Active Learning for Single-Stage Object Detection in UAV Images

Asma Yamani[1], Albandari Alyami[1], Hamzah Luqman[1,2], Bernard Ghanem[3], and Silvio Giancola[3]
[1]Information and Computer Science Department, KFUPM
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, KFUPM
{g201906630,g201901850 , hluqman}@kfupm.edu.sa
[3]Image and Video Understanding Laboratory (IVUL), KAUST
{silvio.giancola, bernard.ghanem}@kaust.edu.sa,

Abstract

Unmanned aerial vehicles (UAVs) are widely used for image acquisition in various applications, and object detection is a crucial task for UAV imagery analysis. However, training accurate object detectors requires a large amount of annotated data, which can be expensive and time-consuming. To address this issue, we propose an active learning framework for single-stage object detectors in UAV images. First, we introduce Diverse Uncertainty Aggregation (DUA), a novel uncertainty aggregation method that aims to select images with a more diverse variety of object classes with high uncertainties. Second, we address the problem of class imbalance by adjusting the uncertainty calculation based on the performance of each class. Third, we illustrate how reducing the number of images for labeling does not necessarily lead to a lower labeling cost. Evaluation of our approach on a common UAV dataset shows that we can perform similarly (within 0.02 0.5mAP) to using the whole dataset while using only 25% of the images and 32% of the labeled objects. It also outperforms Random Selection and some other aggregation methods. Evaluation on VOC2012 show also consistent results utilizing only 25% of the labeling cost to reach a performance within 0.1 0.5mAP of using the whole dataset. Our results suggest that our proposed active learning framework can effectively reduce the annotation cost while improving the performance of single-stage object detectors in UAV image settings. The code is available on: https://github.com/asmayamani/DUA

1. Introduction

Unmanned aerial vehicles (UAVs) have revolutionized the way of collecting images for various applications, from monitoring crop growth in agriculture [15] to disaster management [7] and surveillance [2]. However, despite the great potential of UAV imagery, the diversity of objects and scenes captured in UAV images poses a significant annotation challenge. Annotating these images can be a costly and time-consuming task, hindering the use of UAV imagery for several practical applications.

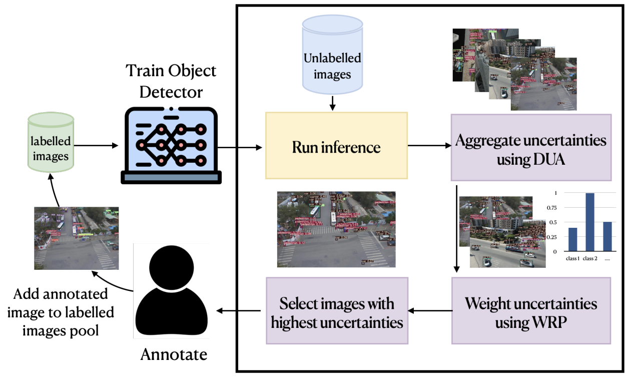

Train Object Detector
labelled images
Unlabelled images
Run inference
Aggregate uncertainties using DUA
Weight uncertainties using WRP
Select images with highest uncertainties
Annotate
Add annotated image to labelled images pool
1
0.75
0.5
0.25
0
class1 class2 ...

Figure 1. **Active Learning for Object Detection.** We train an object detector on a set of labeled images. Then, we select the most informative images from a set of unlabeled images through our new DUA active sampler and WRP weighting mechanism. An oracle annotates the unlabeled images used to re-train the object detector.

To circumvent these limitations, active learning has emerged as a promising solution to reduce the annotation cost by selecting the most informative and diverse samples to annotate next. Active learning algorithms aim to select a small subset of unlabeled data that experts can annotate and then add to the training set to improve the model's performance. Active learning methods have been successfully applied to various computer vision tasks [21], including object detection [4], segmentation [26], and classification [16]. However, the existing active learning methods for object detection in UAV image settings mainly focus on two-stage detectors, where the detection is divided into the regional proposal step and the classification step. These detectors are more suitable for active learning than single-stage detectors, where the bounding box regression and object classification are done simultaneously. Moreover, most of these methods do not address the problem of class imbalance, which is prevalent in UAV imagery due to the high variation in object sizes and frequencies. Two-stage frameworks divide the detection process into the region proposal and the classification stage. These models first propose several object candidates, known as regions of interest (RoI), using reference boxes (anchors). In the second step, the proposals are classified, and their localization is refined.

While accelerating improvement is happening in single-stage object detectors, much of the current work on active learning for object detection focuses on two-stage detectors [1,8,12,17,19,23,28], or requires training auxiliary models [27]. In previous work [1,4,8,12,17,19,23,28], the budget is considered on the image level without considering the label level, so approaches favoring highly dense images are usually selected. However, they lead to a higher cost per image [4]. On the application level, object detection for images shot from a UAV holds special challenges to other object detection models. These challenges include small object inference, background clutter, and wide viewpoint [5]. In addition, annotating only one view of the object may not be sufficient to train a reliable object detection model since drone images are captured from different angles.

