MUS-CDB: Mixed Uncertainty Sampling with Class Distribution Balancing for Active Annotation in Aerial Object Detection

Dong Liang[1], Jing-Wei Zhang[1], Ying-Peng Tang[1], Sheng-Jun Huang*

***Abstract*—Recent aerial object detection models rely on a large amount of labeled training data, which requires unaffordable manual labeling costs in large aerial scenes with dense objects. Active learning is effective in reducing the data labeling cost by selectively querying the informative and representative unlabelled samples. However, existing active learning methods are mainly with class-balanced setting and image-based querying for generic object detection tasks, which are less applicable to aerial object detection scenario due to the long-tailed class distribution and dense small objects in aerial scenes. In this paper, we propose a novel active learning method for cost-effective aerial object detection. Specifically, both object-level and image-level informativeness are considered in the object selection to refrain from redundant and myopic querying. Besides, an easy-to-use class-balancing criterion is incorporated to favor the minority objects to alleviate the long-tailed class distribution problem in model training. To fully utilize the queried information, we further devise a training loss to mine the latent knowledge in the undiscovered image regions. Extensive experiments are conducted on the DOTA-v1.0 and DOTA-v2.0 benchmarks to validate the effectiveness of the proposed method. The results show that it can save more than 75% of the labeling cost to reach the same performance compared to the baselines and state-of-the-art active object detection methods. Code is available at *https://github.com/ZJW700/MUS-CDB*.**

***Index Terms*—Active Learning, semi-supervised learning, object detection, aerial remote sensing image.**

I. Introduction

AERIAL object detection has received much attention in recent years, due to its important role in land and resources survey, geographic mapping, and other fields [1]. However, existing state-of-the-art aerial object detectors usually require a large amount of training data with bounding-box annotations for model training to reach the desired performance, which is typically expensive to obtain [2]. Active learning (AL) is a machine learning technique that selectively queries the informative unlabeled examples from the oracle for annotation to reduce the labeling cost. It has been successfully applied to the generic object detection problem for data-efficient learning [3]–[6]. However, existing generic active learning methods can hardly be applied to the remote sensing images. As shown in Figure 1, the objects in aerial images (i.e., remote sensing data) are usually small, blurred, and densely distributed in the complex background. Such characteristics are not well-considered by the existing methods, thus they may lead to unsatisfied performances. For these reasons, we strive to design an effective active learning method for aerial object detection that sufficiently takes into account the challenges of remote sensing images.

There are basically two aspects needed to be considered in active object detection, i.e., **query strategy** and **query type**. The former investigates the measurements of the informativeness of the data, and the latter designs an efficient manner to acquire knowledge from the oracle.

For the query strategy, most existing methods try to evaluate the uncertainty criteria of the unlabeled data. However, these common criteria neglect one of the most notable problems of aerial data – the severe class imbalance problem [7]. As a result, query by uncertainty may further intensify the imbalance problem and bring challenges to the model training. Since the uncertain patterns are more likely to come from uncertain classes (the classes with rare samples), class preference should be considered in the active selection for the aerial object detection task. However, implementing such a criterion can be highly challenging according to the active learning literature [8], since it is hard to accurately predict the class labels for the unlabeled images with only limited training data.

For the query type, existing solutions can be roughly divided into the following two categories, image-based as shown in Figure 2 (a) and object-based methods as shown in Figure 2 (b). Image-based methods estimate the uncertainty of the whole image and require the bounding-box annotation of all the objects from the image [9], [10]. Although they are adept at capturing the comprehensive contextual information of the image, these approaches suffer from inefficient and redundant labeling problems for aerial object detection. Because an aerial image usually has plenty of similar objects with information redundancy, annotating all of them may lead to a waste of labeling costs. Object-based sampling methods have recently been studied for object detection [4], [5]. They query a specific object (i.e., bounding-box) rather than the whole image for more fine-grained and cost-effective supervision, which are more suitable to the aerial data. However, these approaches also confront limitations. On the one hand, they ignore the contextual information of the objects, i.e., they only consider the uncertainty of the object, but neglect the spatial information and the semantic structure of the image. On the other hand,

All authors are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China.

