inspired by [9], we compute the training difficulty of every predicted box during training as:

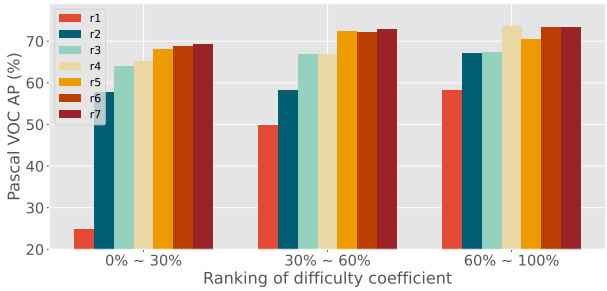


Figure 2. Illustration of how the category-wise difficulty coefficients correspond to the evaluated detection APs on Pascal VOC at each active learning round, in which the difficulty coefficients are sorted in descending order. Objects in categories with high-difficulty coefficients are harder to be detected than those in categories with low-difficulty coefficients.

$$q(b|\hat{b}) = 1 - P\big(b|\hat{b}\big)^{\xi} \cdot \mathrm{IoU}\big(b, \hat{b}\big)^{1-\xi} \tag{1}$$

where b is the predicted box, $\hat{b}$ is the ground-truth box that b is assigned to, $P\big(b|\hat{b}\big)$ is the classification probability w.r.t. the class of $\hat{b}$, $\mathrm{IoU}\big(b, \hat{b}\big)$ is the IoU between b and $\hat{b}$, and $0 \leq \xi \leq 1$ is a hyper-parameter. The detection difficulty defined in Eq. 1 takes both classification and localisation into account, hence both of them will contribute to the uncertainty sampling. Then, the recorded class-wise difficulties are updated by the averaged object difficulty using an exponential moving average (EMA) during training:

$$d_i^k \leftarrow m_i^{k-1} d_i^{k-1} + (1 - m_i^{k-1}) \frac{1}{N_i^k} \sum_{j=1}^{N_i^k} q_j \tag{2}$$

$$m_i^k \leftarrow \begin{cases} m^0 & \text{if } N_i^k > 0 \\ m^0 \cdot m_i^{k-1} & \text{if } N_i^k = 0 \end{cases} \tag{3}$$

where k is the training iteration, d_i^k is the updated difficulty for category i, N_i^k is the number of predicted objects of class i in the training batch, q_j is the j-th object's training difficulty, m_i^* is the EMA momentum, m^0 is the initial momentum for all categories. As shown in Eq. 3, if a training batch does not contain objects in category i, we decrease its class-wise momentum by multiplying m^0, which ensures a similar updating pace of difficulties of different classes. In Fig. 2, we show how the class-wise difficulties correspond to the class-wise detection APs during active learning.

After training the detector on the labelled dataset, we compute the category-wise difficulty coefficient $W^r = \{w_i\}_{i\in[C]}$ for the r-th AL round as:

$$\begin{aligned} w_i &= 1 + \alpha\beta \cdot \log\left(1 + \gamma \cdot d_i\right) \\ \text{where} \quad \gamma &= e^{1/\alpha} - 1 \end{aligned} \tag{4}$$

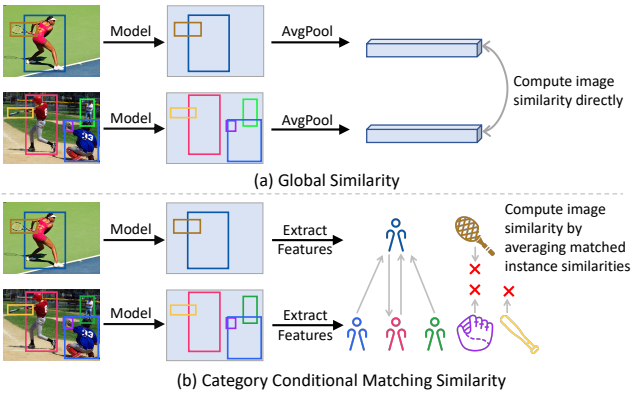


Figure 3. Comparison of global similarity and our CCMS. The global similarity is computed using the averaged image feature maps, failing to capture the fine-grained spatial information of multi-instance images. On the other hand, in CCMS, each object in an image finds its most similar counterpart with the same category in another image to compute similarities. Then image-wise similarity is computed by averaging the object similarities.

where α controls how fast the difficulty coefficient changes w.r.t. the class-wise difficulty, β controls the upper bound of the difficulty coefficient. Finally, we compute the image-wise uncertainty of every unlabelled image by summing the entropy of each detected object weighted by the corresponding difficulty coefficient:

$$U(I) = \sum_{i=1}^{M_I} w_{c(i)} \cdot \sum_{j=1}^{C'} -p_{ij} \cdot \log(p_{ij}) \tag{5}$$

where M_I is the number of detected objects from image I; $w_{c(i)}$ is the weight of object i's predicted category; C' is the number of classification ways, which is usually $C+1$ for two-stage detectors and 2 for one-stage detectors; p_{ij} is the predicted probability of category j. Finally, we use a simple strategy to select the candidate pool: for a given AL budget b, we sort the images by their uncertainties and select $\delta \cdot b$ most uncertain ones from the unlabelled set. We call the hyper-parameter $\delta > 1$ *budget expanding ratio*.