This paper proposes an active learning framework for single-stage object detection in UAV images. Our approach aims to reduce the annotation cost while maintaining high accuracy by selecting the most informative and diverse samples for the annotation process. We introduce a novel uncertainty aggregation method called Diversity in Uncertainty Aggregation (DUA) approach that sums each class's average uncertainty per image. Additionally, we propose different weighting methods to address the class imbalance issue. Our approach uses significantly fewer labeled samples while outperforming several baseline methods, demonstrating the effectiveness of our proposed method. Figure 1 illustrates the main ideas of our approach.

Contributions. We summarize our contributions as follows: **(i)** We propose a DUA technique to query images for active learning to aggregate uncertainties while ensuring that images that carry information on various classes are selected. **(ii)** We propose two methods for addressing the problem of class imbalance by adjusting the uncertainty calculation based on the performance of each class. **(iii)** Provide analysis on the labeling cost at different aggregation and weighting methods.

2. Related Work

Early works on active learning relied on kernel methods by feeding image pairs through different kernels to capture image similarity as an input to a Support Vector Machine (SVM) [16]. With the use of deep learning in computer vision tasks comes the need to re-purpose appropriate acquisition functions to accommodate higher dimensional data [10]. Deep active learning can be categorized into *Uncertainty-based methods*, *Diversity-based methods*, and other *hybrid approaches*. The following sections will explain some of those methods and how previous works dealt with an imbalanced class problem.

Uncertainty-based Methods. Multiple studies focused on approximating aleatory and epistemic uncertainty, as the Convolution Neural Network (CNN) model's uncertainties are poorly captured. Early work [3, 10] in this field focused on the image classification task. In [10], Monte-Carlo (MC) dropout captures the epistemic uncertainty, keeping the dropout during testing with multiple runs to approximate the posterior. Different acquisition functions relying on Bayesian CNN uncertainty are explored. The study shows that using Bayesian CNN outperforms deterministic CNN in capturing uncertainty. It also shows that using variation ratio as the acquisition function performed better with distinct classes. In contrast, BALD [13], which maximizes the mutual information between predictions and model posterior, performs better when the difference between the classes is very narrow. Such performance is attributed to the fact that BALD avoids selecting noisy points and selects points that reduce the epistemic uncertainty. Inspired by Deep Ensembles [18], Beluch *et al.* [3] proposed averaging the Softmax vector of five CNN image classification models with different parameters to form an ensemble. The images with the highest uncertainties are queried by measuring the predictive variance between the vectors. This approach outperformed MC-dropout and the single CNNs regardless of the acquisition functions. However, this method is computationally expensive for large datasets.

One of the earliest works related to active learning on object detection is [4]. It studied the different aggregation methods of uncertainties using the sum, average, and maximum uncertainties. The reported results show that the sum aggregation methods perform better mean average precision (mAP) and area under the mean squared error learning curve (AULC). To suppress high uncertainties of noisy negative instances from the background, Yuan *et al.* [29] leveraged the discrepancy of two adversarial instance classifiers to learn each object's uncertainty. The image uncertainty was estimated by treating the image as a bag of instances and utilizing a classifier to estimate the labels. The instance uncertainty scores were iteratively re-weighted to minimize the image classification loss. Ensemble uncertainty-based methods lowered the labeling cost significantly in models built for autonomous driving [8]. More recent studies attempted to capture different uncertainties in two-stage detectors. Choi *et al.* [6] proposed mixture density networks to learn a probabilistic distribution of both the localization and classification to estimate the aleatoric and epistemic uncertainty in a single forward pass of a single model. Yu *et al*. [28] proposed a Consistency-based Active Learning approach for object Detection (CALD), which integrates the box regression uncertainties and classification of uncertainties of two-stage detectors with a single metric. It also focuses on the informative local region in an image rather than the whole image to better estimate the image uncertainties. CALD did outperform [1, 22] based on a recent survey [9].

Diversity-based Methods. This approach attempts to overcome a limitation in uncertainty-based methods in which highly correlated images are selected. Therefore, the images are clustered in a diversity-based approach, and a representative image is selected. Sener *et al*. [24] is one of the earliest works that recognized the limitation of selecting highly correlated images. To mitigate this issue, the study formulates the active learning problem as a core-set problem in which the algorithm selects the optimum cover set for random batches of images. It then converts this problem to the K-Center problem to choose the centers of the sets. By this conversion, the optimum cover set is the image that minimizes the L2-norms between the representation in the last fully connected layer in a CNN and the representation of the rest of the images in the set. The core-set Diversity-based approach showed improvement of 6% across subsets in the satellite images study [11].

Other Methods. Focusing on binary class detection, Aghdam *et al*. [1] calculates the posterior probability for each pixel and aggregates pixel-level scores per image and thresholds the distance between the representation of the selected images. This is based on the hypothesis that patches in different images will have similar prediction probability distribution if they have been seen adequately during training, ensuring the diversity of the selection. A task-agnostic approach is proposed by Yoo *et al*. [27] by introducing a loss prediction module. This component is a small parametric module to a target neural network to learn to predict the loss. It then predicts the loss over the unlabeled dataset to select the images with higher loss.

Dealing with Imbalanced Classes. One challenge that introduces biases in the object detection model is the imbalances of objects' classes within the training data. To tackle this issue, Brust *et al*. [4] weighted the uncertainties by the object's presence in the training data of the previous iterations. Other studies used the ratio of the selected labels [23] or the loss of the background and the other objects' classes [19] as weights for the class with a weighted cross-entropy loss function. A weighting filter tailored for object detection is proposed by Huang *et al*. [14], which calculates the frequency domain information of images and removes similar ones in selected data.