1 These authors are contributed equally.

∗ Corresponding author: Sheng-Jun Huang (huangsj@nuaa.edu.cn).

Manuscript received December 3, 2022.

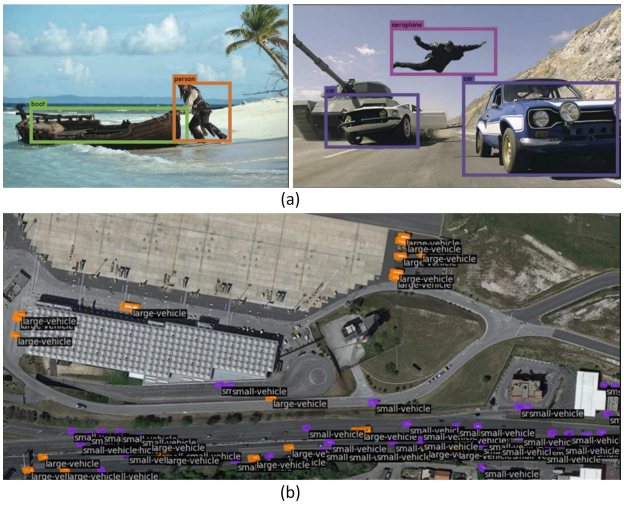

Fig. 1: Comparison of generic object detection and aerial object detection. (a) generic object detection: objects in general scenes are usually sparsely distributed and easy to identify. (b) aerial object detection: objects in aerial images (i.e., remote sensing data) are usually small, blurred, and densely distributed in the complex background.

they may introduce training noise. Because each image is partially annotated, i.e., only the supervision of certain objects can be obtained, the remaining unknown regions are usually treated as background, which can mislead the model due to the latent foregrounds.

To address the problems raised above, in this paper, we present a novel active learning method for aerial object detection, Mixed Uncertainty Sampling with Class Distribution Balancing (MUS-CDB). Specifically, for the query type, we propose an object-based mixed uncertainty sampling method to label the most informative and representative objects. Different from the existing works, our method addresses the limitations of both conventional object-based and image-based methods as analyzed above, i.e., the redundant and myopic information querying, by considering both object-level and image-level uncertain information. For the query strategy, we propose a hybrid selection criteria, which considers both uncertainty and class distribution balancing. The former identifies the objects that are most helpful to improve the performance of current detection model, while the latter tries to improve the class distribution balance of training samples to further enhance the model's capability on rare classes. Finally, in order to fully utilize the weakly supervised data, i.e., partially labeled images, we propose an effective training scheme associated with a loss function, which can effectively mine the latent knowledge in the unlabeled image regions that have not been queried as positive samples. Extensive experiments are conducted on DOTA-v1.0 [11] and DOTA-v2.0 [12], and the results show that the proposed method can significantly outperform the conventional image-based and object-based active learning methods for aerial object detection.

We summarize our contributions as the following:

1) We propose an object-based active learning method for aerial object detection, which considers the characteristics of remote sensing data. It incorporates both uncertainty and class balancing in active select informative and representative object samples.
2) We propose a training scheme associated with a loss function for object-based querying with partially labeled data. It sufficiently and robustly exploits the queried information from partial labels and can effectively improve the model's capability.
3) Extensive experiments on DOTA-v1.0 and DOTA-v2.0 validate the effectiveness and practicability of the proposed method, which significantly outperforms the conventional image-based and object-based active learning methods for aerial object detection.

The rest of this article is structured as follows. In Section II, We discuss the related work. In Section III, we introduce the proposed method in detail. The experimental results are presented in Section IV. Section V presents the conclusion.

II. RELATED WORK

A. Aerial Object Detection

Aerial object detection is an important task in computer vision [1]. Compared with the conventional object detection task, aerial object detection is more challenging for the following reasons. Firstly, the objects in remote sensing images have various orientations and aspect ratios. The detector often needs to predict the angle of the bounding box in order to locate the object accurately. Secondly, the objects in remote sensing images are usually small, and their visual features are easily affected by complex imaging processes, noise, and occlusion. Many works have been devoted to solving these specific problems of aerial object detection.

