Entropy-based Active Learning for Object Detection with Progressive Diversity Constraint

Jiaxi Wu[1,2], Jiaxin Chen[2], Di Huang[1,2*]
[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
{wujiaxi,jiaxinchen,dhuang}@buaa.edu.cn

Abstract

Active learning is a promising alternative to alleviate the issue of high annotation cost in the computer vision tasks by consciously selecting more informative samples to label. Active learning for object detection is more challenging and existing efforts on it are relatively rare. In this paper, we propose a novel hybrid approach to address this problem, where the instance-level uncertainty and diversity are jointly considered in a bottom-up manner. To balance the computational complexity, the proposed approach is designed as a two-stage procedure. At the first stage, an Entropy-based Non-Maximum Suppression (ENMS) is presented to estimate the uncertainty of every image, which performs NMS according to the entropy in the feature space to remove predictions with redundant information gains. At the second stage, a diverse prototype (DivProto) strategy is explored to ensure the diversity across images by progressively converting it into the intra-class and inter-class diversities of the entropy-based class-specific prototypes. Extensive experiments are conducted on MS COCO and Pascal VOC, and the proposed approach achieves state of the art results and significantly outperforms the other counterparts, highlighting its superiority.

1. Introduction

During the past decade, visual object detection [23, 30] has been greatly advanced by deep Convolutional Neural Networks (CNN) [12, 27] with persistently increasing performance reported. Unfortunately, strong CNNs generally make use of huge amounts of annotated data to fit extensive numbers of parameters, and training such detectors requires bounding-box labels on images, which is quite expensive and time-consuming. As one of the most promising alternatives to alleviate this dilemma, active learning [25, 38] aims to reduce this high cost by consciously selecting more informative samples to label, and it is expected to deliver a higher accuracy with much fewer annotated images compared to that conducted in the random way.

In the community of computer vision, active learning is mainly discussed on image classification [15, 25, 28], where current methods roughly go into two categories, *i.e.* uncertainty-based [9, 36] and diversity-based [22, 25]. Uncertainty-based methods [9, 36] screen informative samples from entire databases according to their ambiguities [3, 9, 15, 36]. As the samples are separately predicted, they are efficient but tend to incur high correlations. Diversity-based methods [1, 22, 25] claim that informative samples are the representatives of the whole data distribution and identify a subset using distance metric [25] or class probability [1]. They prove effective for small models, but suffer from high computational complexity. In addition, there exists another trend to combine the uncertainty- and diversity-based methods as hybrid ones [2, 6, 35], and the achieved superiority figures out a promising alternative to other tasks.

As we know, object detection is more complicated than image classification, where object category and location are simultaneously output. In this case, active learning is desired to deal with various numbers of objects within images and the essential issue is to make image-level decisions according to instance-level predictions. The diversity-based method, CDAL [1], applies spatial pooling to roughly approximate instance aggregation and formulates image selection as a reinforcement learning process. Regarding uncertainty-based methods, Learn Loss [36] designs a task-free loss prediction module, and computes the image uncertainty by image-level features instead of instance-level ones, while MIAL [37] defines the image uncertainty as that of the top-K instances and estimates it with multiple instance learning based re-weighting. Since the diversity-based methods do not fully make use of categorical information and the uncertainty-based ones do not well measure the discrepancy of informative samples, the two types of methods leave room for performance improvement.

In this study, we propose a novel hybrid approach to

[*] Corresponding author.

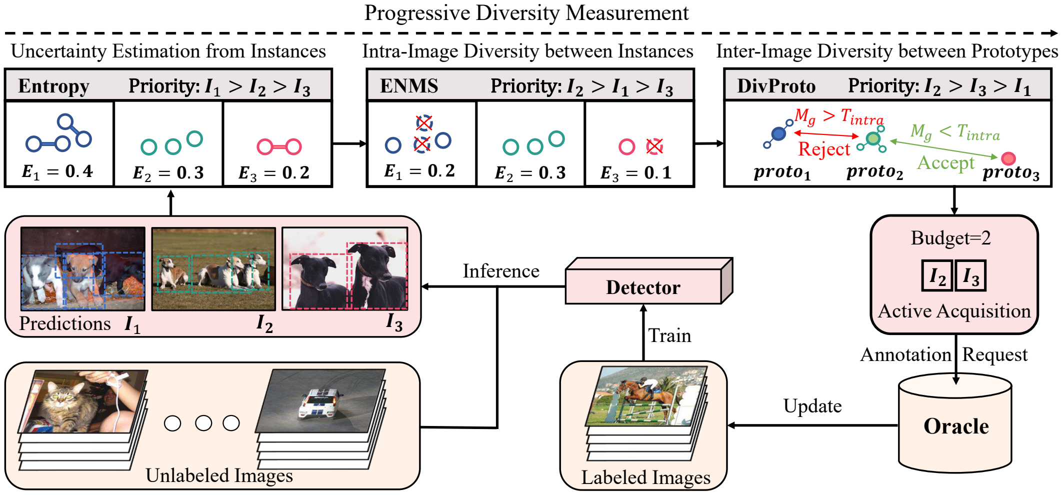


Figure 1. Framework overview. The hollow circles refer to uncertainty predictions and the solid ones denote the aggregated prototypes. At each cycle, the detector is trained with labeled images and infers the unlabeled ones. Instance uncertainty is first computed based on the entropy. ENMS then performs on each image to remove redundant instances. DivProto aggregates the instances of each image to prototypes and rejects the images close to the selected ones. The priority of active acquisition is illustrated with 3 examples: I_1, I_2, I_3. At the end of each cycle, the selected images (*e.g.* I_2, I_3) are labeled by an oracle.