3.3. Diversity Sampling for Multi-instance Images

In the second stage of PPAL a diverse and representative set will be selected from the candidate pool to serve as the AL query set. In previous works [1, 39], diversity-based sampling is achieved by minimising the similarities of every pair of the selected samples. Such similarities are often computed by the cosine or L_2 similarities of the averaged convolutional feature maps [39], which we call *global similarity*. This practice is simple and works well for object-centric datasets like ImageNet [37]. However, object detection usually take multi-instance images as input, and in such cases the averaged feature maps are difficult to capture the fine-grained spatial information in those images [46]. Therefore, we design a new similarity computing method to compute similarities of multi-instance images, which can be used off-the-shelf for diversity-based sampling. We show the differences between those two similarities in Fig. 3.

Category Conditioned Matching Similarity. The intuition behind our CCMS is that the similarity of two multi-instance images can be computed by measuring how similar their contained objects are. Formally, for two multi-instance images I_a and I_b, the object detector detects several objects from them $O_a = \{o_{a,i}\}_{i\in[M_a]}$ and $O_b = \{o_{b,i}\}_{i\in[M_b]}$, in which each object $o_{*,i}$ is a triplet $o_{*,i} = (f_{*,i}, t_{*,i}, c_{*,i})$: $f_{*,i}$ is the object's visual features extracted from the feature maps; $t_{*,i}$ is the detection score; $c_{*,i}$ is the predicted class label. We use the similarity of O_a and O_b as a proxy for the similarity of I_a and I_b, which is computed by matching every object to its most similar counterpart in the same category in the other image. Specifically, for an object $o_{a,i}$ in O_a, its similarity to O_b is computed as:

$$s(o_{a,i}, O_b) = \begin{cases} \max\limits_{c_{b,j}=c_{a,i}} \frac{f_{a,i} \cdot f_{b,j}}{||f_{a,i}|| \cdot ||f_{b,j}||} + 1 \\ 0 \quad \text{if no } c_{b,j} = c_{a,i} \end{cases} \tag{6}$$

where we set $S(o_{a,i}, O_b)$ to 0 if no object in O_b is in the same category as $o_{a,i}$. Then the similarity of O_a to O_b is computed by averaging the object similarities weighted by their detection scores:

$$\begin{aligned} S'(O_a, O_b) &= \frac{1}{\sum_i t_{a,i}} \sum_{i=1}^{M_a} t_{a,i} \cdot s(o_{a,i}, O_b) \\ S(O_a, O_b) &= \frac{1}{2} \cdot \big(S'(O_a, O_b) + S'(O_b, O_a)\big) \end{aligned} \tag{7}$$

where the final similarity is the average of $S'(O_a, O_b)$ and $S'(O_b, O_a)$, which ensures the symmetry of the similarity.

Sampling AL Queries. We use the proposed CCMS to sample representative image set from the candidate pool as the AL query Q. The objective of the diversity-based sampling is formally written as:

$$Q = \min_{Q' \subseteq H : |Q'| = b} \Big\{ \max_{I_i, I_j \in Q'} S(O_i, O_j) \Big\} \tag{8}$$

where H is the candidate pool output from the first stage, and b is the AL budget. However, Eq. 8 is an NP-Hard problem [39], so we follow [39] to use the k-Center-Greedy algorithm to get a $2 - OPT$ solution. Another problem of the objective in Eq. 8 is that it only maximises the diversity of the selected samples while ignoring how representative they are, which may cause the selection to favour data outliers. To deal with the problem, we further run a k-means++ algorithm using the results of k-Center-Greedy algorithm as the initial centroids. However, here we can only compute the similarities of every pair of images and are not able to compute the actual *mean* of every cluster to update its centroids.