This work proposes a hybrid method that aggregates uncertainties while ensuring the object diversity of the selected images. It applies to single-stage object detection models and overcomes limitations of summing uncertainties that favor images with a high population of the over-represented class. The proposed approach also overcomes the limitations of averaging uncertainties, selecting images with few classes and high background noise. The proposed approach also deals with the class imbalance issue as it adjusts the ranking of the images chosen by weighting the uncertainties obtained by the class performance. This adjustment aids in prioritizing objects with lower AP in a validation subset due to their rareness, limited intra-class variance representation, or low inter-class variance.

3. Methodology

Figure 1 shows a high-level flow of the proposed active learning framework. The system starts with an object detection (OD) model trained on a small subset of labeled images. Then, it runs inference on the set of unlabeled images. The uncertainties of the detected objects are aggregated using Diversity in Uncertainty Aggregation (DUA), and then weighted using Weighting by Random Performance (WRP). The images are ranked based on the weighted uncertainty and, subsequently, based on the budget, are annotated. Finally, the model is re-trained using the updated subset of annotated images. You Only Look Once (YOLO) [20] will be used to demonstrate how our querying approach can be applied to single-stage object detectors. We will discuss the general details of YOLO and how it calculates the confidence scores. Then, we will discuss how to calculate and aggregate the uncertainties. Finally, we discuss how we weigh the aggregated uncertainties to account for class imbalance.

3.1. Backbone Model

This work uses YOLOv7 as a backbone for the proposed approach. YOLO is one of the fastest and most highly accurate real-time object detection techniques in the computer vision field [25]. YOLO is a single-stage object detector that formulates object detection as a regression problem [20]. The model outputs the bounding box coordinates of the detected objects with their class probabilities. YOLO works by dividing the image into an $S \times S$ grid, in which each cell is responsible for predicting the output of B bounding boxes after extracting the features from the whole image. The features are extracted through multiple layers, and the last layer produces the output vector. This vector contains the coordinates and size of the bounding box, the probability of the bounding box containing an object ($confidence_{box}$), and C conditional class probabilities ($Pr(class_i|object)$). The $confidence_{box}$ is calculated during the training process as follows:

$$confidence_{box} = Pr(object).IoU \tag{1}$$

where *Pr(object)* is the probability of the box containing an object, and IoU is the intersection over the union between the bounding box and the ground truth.

3.2. Calculating Uncertainties

As mentioned, YOLO formulates object detection into a regression problem. At testing, the model outputs the confidence scores encoding the probability of that class appearing in a bounding box and how well the predicted box fits the object. This is calculated by multiplying the conditional class probabilities, $Pr(class_i|object)$, and the individual box confidence, $confidence_{box}$, predictions learned during training [20]. In this work, the uncertainty of an object is calculated as.

$$uncertainty_{object} = 1 - confidence_{class} \tag{2}$$

The class confidence is computed as follows:

$$confidence_{class} = confidence_{box} \times Pr(class_i|object) \tag{3}$$

3.3. Diversity in Uncertainty Aggregation (DUA)

We propose DUA, a method that encourages class diversity when aggregating the uncertainties. For each unlabeled image in the iteration, we average the $uncertainty_{object}$ per class to obtain $uncertainty_{class}$. Then, we perform summation over all $uncertainty_{class}$. By doing so, images with a larger variety of classes with high average uncertainty per class are obtained. This aims to select images that have distributed contributions to the improvement of AP across classes and limit over-represented classes from overpopulating the selection at each iteration. Figure 2 illustrates how summing uncertainties (Sum) and averaging them *Avg* from [4] compares to our *DUA* approach when selecting between different images after running inference. The figure contains the top selected picture across approaches at the 1^{st} iteration. The Sum approach selects an image with over 300 detected objects belonging to 4 classes, and 200 of these objects belong to the "car" class. The *Average* approach selects an image with only 10 objects across 2 classes, each object with high uncertainty. On the other hand, DUA, with and without weighting, selects images with 7 classes and less than 140 objects with varying uncertainties.

3.4. Weighting by Performance

Weighting the classes by their presence in the training data (WTC), as in [4], may suppress the selection of classes with many labels in the training subset yet perform poorly. This can be attributed to the low inter-class variance that causes high confusion or the low representation of the high intra-class variance. To mitigate such issues, we propose Weighing by Random subset Performance (WRP). In this proposed approach, the uncertainties of different classes at an image are weighted before ranking based on the current model performance on a random validation subset sampled and labeled from the unlabeled training data per the formula:

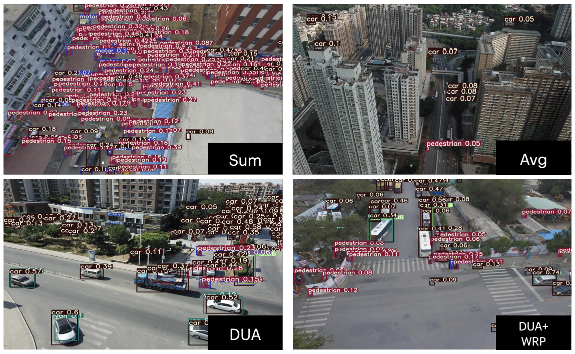


Figure 2. **Most Informative Images at First Iteration.** The top image was selected in the first iteration using different aggregation approaches.