active learning for object detection, which considers both the uncertainty and diversity at the instance level. To balance the computational complexity, the proposed approach works in a two-stage manner, as Fig. 1 displays. At the first stage, we estimate the uncertainty of each image by an Entropy-based Non-Maximum Suppression (ENMS). ENMS performs Non-Maximum Suppression on the calculated entropy in the feature space to remove instances that bring redundant information gains, where a bigger value of the entropy refined by ENMS indicates the selection priority of an unlabeled image. At the second stage, unlike existing uncertainty-based methods [24, 37] which choose the top-K images for annotation, we introduce the diverse prototype (DivProto) strategy to ensure instance-level diversity across images. It employs the prototypes [29, 33] as the image-level representatives by aggregating the class-specific instances, and decomposes the cross-image diversity into the intra-class and inter-class ones. We then acquire the images of the minority classes for inter-class diversity and reject the ones that incur redundancy for intra-class diversity. In this way, the proposed approach combines the advantages of the uncertainty and diversity based ones in a bottom-up manner. We evaluate the proposed approach on MS COCO [19] and Pascal VOC [7, 8] and deliver state of the art scores on both of them, highlighting its effectiveness.

2. Related Work

2.1. Active Learning on Image Classification

As stated, the majority of the studies on active learning in computer vision target on image classification and are mainly categorized into diversity-based [1, 25] and uncertainty-based [9, 36] ones.

The diversity-based methods screen a subset of samples to represent the global distribution by clustering [22] or matrix partition [11] techniques. Core-set [25] defines active learning as a core-set selection problem and adopts k-center approximation. To improve efficiency, CDAL [1] replaces distance based similarity with the KL divergence. Those methods are theoretically complete but computationally inefficient when dealing with high-dimensional data.

The uncertainty-based methods select ambiguous samples which are regarded as the most informative to the entire dataset [3, 9, 15, 36]. Many efforts are made to estimate the data uncertainty, *e.g.* the entropy of class posterior probabilities [15]. In this case, [9] introduces Bayesian CNNs as an expert; [3] employs deep ensembles and Monte-Carlo dropout; and Learn Loss [36] proposes a task-free image-level loss prediction module. The methods above are efficient, but bring in redundant samples for annotation.

Some alternatives [6, 14, 34] combine the advantages of both types. With the uncertainty and diversity scores, [6] simply chooses the minimal; [35] emphasizes the diversity at early cycles and moves to the uncertainty gradually; [13] views the fusion as a multi-armed bandit problem and reweights different scores. VAAL [28] performs uncertainty estimation on whether a data point belongs to the labeled or unlabeled pool, and acquires the samples most similar to the latter. SRAAL [38] further exploits the uncertainty estimator and the supervised learner to enclose annotation information. BADGE [2] models the uncertainty and diversity by the gradient magnitudes and directions from the last layer respectively. The hybrid methods achieve promising results and suggest a new fashion for other tasks.

2.2. Active Learning on Object Detection

Object detection has been greatly progressed by CNNs [12, 27] in the past few years mainly under the one-stage [18, 20, 30] and two-stage [10, 23] frameworks. As detection annotation is more expensive and time-consuming, active learning comes into focus in this branch, and preliminary attempts demonstrate its necessity [4, 16, 21, 24, 32]. Meanwhile, with both object category and location to predict, this task is more challenging.

Both the diversity-based and uncertainty-based methods have been recently adapted to object detection, and they extend direct image-level decision by integrating instance-level predictions. For the former, CDAL [1] represents the image using the detection features after spatial pooling to approximate this process. In spite of certain potential, a substantial performance gain requires global instance-level feature comparison, which incurs a vast complexity. For the latter, Learn Loss [36] employs holistic image-level features for uncertainty estimation and with the task-free loss prediction module, it directly evaluates how much information an unlabeled image contributes. MIAL [37] selects the image by measuring its uncertainty based on that of the top-K instances re-weighted in a multiple instance learning framework, with noisy ones suppressed and representative ones highlighted. They ignore instance-level correlation within the whole data pool and thus deliver much redundancy. To address the issues above, this paper presents a way to jointly use their strengths to advance object detection.

3. Problem Statement

This section starts with the formulation of active learning for object detection. The generic pipeline can be roughly grouped into three steps: (1) inference on unlabeled images with the existing detector, (2) image acquisition and annotation under a budget, and (3) detector training and evaluation on newly labeled images. These three steps execute in a loop and each iteration is viewed as a cycle (or stage). After each active learning cycle, the performance of the detectors represents the ability of the active acquisition methods since they select different images to annotate, where a fixed image amount is adopted as the annotation budget [1, 36, 37]. As detector training and evaluation are set in the same way, we focus on exploring a more effective acquisition method.

Suppose we have a large collection of candidate images $\{I_i\}_{i\in[n]}$ as well as a selected image set $\mathcal{S} = \{I_{s(j)}|s(j) \in [n]\}_{j\in[m]}$, where $[n] = \{1, \cdots, n\}$ and $[m] = \{1, \cdots, m\}$. Note that $\mathcal{S}$ denotes the labeled subset before each active acquisition cycle, which is initially chosen at random. Given a budget b, the batch active learning algorithm aims to acquire an image subset $\Delta\mathcal{S}$ in each cycle such that $|\Delta\mathcal{S}| = b$. $\Delta\mathcal{S}$ is subsequently labeled by an oracle, and is applied to update $\mathcal{S}$ as $\mathcal{S} := \mathcal{S} \cup \Delta\mathcal{S}$. The oracle is requested to provide labels $\mathcal{Y} = \{y_{s(j)}\}_{j\in[m]}$ for each selected image. The learning model $D_{\mathcal{S}}$ is successively trained by $\mathcal{S}$ and $\mathcal{Y}$.