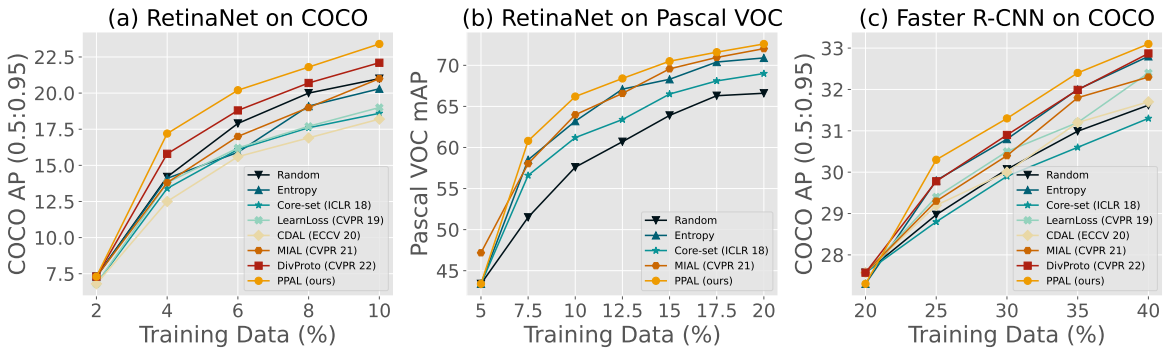


Figure 4. Comparison between the proposed method and the state-of-the-art active learning algorithm for object detection in three different benchmark settings. (a) RetinaNet on COCO; (b) RetinaNet of Pascal VOC; (c) Faster R-CNN on COCO.

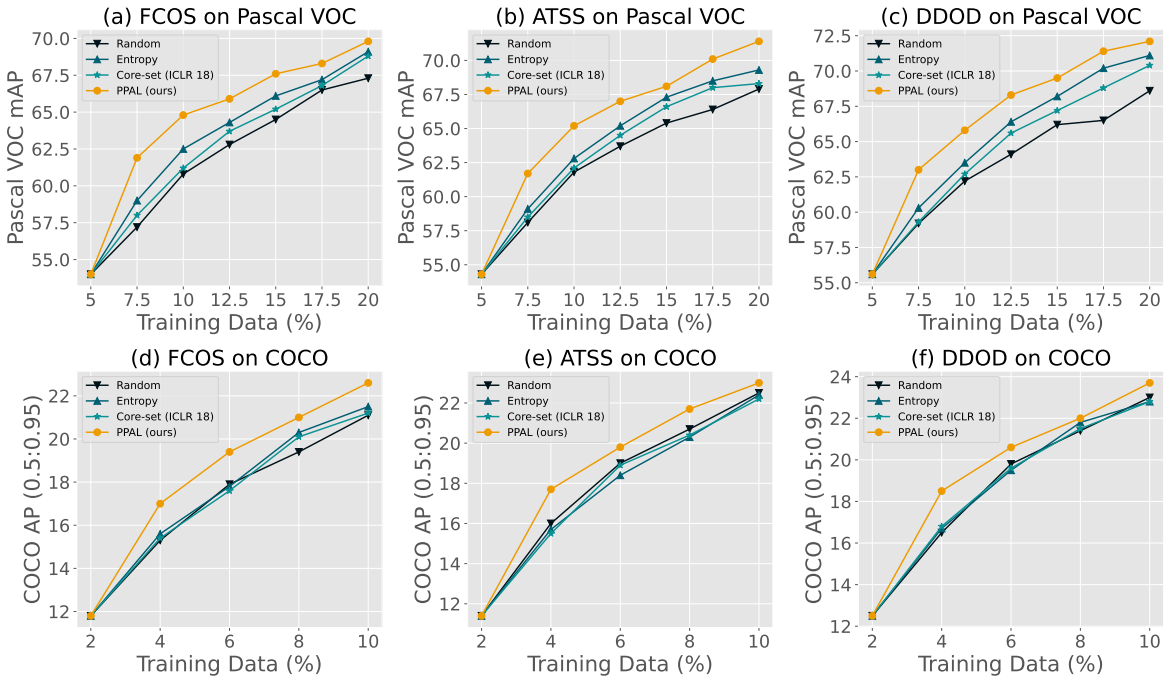


Figure 5. Active learning on Pascal VOC and COCO using (a) Anchor-free FCOS; (b) Anchor-based ATSS; (c) Anchor-based DDOD.

We solve this problem by assigning the new centroids as the images whose summed similarities to other images in their cluster are maximised. Finally, the resulting AL queries are sent to human experts for annotation.