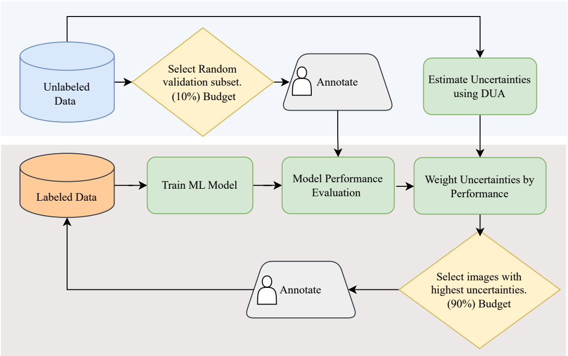

Unlabeled Data
Select Random validation subset. (10%) Budget
Annotate
Estimate Uncertainties using DUA
Labeled Data
Train ML Model
Model Performance Evaluation
Weight Uncertainties by Performance
Select images with highest uncertainties. (90%) Budget
Annotate

Figure 3. Weighting by random performance (WRP) approach.

$$W = MinMaxScaler(1 - AP_{val}(C)), \tag{4}$$

where W is the weight vector for all classes, and C is the class vector. The flow of the algorithm is illustrated in Figure 3.

Although weighting randomly from the unlabeled data is the optimal choice to avoid overfitting, sacrificing 10% of the budget for Random labeling to test on could be costly. Another approach would be to weight based on the training performance (WTP). For this, we use the AP of the model on the training subset, and we perform the same scaling as in Equation 4. With this, all the labeling budget goes to labeling images selected by the approach. Illustrated in Figure 2, weighting the uncertainties by performance minimizes the selection of images of objects on the high-performing side. DUA without weighting pulls images with a higher number of cars, 114, whereas when weighting by performance using WRP only 35 cars are retrieved.

4. Experiments

4.1. Experimental Setup

Datasets. We examined and evaluated our proposed approach on the VisDrone2019 dataset [5]. This dataset was collected by the AISKYEYE team. The dataset contains 288 videos with $261,908$ frames and $10,209$ static images. These images and videos are collected using various drone platforms in different scenarios and cover different locations, environments, objects, and densities. The dataset is prepared for object detection in drone images and videos, single and multi-object tracking, and crowd counting. The dataset consists of $6,471$ images for training, of which $4,279$ are annotated. The sizes for the validation, test-dev, and testing subsets are 548, $1,610$, and $1,580$, respectively. VisDrone2019 dataset contains 10 classes: person, car, bus, truck, bicycle, motor, pedestrian, van, awning-tricycle, and tricycle. Table 1 shows the distribution of object instances across classes. We also evaluate our approach on VOC2012 dataset. VOC2012 consists of $17,125$ images for training and validation. We use 90% as the training subset where we perform the active learning proposed approaches and report the results of the remaining 10% as the validation subset.

Fine-tuning and Incremental Learning Setup. For the experiments the VisDrone dataset, we initialized the process by selecting 100 random points (images). At each iteration, we included an extra 100 instance selected by the active learning approach. In the case of the first experiment related to the uncertainty aggregation approach, the active learning approach is applied on 500 random instances as preliminary experiments showed that applying active learning approaches to a large random sample of the training data yields results through increased diversity. 500 was chosen as it is sufficiently large for the size dataset and the queried images per iteration. However, the whole dataset is considered in the second experiment related to the different weighting approaches to minimize the confounding factor of randomization. When evaluating the proposed approaches on VOC2012 we initialized the process by selecting 500 random points due to the larger training set and number of classes. We then add 500 extra images based on the active learning approach performed on the large pool of 5000 images randomly selected from the training data.

Baseline. We employ several baselines to compare the performance of our proposed method when it comes to the VisDrone dataset. The first is to use *Random* sampling of new images for each iteration. The second baseline is to aggregate uncertainties by summing them (*Sum*), and the third is to aggregate uncertainties by averaging them (*Avg*) [4]. We also compare with training on the whole annotated training dataset (*Whole*). As for treating class imbalance, the baseline considered is weighting the classes based on their presence in the training data (*WTP*) [4]. When evaluating on VOC2012, we only consider *Whole*, *Random*, *Sum* as baselines.

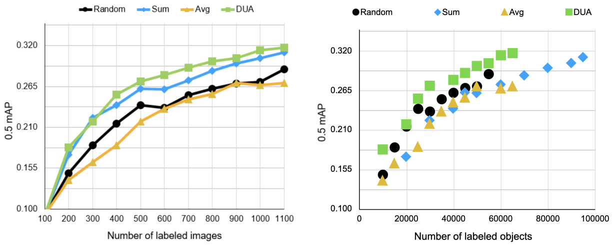


Figure 4. **Active Learning Curves.** Results of comparing different aggregation methods mAP@0.5 and the number of images (Left) and objects (Right) labeled per iteration.