As depicted in Core-set [25], the active learning problem is defined as minimizing the core-set loss $\sum_{i\in[n]} l(I_i, y_i; D_{\mathcal{S}})$, where y_i is the label of I_i. In the setting of object detection, the detector $D_{\mathcal{S}}$ is decomposed to an encoder $P_{\mathcal{S}}$ and a successive predictor $A_{\mathcal{S}}$. $P_{\mathcal{S}}$ encodes a set of spatial positions $\{pos_k\}_{k\in[t]}$ in I_i to a set of features $P_{\mathcal{S}}(I_i) = \{P_{\mathcal{S}}(I_i, k)\}_{k\in[t]}$ by adopting the receptive fields [20, 23] or positional embeddings [5] from $D_{\mathcal{S}}$. Afterwards, $A_{\mathcal{S}}$ predicts $A_{\mathcal{S}}(P_{\mathcal{S}}(I_i)) = \{\tilde{y}_{i,k}, c_{i,k}, p_{i,k}\}_{k\in[t]}$ based on $P_{\mathcal{S}}(I_i)$, where $\tilde{y}_{i,k}$, $c_{i,k}$ and $p_{i,k}$ are the predicted bounding box, object class and confidence score, respectively. The image-level core-set loss $l(\cdot)$ for object detection can be reformulated as: $\sum_{k\in[t]} l_D(P_{\mathcal{S}}(I_i, k), y_{i,k}; A_{\mathcal{S}})$, where l_D is the instance-level loss function. To adopt the Core-set based solution, $l(\cdot)$ should be Lipschitz continuous as required by *Theorem 1* in [25]. However, $P_{\mathcal{S}}(I_i)$ is unordered, and thus is difficult to be explicitly defined, making $l(\cdot)$ not Lipschitz continuous.

To address this issue, inspired by the uncertainty-based studies [24, 37], we alternatively explore the empirical uncertainty from $P_{\mathcal{S}}(I_i)$ and adopt the entropy-based formulation. Specifically, we calculate the following entropy [26] for the k-th instance:

$$\mathbb{H}(I_i, k) = -p_{i,k} \log p_{i,k} - (1 - p_{i,k}) \log (1 - p_{i,k}), \quad (1)$$

where $p_{i,k}$ is the confidence score predicted as the foreground of a certain category and $1-p_{i,k}$ as the background.

From Eq. (1), the image-level *basic detection entropy* is defined by replacing $l_D(\cdot)$ with $\mathbb{H}(I_i, k)$ in $l(\cdot)$ as below:

$$\mathbb{H}(I_i|D_{\mathcal{S}}) = \sum\nolimits_{k\in[t]} \mathbb{H}(I_i, k). \quad (2)$$

Based on $\mathbb{H}(I_i|D_{\mathcal{S}})$, the unlabeled images are sorted, and the top-K images are selected as the acquisition set $\Delta\mathcal{S}$.

As visually similar bounding boxes contain redundant information which are not preferred when training robust detectors, it is desirable to select the most informative ones and abandon the rest. Moreover, such information redundancy occurs not only within each image but also across images, making it more difficult to retain the instance-level diversity. There still lacks a hybrid approach that considers the instance-level evaluation and achieves the image-level acquisition in the mean time.

4. Method

4.1. A Hybrid Framework

In this subsection, we describe the details about our proposed hybrid framework, specifically designed for active learning on object detection.

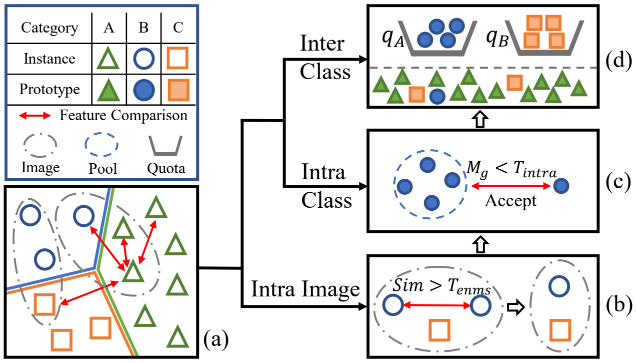


Figure 2. The hierarchy of the instance-level diversity is displayed in (a). (b) refers to the intra-image diversity by removing the instance-level redundancy via ENMS. (c) and (d) strengthen the intra-class and inter-class diversities across images formulated with the class-specific prototypes, respectively.

As shown in Fig. 1, the proposed framework mainly consists of three modules: uncertainty estimation using the basic detection entropy, Entropy-based Non-Maximum Suppression (ENMS) and the diverse prototype (DivProto) strategy. The basic detection entropy in Eq. (2) is adopted to quantitively measure the image-level uncertainty from object instances. ENMS is subsequently presented to remove redundant information according to the entropy, thus strengthening the instance-level diversity within images. DivProto further ensures the instance-level diversity across images by converting it into the inter-class and intra-class diversities formulated with class-specific prototypes.

To be specific, the hierarchy of our method for diversity enhancement is illustrated in Fig. 2. The overall instance-level diversity (a) is divided into the intra-image diversity (b) via ENMS and the inter-image diversity is accomplished by DivProto, which is then decomposed into the inter-class and intra-class ones as shown in (c) and (d), respectively. By virtue of this progressive way, the diversity constraints on the predicted instances are effectively undertaken. The details are elaborated in the rest part.