4. Experiments

4.1. Experiment Settings

Dataset settings. We benchmark the proposed PPAL using two datasets: COCO [26] and Pascal VOC [14]. For COCO we use *train2017* set for training, and evaluate the models on the *mini-val* set. For Pascal VOC, we use *train2007+2012* for training and *test2007* for testing. When comparing with previous works, we follow their dataset split settings [43, 50] to ensure fair comparisons. Note that those settings may vary across different detectors. Ablation studies are conducted on Pascal VOC with a unified setting: 5% of the training data are first sampled as the initial set, then 6 rounds of active learning are conducted where at each round 2.5% extra data are queried. To overcome randomness, we run all experiments using three different initial training sets and report the averaged performance.

Model settings. By default, we follow [50] to set our model and training recipes, which ensures a fair comparison of our method and previous works. We implement our code base using the MMDetection toolkit [8]. For both dataset we train the models for 26 epochs and decay the learning rate by 0.1 at the 20th epoch. We use ResNet-50 [18] as the default backbone network. All experiments are conducted using 8 2080Ti NVIDIA GPUs. We follow [9] to set the ξ in Eq.1 to 0.6; we set base EMA momentum m^0 to to 0.99 following common practices. Other hyper-parameters are empirically set without careful tunning: α and β in Eq.4 are set to 0.3 and 0.2, and the budget expansion ratio δ is set to

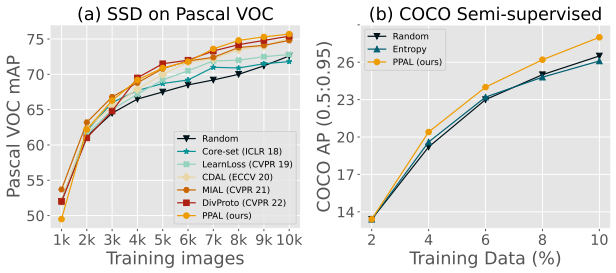


Figure 6. (a) Comparison between the proposed method and the state-of-the-art active learning algorithm for object detection on Pascal VOC with SSD object detector. (b) Active Learning performance on COCO using semi-supervised Soft Teacher detector.

4. We run the kmeans++ algorithm for 100 iterations.

4.2. Main Results

Comparing with the state-of-the-art. In Fig.4, we compare our PPAL with previous state-of-art active learning strategies on three different settings: COCO RetinaNet [50], Pascal VOC RetinaNet [50], and COCO Faster R-CNN [43]. We follow the training and dataset settings in [50] for the first two comparisons and follow [43] for the last. For the entropy-based sampling baseline, we sum up the classification entropy of all detected objects as the image uncertainty, which results in a better performance than previous works [10, 12, 50] that compute image-wise uncertainty using the averaged or the maximum object entropy. As shown by the results, our PPAL outperforms all computing methods in all three settings. Specifically, in the initial rounds (≤ 3), our method outperforms all competing methods by large margins, e.g., it outperforms MIAL [50] by 3.4 AP and 2.8 mAP on COCO and Pascal VOC respectively, which suggests that PPAL can provide the detectors with highly informative samples when they are not well trained. In the later rounds (> 3), although the AP gaps between PPAL and the competing methods decrease, it still maintains the lead. In addition to PPAL's ability to mine informative data samples, it also owns a better generalisation ability than the competing methods [48, 50], because it does not modify the model architecture or training pipeline.

Performance on more detectors. Our PPAL is highly generalisable, which allows us to easily apply it to different object detectors. In Fig.5, we show the Pascal VOC and COCO results when applying it to three recently proposed object detectors: FCOS [41], ATSS [52] and DDOD [9], in which FCOS is anchor-free, and the rest are anchor-based. We also shows the results of three baselines: *Random*, *Entropy* (uncertainty-based), and Core-set [39] (diversity-based). The results show that PPAL is consistently better than the baselines for all three detectors, verifying its effectiveness and good generalisation ability.

SSD Experiments. In Fig. 6 (a), we compare PPAL with previous works on Pascal VOC using the SSD [27] detector. We follow the training recipes in MIAL [50] to train the model for 300 epochs at each active learning round. Unlike other detectors [25, 34] that were used in the main paper, in SSD objects detected from different feature levels are computed using different convolutional kernels. In this case, we are unable to compute those objects' distances using their visual features, so we follow CDAL [1] to compute the distances of their classification probability vectors using KL-divergence. The results show that PPAL is able to achieve a better performance than all competing methods. Notably, although the initial performance of our SSD model is inferior to others, our active learning approach can still enable the model to surpass others in the later stages.