4.2. Comparison with Baseline

4.2.1 Uncertainty Aggregation

Table 2 shows the MAP results of the proposed approach for aggregating uncertainties compared to baselines on the validation set after the 10^{th} iteration of the training. *DUA* has a higher average mAP@0.5 across classes than the baseline querying methods, even compared to training on *Whole*. It also has a higher average mAP@0.5 per class than *Whole* for 7 out of 10 classes reducing the object labeling cost to a third. We can also notice that the number of objects of the 1100 queried images is significantly less for *DUA* at $95k$ and $72k$ for *Sum* and *DUA*, respectively. Figure 4 captures this difference further and shows that *Sum* and *Avg* have similar accuracy when considering a similar number of labels. At a similar labeling cost of $52k$ for Random, *Sum*, and *Avg*, the mAP@0.5 is at 0.271, 0.261, and 0.27 for the approaches, respectively. In contrast, at also $52k$ object labeling cost, *DUA* achieves 0.299 mAP@0.5. It is worth noting that this labeling cost accumulates at different iterations, implying that some approaches query images with higher object densities. Random pulls $5k$ per 100 images (per iteration)), reaching 50k labels at the 9^{th} iteration. On the other hand, *Sum* accumulates this labeling cost faster and reaches 50k at the 5^{th} iteration. For *Avg* and *DUA*, this accumulation occurs at the 8^{th} and 7^{th} iterations, respectively. As for reaching a similar performance of about 0.27, as an example, *Sum* reaches this performance at the 6^{th} iteration with $60k$ labels. In contrast, *DUA* reaches this performance at the 4^{th} iteration with around $30k$ labels, highlighting the enhanced selection process of *DUA*.

4.2.2 Weighting by Performance

From the results in Table 3, weighting improved the performance of underrepresented classes (tricycle, awning-tricycle, bus); however, with different behaviors. We can notice that WTC improved the performance of the "bus" class the most, being a large object under-represented class.

Table 1. **Number of object instances per class.** In **bold** are over-represented classes, in *italic* under-represented classes.

object class	all	pedestrian	people	bicycle	car	van	truck	tricycle	awning-tricycle	bus	motor
#instances	226189	**52762**	17093	6950	**95977**	16866	8652	*2877*	*1968*	*4148*	18896

Table 2. **Main Results Comparison.** Results of comparing the proposed approach for aggregating uncertainties concerning previous methods and baseline on the validation set.

Class	Whole	Random	*Sum*	*Avg*	*DUA*
Images	**4279**	**1100**	**1100**	**1100**	**1100**
Objects	**226k**	**57.3k**	**95.8k**	**63.4k**	**72.7k**
All	0.313	0.288	0.311	0.27	**0.317**
Pedestrian	0.341	0.330	**0.351**	0.325	0.343
People	0.311	0.305	0.332	0.305	**0.333**
Bicycle	0.066	0.042	0.061	0.026	**0.066**
Car	0.726	0.716	0.727	0.684	**0.730**
Van	0.315	0.279	**0.311**	0.234	0.309
Truck	0.284	0.262	0.253	0.239	**0.294**
Tricycle	0.174	0.122	**0.191**	0.133	0.188
Awning-tricycle	0.086	0.066	0.080	0.070	**0.092**
Bus	0.438	0.415	0.411	0.336	**0.418**
Motor	0.384	0.346	0.388	0.353	**0.397**

Table 3. **Analysis of Different Active Learning Weighting.** Performance evaluation results concerning the proposed methods and baseline on the validation set in terms of mAP@0.5.

	DUA	*DUA*+WTC	*DUA*+WRP	*DUA*+WTP
All	0.311	**0.317**	0.308	**0.316**
Pedestrian	0.340	**0.347**	**0.348**	0.346
People	0.326	0.325	**0.327**	**0.34**
Bicycle	**0.063**	0.0588	0.0509	0.0539
Car	**0.728**	0.724	**0.726**	0.723
Van	**0.314**	0.309	0.291	0.306
Truck	**0.29**	**0.295**	0.271	**0.29**
Tricycle	0.186	0.182	**0.187**	**0.189**
Awning-tricycle	**0.0911**	0.0871	0.0897	**0.0956**
Bus	0.39	**0.453**	0.398	0.421
Motor	0.381	0.387	**0.394**	**0.399**

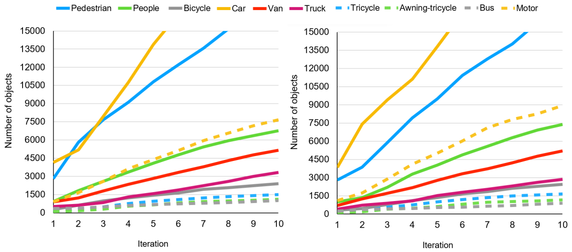


Figure 5. Cumulative object count per iteration for the different weighting approaches. Left: *DUA*+WTC, right: *DUA*+WRP

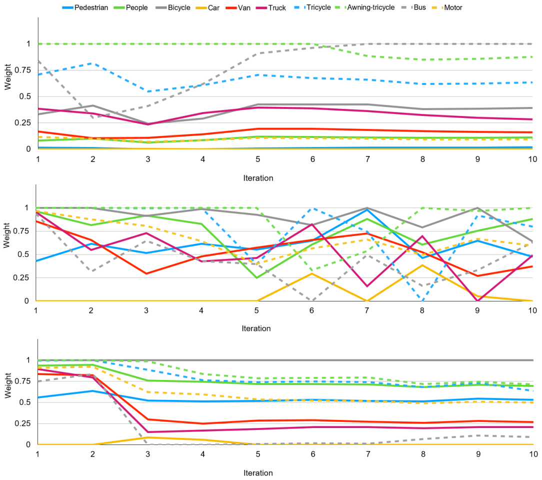


Figure 6. Weights change per iteration for the three weighting approaches. (Top) *DUA*+WTC, (Middle) *DUA*+WRP, (Bottom) *DUA*+WTP.