4.2. ENMS for Intra-Image Diversity

As Eq. (2) depicts, the basic detection entropy $\mathbb{H}(I_i|D_S)$ is a simple sum of the entropy of the candidate bounding boxes. Nevertheless, existing object detectors often generate a large amount of proposal bounding boxes with heavy overlaps, incurring severe spatial redundancy and high computational cost. This issue can be partially mitigated by applying Non-Maximum Suppression (NMS) [20, 23], based on which bounding boxes belonging to the same instance are merged to a unified one. But NMS cannot deal with the instance-level redundancy, *i.e.* instances with similar appearances presenting in the same context, which is supposed to be reduced in active acquisition.

Algorithm 1 Entropy-based Non-Maximum Suppression

```
Input: the predicted classes $\{c_{i,k}\}_{k\in[t]}$
       the confidence scores $\{p_{i,k}\}_{k\in[t]}$
       the instance-level features $\{\boldsymbol{f}_{i,k}\}_{k\in[t]}$
       the threshold $T_{enms}$ (0.5 by default)
Output: the image-level entropy $E_i$ after ENMS
Initialize: $E_i := 0$
Calculate the instance entropy $\{\mathbb{H}(I_i,k)\}_{k\in[t]}$ according to Eq. (1)
Initialize the indicating set $S_{ins} := [t]$
while $S_{ins} \neq \varnothing$ do
    Select the most informative instance $k_{pick}$ according to $k_{pick} := \arg\max_{k\in[S_{ins}]} \mathbb{H}(I_i,k)$ from $S_{ins}$ and update $S_{ins} := S_{ins} - \{k_{pick}\}$
    Update the entropy $E_i := E_i + \mathbb{H}(I_i,k_{pick})$
    for $j$ in $S_{ins}$ do
        if $c_{i,j} = c_{i,k_{pick}}$ and $\mathrm{Sim}(\boldsymbol{f}_{i,j},\boldsymbol{f}_{i,k_{pick}}) > T_{enms}$ then
            Remove the instance $j$ as $S_{ins} := S_{ins} - \{j\}$
        end if
    end for
end while
```

To overcome this shortcoming of NMS, we propose a simple yet effective *Entropy-based Non-Maximum Suppression* (ENMS) as a successive step of NMS for instance-level redundancy removal. Specifically, we first compute the following Cosine distance $\mathrm{Sim}(\cdot,\cdot)$ to measure pair-wise inter-instance duplication: $\mathrm{Sim}(\boldsymbol{f}_{i,k},\boldsymbol{f}_{i,j}) = \frac{\boldsymbol{f}_{i,k}^T\cdot\boldsymbol{f}_{i,j}}{\|\boldsymbol{f}_{i,k}\|\|\boldsymbol{f}_{i,j}\|}$, where $\boldsymbol{f}_{i,k}$ is the feature of the instance k in the image I_i extracted by $P_S(\cdot)$. Subsequently, ENMS is performed on the indicating set S_{ins} initialized as $[t]$, where $[t]$ is the set of all instances in I_i. As summarized in Algorithm 1, the basic idea of ENMS is to select the most informative instance k_{pick} from S_{ins} with the corresponding entropy $\mathbb{H}(I_i,k_{pick})$ being accumulated for the image-level entropy E_i. Meanwhile, the remaining within-class instances that are similar to k_{pick} (*i.e.* the pair-wise similarity is larger than a threshold T_{enms}) are deemed as redundant ones, and further removed from S_{ins}. The procedure aforementioned is iteratively conducted until S_{ins} becomes empty.

It is worth noting that ENMS only compares instances from the same categories w.r.t. the selected informative instance, and is thus computationally efficient. Meanwhile, ENMS extracts the instance-level features on-the-fly, which significantly reduces the memory cost. Additionally, ENMS can mitigate the unbalanced amount of instances per image, by means of redundant instance removal.

4.3. Diverse Prototype for Inter-Image Diversity

ENMS enhances the intra-image diversity, and the inter-

Algorithm 2 Diverse Prototype

```
Input: the labeled images $\mathcal{S}$
       the unlabeled images $\{I_i\}_{i\in[n]} - \mathcal{S}$
       the budget $b$ and the thresholds $T_{intra}$ and $T_{inter}$
Output: the selected image set $\Delta\mathcal{S}$ to be labeled
Initialize: $\Delta\mathcal{S} := \varnothing$
Calculate the entropy $\{E_i\}$ as well as the prototypes $\{\{\boldsymbol{proto}_{i,c}\}_{c\in[C]}\}$ for the set of the unlabeled images $\{I_i\}_{i\in[n]} - \mathcal{S}$ by ENMS and Eq. (3), respectively.
Calculate the quotas $\{q_c\}_{c\in[C_{minor}]}$ based on $\mathcal{S}$
Sort $\{I_i\}_{i\in[n]} - \mathcal{S}$ in descending order according to $\{E_i\}$
for $i$ in $\left[\left|\{I_i\}_{i\in[n]} - \mathcal{S}\right|\right]$ do
    if $M_g(I_i,[C]) < T_{intra}$ and $M_p(I_i,[C_{minor}]) > T_{inter}$ then
        Select $I_i$ and update $\Delta\mathcal{S} := \Delta\mathcal{S} \cup \{I_i\}$
        for $c$ in $[C_{minor}]$ do
            Update $q_c := q_c - 1$ if $p(i,c) > T_{inter}$
            Update $C_{minor} := C_{minor} - 1$ if $q_c = 0$
        end for
    end if
end for
Fill up $\Delta\mathcal{S}$ with the rest images from the sorted set $\{I_i\}_{i\in[n]} - \mathcal{S}$ until $|\Delta\mathcal{S}| = b$
```

image diversity, *i.e.* the redundancy across images, still remains. Most conventional approaches [1] address this issue based on holistic image-level features, which are too coarse to fulfill instance-level processing in object detection. Some re-weighting based methods [13, 35] can be adapted from image-level to instance-level, mitigating the inter-image redundancy. However, they need to compute the distances between all instance pairs, which incurs high memory and computational cost. Besides, current studies rarely consider the imbalanced classes of instances, making it hard to estimate the normalized diversity.