To mine the direction information, some methods use multiple anchors with different angles, scales, and aspect ratios for regression tasks [13], [14]. However, these methods are usually computationally expensive. Another group of methods uses only horizontal anchors to detect rotated objects to improve efficiency. For example, Ding et al. [15] propose RoI Transformer to convert Horizontal RoIs (HRoIs) into RRoIs, which can effectively reduce the number of anchors. DRN [16] performs orientated object detection through dynamic feature selection and optimization. Wei et al. [17] propose a calibrated-guidance (CG) scheme to enhance channel communications. CSL [18] treats angle prediction as a classification task to avoid discontinuous boundary problems. ReDet [19] is committed to improving the model's feature representation by modifying the backbone to generate rotation-equivariant features, and employing RiRoI Align in the detection head to extract rotation-invariant features.

For the challenge of small objects, Qiu et al. [20] propose an adaptive aspect ratio network. It on the one hand assigns different weights to the feature maps, and on the other hand predicts proper ratio aspects for different objects. Li et al. [21] design a perceptual generative adversarial network to enhance feature response, which converts the features of small targets into large ones with homogeneous attributes. Zheng et al. [22] propose a hyperscale detector to learn scale-invariant representations of objects. Liang et al. [23] present a dynamic enhancement

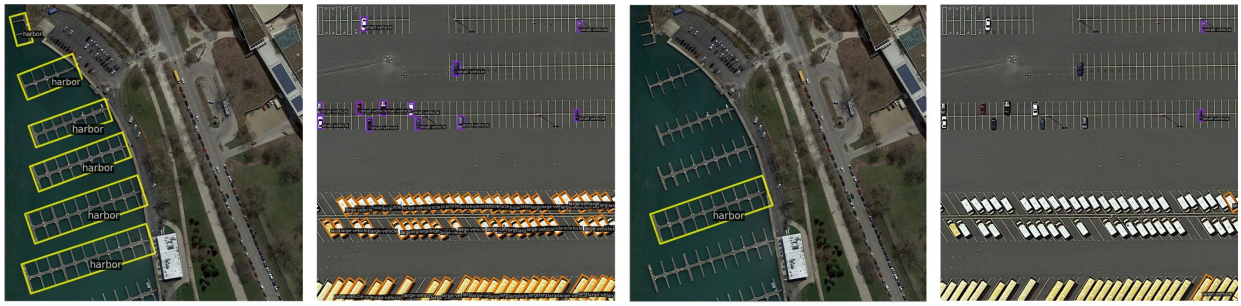

Fig. 2: Illustration of different types of AL methods. (a) Image-based sampling methods fully annotate the whole image, which may result in the inefficient and redundant annotation. (b) Object-based sampling methods select objects for which the model is most uncertain.

anchor network (DEA-Net), it contains a sample generator to augment the training data, and a discriminator to distinguish the samples between an anchor-based unit and an anchor-free unit to help train the sample generator. Liu et al. [24] design an enhanced effective channel attention (EECA) mechanism, an adaptive feature pyramid network (AFPN), and a context enhancement module (CEM) in an adaptive balanced network (ABNet) in order to capture more discriminative features.

Most of the existing methods are based on the assumption that a large number of labeled data for model training is usually unavailable in real applications due to the prohibitive labeling cost of annotating remote sensing images.

B. Active Learning and Its Applications in Object Detection

Active learning has received more and more attention in recent years due to its potential to reduce the labeling cost [25]. It selectively queries the labels of the most informative unlabeled examples from the oracle. It has been successfully applied to many important tasks in computer vision, such as classification [26]–[28], detection [6], [29], [30] and segmentation [31]–[33]. There are two settings in active learning literature, one is the membership query, and the other is the pool-based setting. Among them, the membership query method directly generates informative samples from the feature space for querying [34]–[36], while the pool-based active learning methods assume that there is a large pool of unlabeled data for sampling. They try to select the most informative examples from the pool for querying. In this paper, we mainly study the latter setting.