Figure 5 illustrates that WTC did query an equal amount of "bus" objects compared to "awning-tricycle" objects despite its reasonable performance. Weighting by performance approaches overcomes this issue as the weighting is done on how a certain class performs rather than just the presence. Illustrated in Figure 5, WRP queries "bus" objects at a slower rate than "tricycle" and "awning-tricycle". In addition, the "people" class is being queried at a high rate despite not being an underrepresented class due to the confusion with the "pedestrian" class. As for over-represented classes such as "car", the increase between the 5^{th} and 10^{th} class is 200% when using *DUA* only without weighting, whereas using the different weighting approaches, only 90% of this quantity of labels are queried leading to a less than 1% mAP@0.5 difference. This reduction comes from the weight for the "car" confidence score being zero across weighting mechanisms, so no image was queried because it had a car. The increase in the "car" object results from querying images to obtain the representation of other objects, and cars are coincidentally present.

Looking at the weights per iteration across approaches, in Figure 6, we can see that the weights calculated by the WRP approach, fluctuates even in later iterations. This could be due under-represented classes not appearing in the sample resulting in a weight of one. It also indicates a higher potential for growth as it is exploring diverse forms of the class. In contrast, WTC will continue to query objects with the same distribution as the previous queries do make a lasting impact. Another concern that arises with WTP is that once the model is trained, it could minimize further exploration of objects from the same class with a different view as long as the model achieves high performance on the training data.

Table 4. Performance evaluation results with respect to the proposed methods and baseline on the test-dev set in terms of mAP@0.5.

Class	**Whole**	**Random**	***DUA***	***DUA*+WTC**	***DUA*+WRP**	***DUA*+WTP**
All	0.273	0.211	0.243	**0.248**	**0.253**	0.246
Pedestrian	0.21	0.187	0.196	**0.215**	**0.208**	**0.208**
People	0.16	0.134	0.155	**0.165**	**0.161**	**0.161**
Bicycle	0.0702	0.0406	0.0532	0.0571	**0.0607**	**0.0669**
Car	0.664	0.634	**0.644**	0.638	**0.648**	0.639
Van	0.308	0.205	**0.264**	0.25	0.252	**0.254**
Truck	0.318	0.183	0.229	**0.242**	**0.275**	0.24
Tricycle	0.112	0.0783	**0.126**	**0.133**	0.119	0.115
Awning-tricycle	0.134	0.0555	**0.131**	0.105	**0.113**	0.109
Bus	0.511	0.423	0.425	**0.451**	**0.461**	0.436
Motor	0.243	0.173	**0.223**	0.218	**0.223**	**0.231**

Table 5. Performance evaluation on VOC2012

	Whole	**Random**	**Sum**	**DUA**	DUA+WRP
Labels at 5$^{th}iteration$	**32273**	**6326**	**10510**	**8336**	**8005**
0.5mAP (All classes)	0.807	0.692	0.718	0.714	0.708
Close labeling cost	-	-	**7398**	**7174**	**6732**
0.5mAP (All classes)	-	-	0.643	0.698	0.695
0.5mAP of minimum across classes	-	-	0.261	0.349	0.424

4.2.3 Performance evaluation on VisDrone dataset

The performance evaluation results in terms of mAP@0.5 per class on the test-dev split of the dataset are present in Table 4. When evaluating the model built using the annotated training dataset (4279 images), it yields 0.273 mAP@0.5 on the dev-test set. Meanwhile, using one-fourth of the images with *DUA* acquisition function, the model reaches a 0.243 mAP@0.5, with only a 0.03 mAP@0.5 difference. Using the proposed weighting approaches (WRP and WTP) further narrows the difference. With *DUA*+WRP, the model reaches a 0.253 mAP@0.5. Evaluation of the model on the test-dev set also shows the effectiveness of *DUA*+WRP over Random sampling, as the model built using images queried by Random sampling scores 0.211 mAP@0.5 only. Visualization of the different approaches is in Figure 7.

4.2.4 Performance evaluation on VOC2012 dataset

The results of evaluating our proposed approach on the validation set of the VOC2012 dataset are shown in Table 5. As shown in the table, the performance is comparable to using the *Whole* dataset with only 20% of the training images (3000 labeled images) with all active learning approaches. The 0.5 mAP was less than using *Whole* by less than 0.1 0.5mAP. However, the main variance between the three active learning approaches is the number of labels within the images as *Sum* reaches that using 30% of the labels. In contrast, DUA and DUA+WRP use only 25% of the labeling cost in terms of objects. The results also show that when selecting iterations with close labeling cost per object, iteration 3 for *Sum* and 4 for DUA and DUA+WRP, DUA and DUA+WRP achieve higher performance than *Sum*. The results are further amplified when looking at the least 0.5mAP among the different classes where DUA+WRP had 161% higher 0.5mAP than *Sum* and 121% higher 0.5mAP than *DUA*. This confirms our results on VisDrone, where DUA+WRP is helping the least-performing classes the most.