Inspired by the previous attempts [29, 33, 39], we introduce the prototypes to address the drawbacks above. Concretely, the i-th prototype of class c is formulated as:

$$\boldsymbol{proto}_{i,c} = \frac{\sum_{k\in[t]} \mathbb{1}(c, c_{i,k}) \cdot \mathbb{H}(I_i, k) \cdot \boldsymbol{f}_{i,k}}{\sum_{k\in[t]} \mathbb{1}(c, c_{i,k}) \cdot \mathbb{H}(I_i, k)}, \tag{3}$$

where $\mathbb{1}(c, c_{i,k})$ equals to 1 if $c = c_{i,k}$, and 0 otherwise.

As shown in Eq. (3), the prototype is formulated based on the entropy and the predicted class instead of the confidence score as in existing work [33], since our framework focuses on the information gain. Therefore, the instances with high classification confidence scores contribute less to the prototype than the uncertain ones.

Based on ENMS and the prototypes, we propose the *Diverse Prototype* (DivProto) strategy to enhance the instance-level diversity across images. Specifically, we firstly sort the unlabeled images according to their entropy $\{E_i\}$ via ENMS in descending order. Subsequently, we improve the intra-class diversity via intra-class redundancy rejection as shown in Fig. 2 (c) and the inter-class diversity via inter-class balancing as in Fig. 2 (d).

Intra-class Diversity. Given a candidate image I_i and the prototypes of the acquired set $\Delta\mathcal{S}$, the intra-class diversity of I_i is measured by the following metric:

$$M_g(I_i, [C]) = \min_{c\in[C]} \max_{j\in|\Delta\mathcal{S}|} \mathrm{Sim}(\boldsymbol{proto}_{j,c}, \boldsymbol{proto}_{i,c}) \tag{4}$$

In Eq. (4), we can observe that: 1) by using M_g, the inter-image diversity is measured by the similarity between the prototypes instead of instance-level pair-wise comparison, thereby remarkably reducing the computational complexity, and 2) $M_g(I_i, [C])$ encodes the similarity between intra-class prototypes across images and increases if I_i is more similar to the picked image set $\Delta\mathcal{S}$.

Based on the observations above, we thus adopt the following *intra-class redundancy rejection* process to enhance the intra-class diversity across images: reject the image I_i when $M_g(I_i, [C])$ is larger than a threshold T_{intra} (0.7 by default), and accept otherwise.

Inter-class Diversity. Though the intra-class diversity can be enhanced based on $M_g(I_i, [C])$, the image set acquired by the intra-class rejection process tends to favor certain classes (*i.e.* the majority classes), leading to severe class imbalance. To deal with this issue, we increase the inter-class diversity by introducing the *inter-class balancing* process, *i.e.* adaptively providing more budgets for the minority classes than the majority ones.

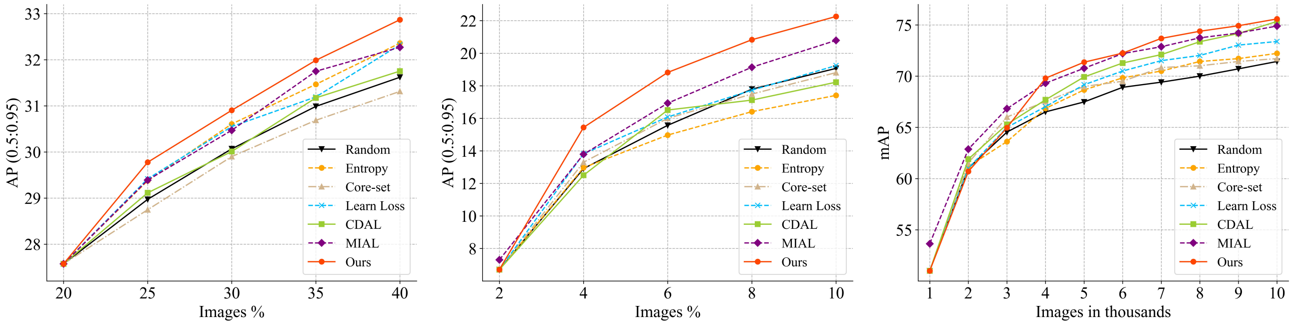
Concretely, we first build the minority class set $[C_{minor}]$ by sorting the overall classes according to the frequency of occurrences in the labeled image set $\mathcal{S}$ and selecting the classes with the C_{minor} fewest instances, where $C_{minor} = \alpha C$ $(0 < \alpha < 1)$. We assign each minority class $c \in [C_{minor}]$ a relatively large quota $q_c = \frac{\beta}{\alpha C} b$ $(\alpha < \beta < 1)$ as the class-specific budget.