Active learning mainly studies effective query strategies and query types. For the former research topic, many selection criteria are proposed to evaluate the informativeness of the unlabeled data from different aspects. Existing methods can mainly be divided into the uncertainty-based approaches [37], [38], representativeness-based approaches [26], [39] and hybrid approaches [40], [41]. Uncertainty-based methods prefer the examples with high prediction uncertainty of the model. For example, Yoo et al. [30] add a loss prediction branch to the neural network to predict the loss of unlabeled samples. The data with large predicted loss will be queried from the oracle. Roy et al. [6] estimate the uncertainty by the difference between the convolutional layers of the object detector backbone. The data with high divergence are preferred in active selection. Kao et al. [3] propose "localization tightness" and "localization stability" criteria. They measure the overlap ratio between region proposals and final predictions, and the prediction changes of the original and corrupted images to evaluate the uncertainty. Representativeness-based methods prefer data that can well represent the latent data distribution. Sener et al. [39] propose a coreset method to query the data that can cover the whole dataset with a minimum radius. Sinha et al. [26] train a variational autoencoder (VAE) and an adversarial network to classify the unlabeled and labeled data, then select the data which is predicted as the unlabeled one. For the hybrid methods, which consider multiple criteria simultaneously, Ash et al. [42] propose a hybrid AL method that clusters the gradient of the target model's final output layer as the feature of the unlabeled samples that contain the uncertainty information. Sharat et al. [43] use the probability vector predicted by CNN to select objects located in different backgrounds.

For the query type, many methods are proposed to improve the querying efficiency by an effective query type, i.e., taking an object rather than the image as the basic unit for querying [4], [5], [44]. For example, Tang et al. [4] consider the partial transfer object detection task, and query the source objects which are informative and transferrable to the target domain. Laielli et al. [45] propose a region-level sampling method, which calculates the score of each region according to an accumulation of the informativeness and similarity of each query-neighbor pairing within a region and finally selects the image region with the highest score to send to the experts for annotation. Xie et al. [46] propose a region-based active learning approach for semantic segmentation with the existence of domain shift. The authors evaluate the informativeness by the category diversity of pixels within a region, and the classification uncertainty. Liang et al. [47] propose a sampling method combining spatial and temporal diversity to label the most informative frames and objects according to the multimodal information provided by the AV dataset. Most of the existing works consider the common learning tasks, but they are less applicable to aerial object detection due to the neglect of the challenges of this task.