Stage 1	Stage 2	mAP on % of labelled images					
		7.5%	10%	12.5%	15%	17.5%	20%
Random		51.5±1.5	57.6±1.6	60.7±1.3	63.9±1.0	66.3±1.1	66.6±1.0
Entropy	None	58.5±0.7	63.2±0.6	67.1±0.3	68.3±0.3	70.4±0.3	70.9±0.2
DCUS	None	60.5±0.9	64.4±0.6	67.2±0.7	68.7±0.3	70.4±0.2	71.6±0.2
Rand	CCMS	56.6±1.0	61.2±0.5	64.7±0.7	66.9±0.4	68.4±0.3	70.3±0.2
Entropy	CCMS	59.0±0.6	64.5±0.6	67.6±0.4	69.2±0.4	71.1±0.3	72.0±0.3
D-Freq	CCMS	59.3±0.6	64.2±0.5	67.9±0.5	68.8±0.5	71.4±0.3	71.8±0.3
DCUS	Rand	60.3±0.7	64.1±0.5	67.6±0.5	68.9±0.4	70.5±0.3	72.0±0.2
DCUS	Global	58.8±0.6	64.5±0.4	66.9±0.2	68.6±0.2	69.3±0.4	71.8±0.2
DCUS	FPN	59.4±0.6	64.3±0.3	67.3±0.3	68.6±0.2	70.0±0.1	71.6±0.1
DCUS	Jaccard	60.3±0.4	65.1±0.4	67.3±0.4	68.6±0.3	70.5±0.3	71.9±0.2
DCUS	CCMS	60.8±0.5	66.2±0.4	68.4±0.2	70.5±0.4	71.6±0.3	72.6±0.2

Table 1. Ablation studies of Difficulty Calibrated Uncertainty Sampling (DCUS) and Category Conditioned Matching Similarity (CCMS) using VOC RetinaNet. The 1st round mAP is 43.4±2.2.

Semi-supervised Experiments. In Fig. 6 (b), we compare PPAL with *Random* / *Entropy* baselines using the semi-supervised detector SoftTeacher [45]. We followed the experiment settings in Sec.4.1 to run five rounds of active learning on COCO. In each round, the model was trained for 26 epochs (counted using labelled images), and the class difficulties were computed using the labelled images. We observe that *Entropy* active learning strategy achieves similar performance with the *Random* baseline, but our PPAL is better than both of them. The result validates PPAL's effectiveness in a semi-supervised learning setting.

4.3. Ablation Studies and Discussions

Effectiveness of PPAL. As shown in Tab.1, we start by comparing DCUS with random sampling and entropy-based sampling by using those strategies to sample the candidate pool and run the diversity-based sampling using our CCMS for similarity computing. We find that DCUS achieves a better performance, which validates DCUS's effectiveness as an uncertainty-based AL strategy. We also try to replace the difficulty coefficient described in Eq. 4 with $1 - f_c$ where f_c is class c's frequency in the training set, which is denoted as *DCUS-Freq*. Therefore the uncertainty-based sampling will favour those *long-tailed* classes. The result shows this strategy is inferior to our DCUS. A possible reason is that we find the class-wise AP does not strictly correspond to the training class frequencies. Therefore, in our proposed DCUS, we turn

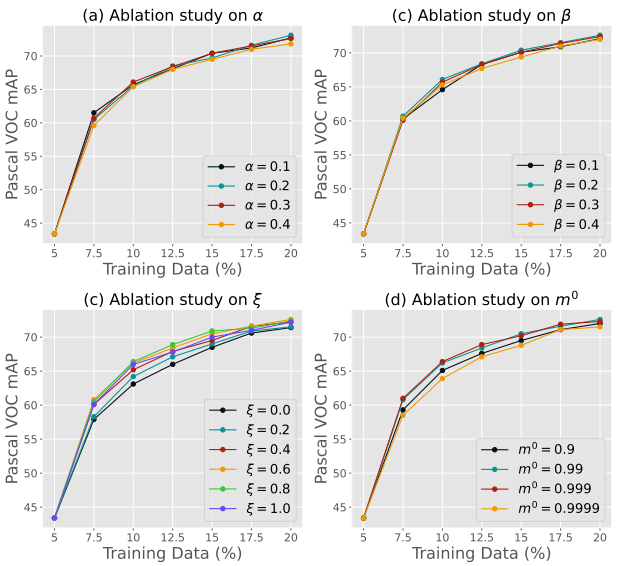


Figure 7. Ablation study on four hyper-parameter ablation studies in the proposed DCUS using Pascal VOC RetinaNet.