For an unlabeled image I_i, we check if it contains the instances from the minority classes by computing

$$M_p(I_i, [C_{minor}]) = \max_{c\in[C_{minor}]} p(i, c), \tag{5}$$

where $p(i,c) = \max_{k\in[t]} \mathbb{1}(c, c_{i,k}) \cdot p_{i,k}$ estimates the probability about the existence of instances from the class c in I_i.

In this work, we adopt a threshold T_{inter} (0.3 by default), where the image I_i is accepted as containing minority classes if $M_p(I_i, [C_{minor}]) > T_{inter}$, and is rejected otherwise. Once I_i is accepted, the quota $\{q_c\}$ will be updated as $q_c := q_c - 1$ if $p(i,c) > T_{inter}$. During image acquisition, the class with $q_c = 0$ will be removed from the minority class set $[C_{minor}]$, while the number of the minority classes



(a) FRCNN on MS COCO

(b) RetinaNet on MS COCO

(c) SSD on Pascal VOC

Figure 3. Comparison results. (a)/(b) AP (%) on MS COCO by using different portions of training data; (c) mAP (%) on Pascal VOC. (a), (b) and (c) adopt Faster R-CNN and RetinaNet with ResNet-50, SSD with VGG-16, respectively.

C_{minor} is updated by $C_{minor} := C_{minor} - 1$. The whole process is terminated when C_{minor} reaches 0.

Since an image may contain instances from multiple minority classes and $\beta < 1$, the number of acquired images by using the inter-class balancing process should not exceed the budget b. We therefore fill up $\Delta\mathcal{S}$ with the remaining unlabeled images until the budget b runs out.

By performing the process above, we can make a balance w.r.t. the number of instances from various classes, and finally increase the inter-class diversity. The details of DivProto are summarized in Algorithm 2.

5. Experiments

5.1. Experimental Settings

Datasets. We evaluate the proposed method on two benchmarks for object detection: MS COCO [19] and Pascal VOC [7, 8]. **MS COCO** has 80 object categories with 118,287 images for training and 5,000 images for validation [19]. Similar to [28] in dealing with large-scale data, we report the performance with 20%, 25%, 30%, 35%, 40% of the training set, where the first 20% subset is randomly collected. At each acquisition cycle, after the detector is fully trained, 5% (*i.e.* 5,914) of the total images are acquired from the rest unlabeled set via active learning for annotation. We adopt the Average Precision (AP) over IoU thresholds ranging from 0.5 to 0.95 as the evaluation metric. **Pascal VOC** contains 20 object categories, consisting of the VOC 2007 trainval set, the VOC 2012 trainval set and the VOC 2007 test set. By following the settings [36], we combine the trainval sets with 16,511 images as unlabeled images, and randomly select 1,000 images as the initial labeled subset. The budget at each acquisition cycle is fixed to 1,000. The mean Average Precision (mAP) at the 0.5 IoU threshold is used as the evaluation metric.

Implementation Details. We set T_{enms} for instance-level redundancy removal in ENMS, T_{intra} for the intra-class diversity and T_{inter} for the inter-class diversity to 0.5, 0.7 and 0.3, respectively. α and β are set to 0.5 and 0.75, ensuring that at least 75% budgets are assigned to 50% of the classes (minority classes). We utilize Faster R-CNN [23] and RetinaNet [18] with ResNet-50 [12] and FPN [17] as the detection models on MS COCO. At all active learning cycles, we train the detector for 12 epochs with batch size 16. The learning rate is initialized as 0.02 and is reduced to 0.002 and 0.0002 after $2/3$ and $8/9$ of the maximal training epoch. On Pascal VOC, we adopt the settings in [36] using SSD [20] with VGG-16 [27] as the base detector.

Counterpart Methods. We make comparison to the state-of-the-art methods. The diversity-based ones include Core-set [25] and CDAL [1]. To adapt Core-set [25] to the detection task, we follow Learn Loss [36] to perform k-Center-Greedy over the image-level features. In regard of CDAL [1], we apply the reinforcement learning policy on the features after the softmax layer. The uncertainty-based methods contain Learn Loss [36] and MIAL [37]. We follow Learn Loss [36] to add a loss prediction module to simultaneously predict the classification and regression losses. The loss prediction module is trained by comparing image pairs, which empirically performs better than the mean square error [36]. Since loss prediction can affect detector training, we separately conduct active acquisition and detector retraining for fair comparison.

5.2. Experimental Results

On MS COCO. The comparison results on MS COCO are summarized in Fig. 3 (a). The detector built with the complete (100%) training set achieves 36.8% AP according to the open-source implementation*, which can be treated as an approximated upper bound. As demonstrated, our method consistently reaches the best performance in all active learning cycles, showing the superiority of the proposed acquisition strategy. In the last cycle with 40% annotated images, our method achieves 32.87% AP with an

*https://github.com/facebookresearch/maskrcnn-benchmark

Method	Entropy	ENMS	DivProto	Annotated Percentage				
				20%	25%	30%	35%	40%
Random				27.57±0.18	28.97±0.12	30.07±0.24	30.99±0.12	31.62±0.29
Ours	✓			27.57±0.18	29.38±0.13	30.61±0.12	31.47±0.17	32.36±0.07
	✓	✓		27.57±0.18	29.76±0.16	30.82±0.23	31.79±0.15	32.56±0.09
	✓		✓	27.57±0.18	29.73±0.16	30.64±0.11	31.86±0.09	32.53±0.14
	✓	✓	✓	27.57±0.18	**29.78**±0.06	**30.90**±0.14	**31.99**±0.05	**32.87**±0.04

Table 1. AP (%) by using different components of the proposed method with Faster R-CNN (ResNet-50) on MS COCO. "Random" refers to uniform acquisition. With various active acquisition strategies applied, the results are reported with standard deviation over 5 trials.