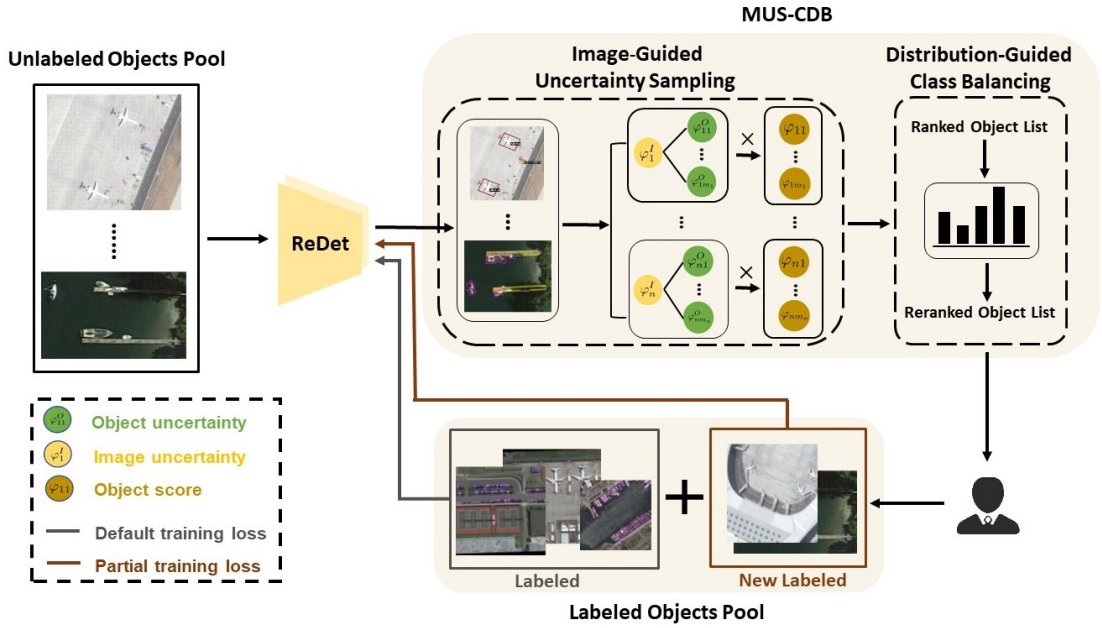

Fig. 3: Overall framework of the proposed method. The active learning sampling strategy consists of two modules: Image-Guided Uncertainty Sampling (IGUS) and Distribution-Guided Class Balancing (DGCB). The IGUS module combines the image uncertainty and object uncertainty to obtain the final object score. The DGCB module uses the class distribution of the labeled objects pool to constrain the class distribution of the sampling results. We take the model prediction object as the basic sampling unit and send it to experts for labeling. ReDet is a remote sensing object detector.

III. The Methodology

A. Problem Definition

In the active learning for aerial object detection task, there is a small set of fully labeled images

$$\mathcal{D}_L = \{(X_i^L, \mathcal{Y}_i^L)\}_{i=1}^{N_L}, \tag{1}$$

to initialize the model. To be specific,

$$\mathcal{Y}_i^L = \{(\mathbf{c}_{ij}^L, \mathbf{b}_{ij}^L)\}_{j=1}^{n_i}, \tag{2}$$

where $\mathbf{c}_{ij}^L$ is the one-hot class label belonging to one of the C known classes in label space and $\mathbf{b}_{ij}^L$ is the bounding box label. Meanwhile, there is also a large set of unlabeled images,

$$\mathcal{D}_U = \{(X_i^U)\}_{i=1}^{N_U} \tag{3}$$

for data selection, where $N_U \gg N_L$. Our goal is to train a well-performed detector f with the least queries from $\mathcal{D}_U$. Denote by

$$\mathcal{P}_U = \{(X_i^U, \hat{\mathcal{Y}}_i^U)\}_{i=1}^{N_U}, \tag{4}$$

the model prediction on the unlabeled data where

$$\hat{\mathcal{Y}}_i^U = \{(\hat{\mathbf{c}}_{ij}^U, \hat{\mathbf{b}}_{ij}^U)\}_{j=1}^{\hat{n}_i}. \tag{5}$$

To be specific, $\hat{\mathbf{c}}_{ij}^U, \hat{\mathbf{b}}_{ij}^U$ represent the model probability prediction of each category, i.e.,

$$\hat{\mathbf{c}}_{ij}^U = \{\hat{c}_{ijk}^U\}_{k=1}^{C} \tag{6}$$

and bounding box position, respectively. Given a sampling function φ, which takes $\mathcal{P}_U$ as input, to select the informative data for labeling. For the proposed object-based active learning method, we rate each predicted bounding box in the unlabeled images set $\mathcal{D}_U$, and select the top N informative predicted bounding boxes for querying. Once labeled, these bounding boxes are added to the partially labeled set

$$\mathcal{D}_P = \{(X_i^P, \mathcal{Y}_i^P)\}_{i=1}^{N_P} \tag{7}$$

to improve the object detector, along with the initial labeled set $\mathcal{D}_L$.

In this work, we design two modules to perform a cost-effective selection for objects, i.e., the Image-guided Uncertainty Sampling module (IGUS) and the Distributed-guided Class Balancing module (DGCB). Next, we will introduce the proposed query strategy in detail, and elaborate on the training scheme to fully utilize the partial labels.