to explicitly modelling the class-wise detection difficulties, which is more suitable for uncertainty re-weighting. Also in Tab.1, we compare the proposed CCMS with four alternatives: (1) *Rand*, using random selection for the second stage; (2) *Global*, global similarity computed as the cosine similarity of the averaged feature maps from the backbone network's last layer; (3) *FPN*, the averaged global similarity of every feature pyramid layer; (4) *Jaccard*, the Jaccard similarity between the predicted category sets of two images. Note that nearly all previous works [39, 50] use global similarity for diversity sampling. We observe that the global similarity works even worse than random selection, demonstrating its failure in measuring the similarity of multi-instance images. The FPN similarity works slightly better than the global similarity and generates similar performance with random sampling. Our CCMS achieves the best result, which is consistently better than the three alternatives. These benchmark results validate that the proposed CCMS is more suitable to measure the similarities of multi-instance images in active learning. In addition, we observe that when diversity-sampling is fixed or disabled, our DCUS can achieve much better performance than entropy-based baseline in early AL rounds, but in later rounds their difference is marginal. On the other hand, although CCMS achieves a similar performance with random sampling in the first two rounds, but their performance gap starts to increase from the third round. From those observations we draw an important conclusion: *In active learning for object detection, uncertainty-based sampling is more critical in the early AL stages while diversity-based sampling is more essential in later rounds.*

Hyper-parameters. In Fig.7, we report the ablation study results to investigate the hyper-parameter settings in PPAL.

Pool Size (δ)	mAP on % of labelled images					
	7.5%	10%	12.5%	15%	17.5%	20%
1	60.5±0.5	64.4±0.4	67.2±0.4	68.7±0.3	70.4±0.4	71.6±0.3
2	60.4±0.6	65.8±0.4	67.6±0.3	69.0±0.4	70.6±0.2	71.5±0.3
3	60.5±0.6	65.4±0.4	67.7±0.5	70.0±0.3	71.3±0.3	71.8±0.2
4	60.8±0.5	66.2±0.4	68.4±0.2	70.5±0.4	71.6±0.3	72.6±0.1
5	61.2±0.5	65.0±0.6	68.0±0.4	70.1±0.5	71.4±0.2	72.6±0.1
6	59.8±0.7	65.9±0.5	67.9±0.3	70.1±0.3	71.4±0.2	72.3±0.2

Table 2. Ablation study using VOC RetinaNet on how the budget expanding ratio δ, which determines the size of the candidate pool in the first stage. The 1st round mAP is 43.4±2.2.

Uncertainty	mAP on % of labelled images					
	7.5%	10%	12.5%	15%	17.5%	20%
Random	51.5±1.5	57.6±1.6	60.7±1.3	63.9±1.0	66.3±1.1	66.6±1.0
Posterior	60.8±0.6	66.0±0.5	68.7±0.3	71.2±0.3	71.5±0.3	72.3±0.2
Margin	59.9±0.5	66.1±0.4	67.8±0.5	70.8±0.2	71.4±0.3	72.8±0.1
Entropy	60.8±0.5	66.2±0.4	68.4±0.2	70.5±0.4	71.6±0.3	72.6±0.2

Table 3. Comparison of different uncertainty measurements on VOC RetinaNet. The 1st round mAP is 43.4±2.2.

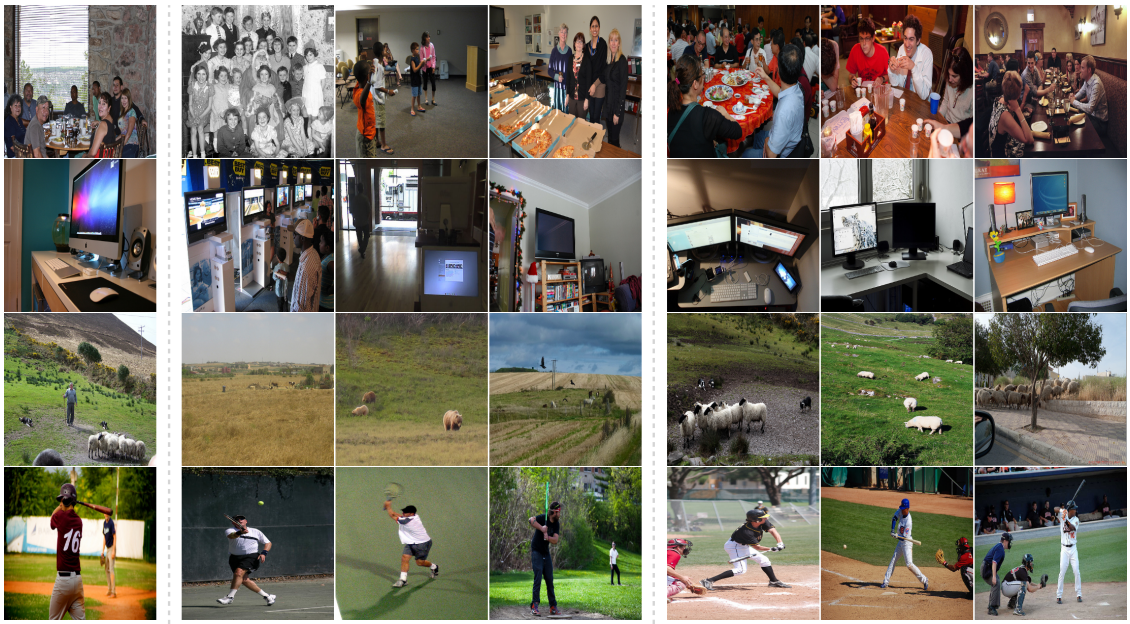
They are ξ in Eq. 1, m^0 in Eq. 3, and α and β in Eq. 4. The results suggest that PPAL's performance is stable although some optimal hyper-parameters settings may exist.