α	β	AP	AP_{50}	AP_{75}
0.50	0.25	30.68	52.48	31.93
0.50	0.50	30.74	52.97	31.86
0.50	0.75	**30.90**	53.08	32.01
0.50	1.00	30.79	53.00	32.01
0.25	0.75	30.71	52.90	31.63
0.75	0.75	30.58	52.62	32.15

Table 2. Results at the 30% cycle using Faster R-CNN (with the ResNet-50 backbone) on MS COCO with various α and β. AP_{50} (%) and AP_{75} (%) refer to AP at the IoU thresholds 0.5 and 0.75, respectively.

Method	Annotated Percentage				
	20%	25%	30%	35%	40%
Random	29.38	31.03	32.10	33.13	33.58
Entropy	29.38	31.48	32.53	33.98	34.12
Ours	29.38	**31.79**	**32.95**	**34.14**	**34.89**

Table 3. AP (%) using Faster R-CNN (ResNet-101) on MS COCO.

increase of 1.25%, compared to the uniform random sampling. The uncertainty based methods, *i.e.* Learn Loss [36] and MIAL [37], deliver almost the same performance as the basic entropy. The diversity-based ones, *i.e.* Core-set [19] and CDAL [1], perform poorly, since they utilize holistic features after spatial pooling without aggregating instance-level information.

Additionally, we report the results with small budgets (no more than 10%) following MIAL [37]. As in Fig. 3 (b), our method still outperforms the counterpart methods by a large margin in those settings. Though the detection performance under such small budgets does not meet the level of real-world applications, the remarkable gains compared with MIAL [37] demonstrate its effectiveness.

On Pascal VOC. We follow the same settings and open-source implementation[†] as reported in previous studies [1, 36, 37], whose result is 77.43% mAP with all (100%) training images. As shown in Fig. 3 (b), our method achieves better results than the other counterparts among the $4k$ to $10k$ cycles. Note that our method reaches a 73.68% mAP by using only 7k images. As a contrast, CDAL [1] and MIAL [37] need 8k and Learn Loss [36] needs 10k, showing the advantage of our method in regard of saving annotations. Indeed, our method does not perform better at the $2k$ and $3k$ cycles. It happens since the detectors are limited due to insufficient training at the initial cycles and cannot distinguish the difference between uncertain queries. This motivates us that the uncertainty and diversity should have varying weights during different active learning periods, but we do not go further here in case of increasing the complexity. MIAL [37] achieves the best performance when using 1k, 2k and 3k images. It should be noted that MIAL performs semi-supervised learning with unlabeled images, which is unfair for comparison, especially when labeled images are far fewer than the unlabeled. On the contrary, our method focuses only on active learning and the semi-supervised part is not introduced temporarily.

5.3. Ablation Study

On ENMS and DivProto. As shown in Table 1, the baseline entropy-based methods (Faster R-CNN with ResNet-50) separately applying ENMS or DivProto outperform that using uniform sampling and entropy only, showing the superiority by ensuring the diversity at the instance level. A combination of both the modules further boosts the overall accuracy, indicating that the intra-image diversity and inter-image diversity provide complementary diversity constraints for performance improvement.

On hyper-parameters in DivProto. α and β control the instance-level balance of classes, which is important to the inter-class diversity. We study the effect of these two hyper-parameters at the 30% cycle on MS COCO, where the same Faster R-CNN detector is used. As summarized in Table 2, our method achieves the best performance when $\alpha = 0.5$ and $\beta = 0.75$. **On different backbones.** To evaluate the effect of backbones on the detection accuracy, we report the performance on Faster R-CNN by using ResNet-101. As shown in Table 3, ResNet-101 can generally improve the performance, since it is a stronger backbone compared to ResNet-50. Our active acquisition strategy still consistently outperforms the random uniform sampling and basic

[†] https://github.com/amdegroot/ssd.pytorch

Method	Inference (s)	Acquisition (s)	Full (s)
CDAL [1]	5,599	2,798	8,397
UB	1,666	9.95×10^7	9.96×10^7
Ours	1,702	1,015	2,717

Table 4. Comparison of time cost for active learning on MS COCO at the 20% cycle. "Inference" refers to prediction on unlabeled images and "Acquisition" refers to image selection. UB denotes the upper bound of raw instance-level diversity computation.

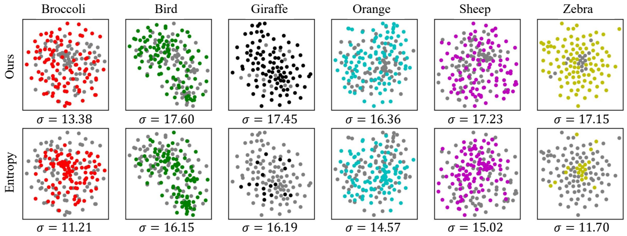


Figure 4. Visualization of the prototypes on MS COCO for 6 classes via t-SNE. Gray points indicate the unselected prototypes. The top and bottom rows are the results by using our method and the Entropy baseline.

entropy, showing its effectiveness with various backbones.