B. Mixed Uncertainty Sampling

As mentioned before, existing object-based sampling methods mainly consider the information of the prediction box itself, i.e., category uncertainty or regression uncertainty, but neglect the spatial information and the semantic structure of the image. To tackle this problem, we propose to consider both uncertainty of the image and the object for more comprehensive data evaluation, which incorporates both global and local information.

As for the image uncertainty, we believe that if there are many predicted objects with high uncertainty in an image, it is not well-learned by the model. Thus, this image should be preferred in active selection. To this end, we evaluate and aggregate the uncertainty values of the most confident model predictions (i.e., objects whose predicted class confidence is greater than a specific threshold) to indicate the uncertainty

value of the whole image. This process can be regarded as a fusion of the spatial and semantic structure information of the image. Specifically, the image uncertainty φ_i^I for a given image X_i^U is formulated as

$$\varphi_i^I = 1 - \frac{1}{|\mathcal{S}_i^\theta|} \sum_{j \in \mathcal{S}_i^\theta} \max \hat{\mathbf{c}}_{ij}^U \,, \tag{8}$$

where

$$\mathcal{S}_i^\theta = \{j | \max \hat{\mathbf{c}}_{ij}^U > \theta, \forall j = 1, \ldots, \hat{n}_i\} \,, \tag{9}$$

and $|\cdot|$ represents the number of elements in the set and θ is the threshold. Here we use the parameter θ to impose preference on the images with distinct information in the active selection phase. It is conceivable that the images with many similar objects are more likely to receive a small φ_i^I due to the sufficient supervision of this pattern. On the contrary, the examples with higher values of φ_i^I are more likely to contain knowledge of the rare patterns, thus are more informative. One phenomenon to imply this motivation is that, the distribution of different categories of remote sensing images has certain regularity. Some head categories, such as small vehicle and large vehicle, are densely distributed and very similar to each other; Some medium categories, such as bridge, are sparsely distributed in the picture, and the background of the picture is blurry and difficult to distinguish; Tail categories, such as helicopter, have a small number and poor performance. These images with similar and densely distributed objects on the one hand may introduce information redundancy, on the other hand may intensify the class imbalance problem in model learning. For these reasons, we should avoid querying these images.

To further consider the instance-level information in our object-based querying, we employ entropy to evaluate the uncertainty of each predicted bounding box. Specifically, the object uncertainty is calculated as follows

$$\varphi_{ij}^O = -\sum_{k=1}^{C} \mathbb{P}(\hat{c}_{ijk}^U | X_i^U) \log \mathbb{P}(\hat{c}_{ijk}^U | X_i^U) \,, \tag{10}$$

where $\mathbb{P}(\hat{c}_{ijk}^U | X_i^U)$ is the predicted probability of the j^{th} bounding box in the image X_i^U on class k.

Although the uncertainty measurement of the objects is straightforward, we note that this criterion is combined with the image uncertainty and class balancing to conduct active selection. The overall query strategy well considers the spatial information and the semantic structure of the image, as well as the instance-level uncertainty. We will show the effectiveness of the selection criteria empirically in the experiments.

C. Class Distribution Balancing

Many remote sensing datasets suffer from the problem of class imbalance [11], as some categories are naturally rare. This phenomenon will significantly jeopardize the model performance, especially for the minority classes. To tackle this problem, we propose a selection criterion to emphasize the categories with low occurrence frequency in active querying. Specifically, we first identify the minority classes by counting the objects of each class on the labeled set. Denote by a_k the number of objects in class k, $\forall k = 1, \ldots, C$. We would like to query more objects in the class with a small a_k, e.g., by imposing a preference ζ_k, which is inversely proportional to the a_k, on each class during the selection phase