Table 6. Results of varying the size of the random pool on which *DUA* is applied

Random size	***DUA* on 500**	instances	***DUA* on 1000**	instances	***DUA* on ∼4000**	Full Dataset
All	**0.303**	60188	**0.31**	62585	**0.3**	63429
Pedestrian	**0.33**	14719	**0.337**	14071	**0.33**	13903
People	**0.308**	5875	**0.324**	6234	**0.322**	6499
Bicycle	**0.0522**	2244	**0.0487**	2234	**0.0412**	2249
Car	**0.721**	21598	**0.723**	22694	**0.716**	23075
Van	**0.289**	3948	**0.307**	4133	**0.291**	4360
Truck	**0.277**	2355	**0.262**	2583	**0.271**	2762
Tricycle	**0.173**	1232	**0.19**	1535	**0.172**	1505
Awning-tricycle	**0.0792**	879	**0.0863**	1035	**0.0808**	1109
Bus	**0.434**	602	**0.432**	671	**0.407**	790
Motor	**0.37**	6736	**0.387**	7395	**0.373**	7177

4.3. Ablation Study

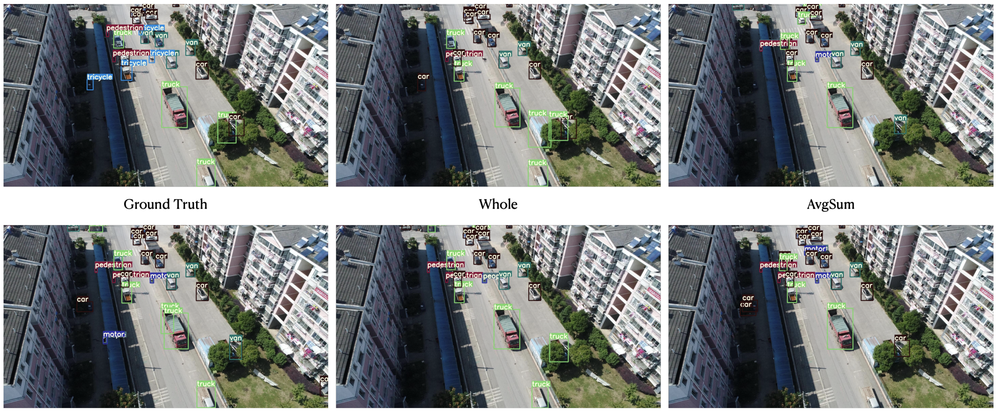
4.3.1 Incremental Learning

Four setups have been examined for selecting the incremental learning approach. The first setup is to continue learning from the weights of the previous iteration while including 20 random data points from the previous iteration and 80 data points from the queried data points for a 100 epoch per iteration. The second setup is to continue learning from the weights of the previous iteration while training on the old data points and 100 new selections for 20 epochs. The third setup is to continue learning from the weights of the previous iteration while training on the old data points and 100 new selections for 100 epochs. The last is to train from scratch for 100 epochs on the previous and new 100 points selection. Some of these setups ran for ten iterations, while the rest ran only for five iterations, as early iterations showed they were not optimal.

As the results show in Table 7, setups 1 and 2 do not provide steady fast growth. This could be due to the small-sized initialization and few numbers of epochs. The primary choice was between setup 3 and setup 4 as the difference between them at the 10th iteration is 0.018 mAP@0.5, with setup 4 leading. Figure 8 emphasizes the use of setup four and shows that when training from scratch, in later iterations, the model does continue to improve with a significant margin. Compared to when adopting continuous training, the improvement has high speed at the start and then plateaus, possibly due to overfitting the points from early iterations.

4.3.2 Sampling Training Dataset

Running the inference per iteration on the whole dataset to calculate uncertainties may be infeasible due to computation costs. In multiple studies, the active learning approach is applied over a large, randomly selected pool of images. In this experiment, we study the effect of varying the size



Ground Truth | Whole | AvgSum

WTC | WTP | WRP

Figure 7. Visualization of the predictions on the test-dev split

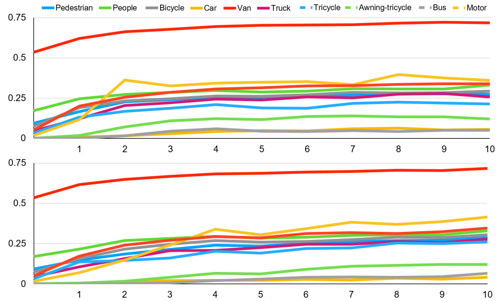


Figure 8. Improvement of the 3^{rd} (up) and 4^{th} (bottom) setups over 10 iterations .