5.4. Analysis

Computational complexity. We introduce the diversity into the uncertainty-based solution, and the computational complexity and time cost grow accordingly. Thanks to the design ensuring the diversity at the instance level, we reduce the massive computations by converting them into all the three steps of active learning for object detection and thus avoid remarkable complexity increase. The time cost of the diversity-based methods are reported in Table 4, evaluated on a server with 8 NVIDIA 1080Ti GPUs. As shown in Table 4, the time cost is unacceptable when comparing each predicted instance pairs (UB), where exponential time increase occurs as each image usually contains multiple objects. By contrast, our proposed method implements the instance-level diversity constraints through the progressive framework and significantly reduces the time cost. Besides, our method spends less time than CDAL [1], since it requires REINFORCE based model training.

Visualization of prototypes. We qualitatively evaluate our method on enhancing the diversity in the presence of prototypes. We choose the basic entropy as the baseline and visualize the prototypes of six categories from MS COCO via t-SNE [31]. We also report the standard deviation σ accordingly. As illustrated in Fig. 4, the prototypes obtained by DivProto are more representative to the whole unlabeled dataset. Besides, these prototypes are more diverse according to the standard deviation.

Discussion on inter-class diversity. The active acquisition subsets of MS COCO can be used to make more evaluation on our method. For instance, as shown in Fig. 5, we calculate the standard deviations of instance amounts for 80 classes from the subsets, to analyze the inter-class diversity. As illustrated, both the proposed ENMS and DivProto modules decrease the standard deviation, indicating that they perform better in selecting class-balanced subsets of images compared to the basic entropy. A combination of ENMS and DivProto further improves the overall performance, confirming that our method improves the inter-class diversity and helps to construct a balanced subset for stronger detectors.

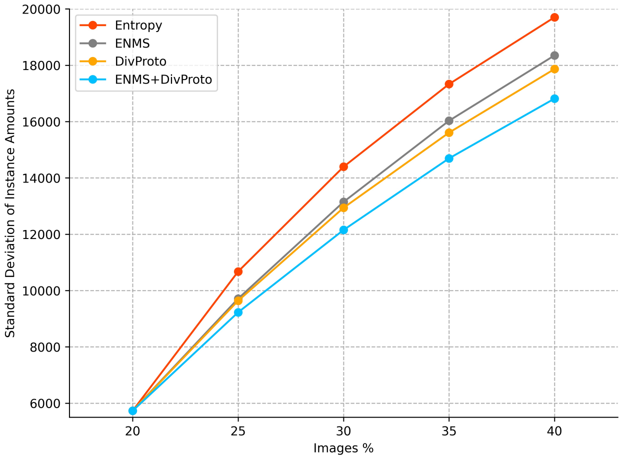


Figure 5. Curves in terms of the standard deviation of instance amounts for 80 classes on MS COCO. Statistics are made on the active acquisition subsets.

6. Conclusion

In this paper, we propose a novel hybrid active learning method for object detection, which combines the instance-level uncertainty and diversity in a bottom-up manner. ENMS is presented to estimate the instance-level uncertainty for a single image, while DivProto is developed to enhance both the intra-class and inter-class diversities by employing the entropy-based class-specific prototypes. Experimental results achieved on MS COCO and Pascal VOC show that our method outperforms the state of the arts.

Acknowledgment

This work is partly supported by the National Natural Science Foundation of China (No. 62022011), the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2021ZX-04), and the Fundamental Research Funds for the Central Universities.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153, 2020. 1, 2, 3, 5, 6, 7, 8

[2] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. 1, 2

[3] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 1, 2

[4] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 181–190, 2019. 3

[5] 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, 2020. 3

[6] Ehsan Elhamifar, Guillermo Sapiro, Allen Y. Yang, and S. Shankar Sastry. A convex optimization framework for active learning. In *IEEE International Conference on Computer Vision*, pages 209–216, 2013. 1, 2

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. 2, 6

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010. 2, 6

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192, 2017. 1, 2

[10] Ross B. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 3

[11] Yuhong Guo. Active instance sampling via matrix partition. In *Advances in Neural Information Processing Systems*, pages 802–810, 2010. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 3, 6

[13] Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *AAAI Conference on Artificial Intelligences*, pages 2659–2665, 2015. 2, 5

[14] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(10):1936–1949, 2014. 2

[15] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. 1, 2

[16] 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, 2018. 3

[17] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017. 6

[18] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2999–3007, 2017. 3, 6

[19] Tsung-Yi Lin, Michael Maire, Serge J. 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, 2014. 2, 6, 7

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *European Conference on Computer Vision*, 2016. 3, 4, 6

[21] Yingying Liu, Yang Wang, and Arcot Sowmya. Batch mode active learning for object detection based on maximum mean discrepancy. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 1–7, 2015. 3

[22] Hieu Tat Nguyen and Arnold W. M. Smeulders. Active learning using pre-clustering. In *International Conference on Machine Learning*, 2004. 1, 2

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 1, 3, 4, 6

[24] Soumya Roy, Asim Unmesh, and Vinay P. Namboodiri. Deep active learning for object detection. In *British Machine Vision Conference*, page 91, 2018. 2, 3

[25] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 6

[26] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. 3

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1, 3, 6

[28] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *IEEE International Conference on Computer Vision*, pages 5971–5980, 2019. 1, 2, 6

[29] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 2, 5

[30] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*, pages 9626–9635, 2019. 1, 3

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

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

[33] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2020. 2, 5

[34] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 399–407, 2017. 2

[35] Tianxiang Yin, Ningzhong Liu, and Han Sun. Self-paced active learning for deep cnns via effective loss function. *Neurocomputing*, 424:1–8, 2021. 1, 2, 5

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

[37] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 6, 7

[38] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8753–8762, 2020. 1, 2

[39] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13763–13772, 2020. 5