Size of the candidate pool. In Tab.2, we show how the budget expanding ratio δ, which determines the size of the candidate pool, affects PPAL's performance using Pascal VOC RetinaNet. We observe that when δ is around 4 our method can achieve the best performance. Specifically, we find that a too large δ, like $\delta \geq 6$, will harm PPAL's performance in early rounds, because in such cases the candidate pool will include many samples that the model is certain on, harming the overall information gain. On the other hand, a too-small expanding ratio, like $\delta < 3$, will harm PPAL's performance in later rounds because of the lacking of sample diversity in such small candidate pools.

Uncertainty Measurement. As presented in Sec.3.2, in PPAL we use entropy as the default uncertainty measurement. In Tab.3, we show that our method can achieve similar performance when generalising to other types of uncertainty measurement. Specifically, we test PPAL on Pascal VOC RetinaNet using two alternative uncertainty measurements: posterior probability [22] and probability margin [11]. We observe that the posterior probability has a very close performance to entropy. However, the probability margin achieves an inferior performance in the early rounds, but can get similar performances with the other two uncertainty measurements in later rounds.

Generalising to different backbones. In Tab.4, we show our approach can well generalise to other backbone networks. We test PPAL on Pascal VOC RetinaNet using the heavy-weighted ResNet-101 [18], light-weighted MobileNet v2 [38] and the high-performing transformer based SwinTransformer-Tiny [28]. The results show that PPAL is consistently better than random sampling by a large margin in all those backbone architectures, validating the strong generalisation ability of our approach.

Why is CCMS better than global similarity? Here we investigate why the proposed CCMS is better than global



Query | Global Similarity | Category Conditioned Matching Similarity (CCMS)

Figure 8. Image retrieval visualisation using global similarity and the proposed CCMS on COCO *mini-val* using RetinaNet.

Backbone	Method	mAP on % of labelled images					
		7.5%	10%	12.5%	15%	17.5%	20%
ResNet101	Rand	59.3±1.4	64.6±1.0	68.4±1.1	69.8±1.1	71.0±0.9	72.3±1.0
	PPAL	64.3±0.5	69.6±0.3	71.1±0.2	72.8±0.3	73.8±0.1	75.0±0.1
MobileNetV2	Rand	31.2±2.8	31.4±2.4	39.7±2.3	42.8±2.0	42.9±1.9	44.7±1.7
	PPAL	35.7±1.7	41.9±1.2	45.6±1.3	46.3±1.2	48.4±0.9	50.7±0.9
SwinTiny	Rand	63.8±1.0	67.9±0.9	70.0±1.0	71.2±0.6	71.8±0.8	73.3±0.6
	PPAL	68.4±0.3	71.2±0.4	72.7±0.2	74.8±0.1	75.4±0.1	76.2±0.1

Table 4. Comparison of PPAL and random sampling on VOC RetinaNet using different backbones. The 1st round mAPs are 51.0±1.8, 20.9±3.4, and 59.0±1.5 for all three backbones.

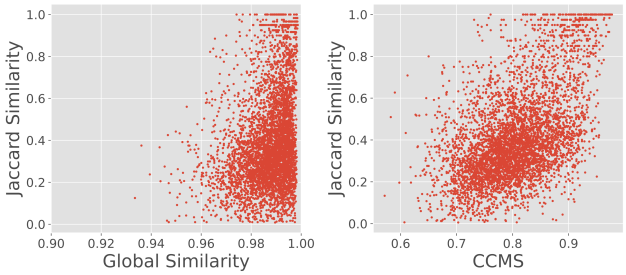


Figure 9. Comparison between global similarity and the proposed CCMS on image retrieval using RetinaNet on COCO *mini-val*. We retrieve 20 most similar images for each anchor image. The similarity numbers are normalised to $(0, 1)$.