Table 7. mAP@0.5 of the fifth iteration of Random Selection

	Setup			
Class	**1**	**2**	**3**	**4**
All	0.19	0.164	**0.258**	0.24
Pedestrian	0.246	0.24	**0.297**	0.296
People	0.187	0.153	0.262	**0.269**
Bicycle	0.00984	0.00982	**0.0426**	0.0233
Car	0.643	0.63	**0.696**	0.682
Van	0.165	0.136	**0.243**	0.208
Truck	0.157	0.144	**0.21**	0.202
Tricycle	0.0305	0.00696	**0.122**	0.0656
Awning-tricycle	0.00343	0.000474	**0.0596**	0.0215
Bus	0.248	0.119	**0.342**	0.34
Motor	0.212	0.204	**0.306**	0.294

of this randomly selected pool of images. We experiment with 500, 1000, and 2000 random instances and the entire dataset (4279 images). Only eight iterations per size are studied due to computational costs. Results in Table 6 show that, as opposed to what was anticipated, some randomness is better for the querying process. It did help in querying images with fewer labels while improving the mAP@0.5 for 1000 and 2000 randomly selected pools. This could be due to the *DUA* approach favoring ideas with a larger number of class objects, ensuring the selection of almost all images with the under-represented object. This sometimes jeopardized information that could be gained from images that don't necessarily have the under-represented object. Also, introducing some randomness reduces images with high uncertainty patterns in the same iteration. Despite the apparent benefits of applying active learning to a random pool of unlabeled data, it makes small differences in the approaches harder to evaluate. Therefore, we used the whole dataset when assessing the proposed weighting approaches.

5. Conclusion

This work studies active learning for single-shot object detection. It proposes the DUA method that sums average uncertainties across each class to query the images based on the confidence score incorporating the bounding box and classification uncertainties. When evaluated on the test-dev split of VisDrone, the proposed approach, DUA, achieves a comparable performance to using the whole annotated training data, with a 0.03 mAP difference, while using a fourth of the annotation cost. The improvement is emphasized in the case of the under-represented classes of the awning-tricycle and bicycle, where DUA did exceed the performance of using the whole dataset. This study also proposes weighting by performance with its two approaches to address the issue of class imbalances. It further minimizes the difference by using all the annotated images to 0.02 mAP. Evaluating the approach on VOC2012 using a validation subset indicates similar patterns. For future work, we consider evaluating the approach on other benchmark datasets and exploring the combination of the WTC and WRP approaches.

Acknowldgment

The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant JRC-AI-RFP-14.

References

[1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3672–3680, 2019. 2, 3

[2] Randal W Beard, Timothy W McLain, Derek B Nelson, Derek Kingston, and David Johanson. Decentralized cooperative aerial surveillance using fixed-wing miniature uavs. *Proceedings of the IEEE*, 94(7):1306–1324, 2006. 1

[3] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018. 2

[4] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv:1809.09875*, 2018. 1, 2, 3, 4, 5

[5] Yaru Cao, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, Steven Hoi, Qinghua Hu, and Ming Liu. Visdrone-det2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2847–2854, October 2021. 2, 5

[6] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021. 2

[7] Milan Erdelj, Enrico Natalizio, Kaushik R Chowdhury, and Ian F Akyildiz. Help from the sky: Leveraging uavs for disaster management. *IEEE Pervasive Computing*, 16(1):24–32, 2017. 1

[8] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 667–674. IEEE, 2019. 2

[9] Zhanpeng Feng, Shiliang Zhang, Rinyoichi Takezoe, Wenze Hu, Manmohan Chandraker, Li-Jia Li, Vijay K. Narayanan, and Xiaoyu Wang. Albench: A framework for evaluating active learning in object detection, 2022. 3

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

[11] Alex Goupilleau, Tugdual Ceillier, and Marie-Caroline Corbineau. Active learning for object detection in high-resolution satellite images. *arXiv preprint arXiv:2101.02480*, 2021. 3

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

[13] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011. 2

[14] Wei Huang, Shuzhou Sun, Xiao Lin, Dawei Zhang, and Lizhuang Ma. Deep active learning with weighting filter for object detection. *Displays*, page 102282, 2022. 3

[15] Kasper Johansen, Mitchell JL Morton, Yoann M Malbeteau, Bruno Aragon, Samir K Al-Mashharawi, Matteo G Ziliani, Yoseline Angel, Gabriele M Fiene, Sónia SC Negrão, Magdi AA Mousa, et al. Unmanned aerial vehicle-based phenotyping using morphometric and spectral analysis can quantify responses of wild tomato plants to salinity stress. *Frontiers in Plant Science*, 10:370, 2019. 1

[16] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pages 2372–2379. IEEE, 2009. 1, 2

[17] Benjamin Kellenberger, Diego Marcos, Sylvain Lobry, and Devis Tuia. Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep cnns and active learning. *CoRR*, abs/1907.07319, 2019. 2

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[19] Ying Li, Binbin Fan, Weiping Zhang, Weiping Ding, and Jianwei Yin. Deep active learning for object detection. *Information Sciences*, 579:418–433, 2021. 2, 3

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3, 4

[21] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 1

[22] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018. 3

[23] Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois Knoll. Advanced active learning strategies for object detection. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 871–876. IEEE, 2020. 2, 3

[24] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 3

[25] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 3

[26] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017. 1

[27] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 93–102, 2019. 2, 3

[28] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3960, 2022. 2, 3

[29] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 2