similarity in measuring image-wise similarity for multi-instance images by running an image retrieval experiment on COCO *mini-val* set using a pre-trained RetinaNet detector: For each anchor image, we use both means of computing similarity to retrieve the 20 most similar images. Then we check whether the retrieved images depict similar scenes with the anchor image by measuring the averaged Jaccard similarity of object categories contained in those images. Specifically, for image I_i and I_j and their contained object category set C_i and C_j, their Jaccard similarity is defined as $\mathcal{J}_{ij} = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$. For example, if an image contains {dog, human} and the other image contains {dog, cat}, their Jaccard similarity 0.33. We show the result on the whole COCO *mini-val* set in Fig.9. We observe that the averaged CCMS and Jaccard similarity are positively correlated, i.e., a high averaged CCMS usually corresponds to a high averaged Jaccard similarity and vice versa. However, global similarity does not hold such correspondence to the class-wise Jaccard similarity. In Fig.8, we show the image retrieval results by visualising the 3 most similar images with the anchor. From it we get two observations: (1) Global similarity is usually biased toward the dominating objects in the image while ignoring other objects (1st and 2nd rows) and CCMS is not; (2) CCMS is better than global similarity in capturing the fine-grained details in the image (3rd and 4th rows). These results validate our argument that CCMS is a more suitable similarity computing method for diversity-based active learning for object detection, which usually uses multi-instance images as input.

5. Conclusion

We introduce a two-stage active learning algorithm for object detection. In the first stage, we propose Difficulty Calibrated Uncertainty Sampling to select a candidate pool of uncertain samples, and in the second stage we select a diverse query set using Category Conditioned Matching Similarity. We show our method can generalise well and outperform previous works in various architectures and datasets.

Acknowledgements. Chenhongyi Yang was supported by a PhD studentship provided by the School of Engineering, University of Edinburgh.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, 2020. 1, 2, 4, 6

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *ICLR*, 2020. 2

[3] Mustafa Bilgic and Lise Getoor. Link-based active learning. In *NeurIPS Workshop on Analyzing Networks and Learning with Graphs*, 2009. 1, 2

[4] Qi Cai, Yingwei Pan, Yu Wang, Jingen Liu, Ting Yao, and Tao Mei. Learning a Unified Sample Weighting Network for Object Detection. In *CVPR*, 2020. 2

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. *ECCV*, 2020. 2

[7] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[9] Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. Disentangle your dense object detector. In *ACM Multimedia*, 2021. 2, 3, 5, 6

[10] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *ICCV*, 2021. 2, 3, 6

[11] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *ICLR*, 2021. 2, 7

[12] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *CVPR*, 2022. 3, 6

[13] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S Shankar Sasrty. A convex optimization framework for active learning. In *ICCV*, 2013. 1, 2

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 2015. 5

[15] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*. PMLR, 2017. 2

[16] Yuhong Guo. Active instance sampling via matrix partition. *NeurIPS*, 2010. 1, 2

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 7

[19] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. DenseBox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 2

[20] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009. 1, 2

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 128(7):1956–1981, 2020. 1

[22] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings*, pages 148–156. Elsevier, 1994. 1, 2, 7

[23] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12. Springer, 1994. 1, 2

[24] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 2

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 6

[26] 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 *ECCV*, 2014. 1, 5

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2, 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 7

[29] Wenjie Luo, Alex Schwing, and Raquel Urtasun. Latent structured active learning. *NeurIPS*, 2013. 1, 2

[30] Mengyao Lyu, Jundong Zhou, Hui Chen, Yijie Huang, Dongdong Yu, Yaqian Li, Yandong Guo, Yuchen Guo, Liuyu Xiang, and Guiguang Ding. Box-level active detection. In *CVPR*, 2023. 2

[31] Oisin Mac Aodha, Neill DF Campbell, Jan Kautz, and Gabriel J Brostow. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, 2014. 1, 2

[32] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *ICCV*, 2021. 2

[33] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 6

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*, 2019. 2

[36] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *ECML*. Springer, 2006. 1, 2

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 4

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018. 7

[39] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *ICLR*, 2018. 1, 2, 4, 6, 7

[40] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012. 1, 2

[41] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 2, 6

[42] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *T-CSVT*, 27(12):2591–2600, 2016. 1, 2

[43] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *CVPR*, 2022. 3, 5, 6

[44] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *ICCV*, 2021. 2

[45] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, pages 3060–3069, October 2021. 6

[46] Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. Contrastive object-level pre-training with spatial noise curriculum learning. *arXiv preprint arXiv:2111.13651*, 2021. 4

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

[48] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. 2, 6

[49] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *CVPR Workshops*, 2022. 2

[50] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *CVPR*, 2021. 2, 3, 5, 6, 7

[51] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ECCV*, 2022. 2

[52] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020. 2, 6

[53] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019. 2

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2