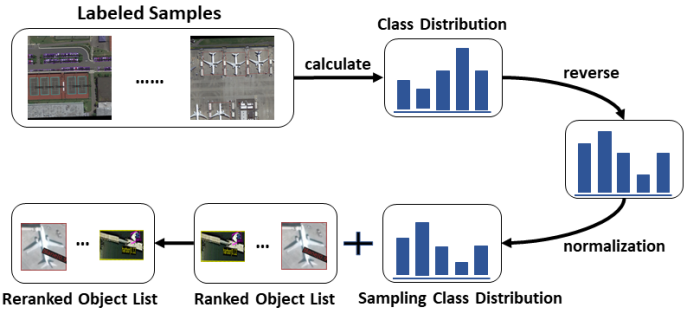

Fig. 4: The procedure of the proposed Distribution-Guided Class Balancing module (DGCB). We reverse the class distribution of the labeled samples and perform a softmax operation to obtain the distribution that the class should satisfy when sampling.

$$\zeta_k = \frac{exp(\beta_k)}{\sum\limits_k exp(\beta_k)} \,, \tag{11}$$

where

$$\beta_k = 1 - \frac{a_k}{\sum\limits_k a_k} \,. \tag{12}$$

Finally, we summarize our active learning method as the following. We first score each predicted object by

$$\begin{gathered} \varphi_{ij} = \varphi_i^I \cdot \varphi_{ij}^O \,, \\ \forall i = 1, \ldots, N^U, \ \forall j = 1, \ldots, \hat{n}_i \,. \end{gathered} \tag{13}$$

Then, we allocate the labeling budget for each class according to the ζ_k^T, and query the annotation of the top-rated candidate objects. If the budget of a specific class has run out, we drop the rest objects which are predicted as this class.

D. Dealing with Partial Labeled Data

In order to fully utilize the weakly supervised data, i.e., partially labeled images, we devise an effective training strategy. For the fully labeled datasets, we follow the default training scheme of the ReDet model to update the detector. For the partially labeled datasets, since the labeling of the images is incomplete, we cannot directly employ the vanilla training scheme, because it will treat the unlabeled region as background. To tackle this problem, we propose a partial training loss to effectively mine the latent knowledge in the unsupervised image regions that have not been queried.

Specifically, for the loss of RPN head $\mathcal{L}_{rpn}$, we take the same form as in the vanilla ReDet model [19]. For the loss of bounding-box head $\mathcal{L}_{bbox}$, inspired by the state-of-the-art semi-supervised learning method [48], in which the unlabeled data is usually exploited by assigning pseudo-labels and its loss is calculated with a relatively small weight. Because the pseudo-labels may be incorrect, down-weighting the loss is equivalent to introducing a noise prior to the model for robust

learning. Following this principle, we employ an adaptive weight to each object when calculating the loss, i.e., if the model prediction is confident, the risk that the pseudo-label is noisy is relatively low, thus the loss weight should be higher for this object. Formally, the $\mathcal{L}_{bbox}$ is defined as follows

$$\begin{aligned}\mathcal{L}_{bbox} =&\lambda_{cls}\sum_{j=1}^{W}\sum_{k=1}^{C+1}\mathbb{I}_i^{weak}\omega[-\mathbf{o_{ijk}}\log(\hat{\mathbf{o}}_{\mathbf{ijk}})]+\\ &\lambda_{cls}\sum_{j=1}^{W}\sum_{k=1}^{C+1}\mathbb{I}_i^{fully}[-\mathbf{o_{ijk}}\log(\hat{\mathbf{o}}_{\mathbf{ijk}})]+\\ &\lambda_{reg}\sum_{j=1}^{W}\sum_{u=1}^{5}\mathbb{I}_{ij}^{obj}\ell_s(\hat{v}_{iju}-v_{iju}),\end{aligned} \tag{14}$$

where

$$\omega=\begin{cases}1 & \mathbb{I}_{ij}^{obj}=1\\ \mu & \mathbb{I}_{ij}^{obj}=0\end{cases} \tag{15}$$

and

$$\ell_s(x)=\begin{cases}0.5(x)^2, & if|x|<1\\ |x|-0.5, & otherwise\end{cases}. \tag{16}$$

Here, λ_{cls} and λ_{reg} are the tradeoff parameters, i and j are the indexes of images and region proposals, respectively. W indicates the number of proposals involved in bounding-box head in training. $\mathbb{I}_i^{weak}$ and $\mathbb{I}_i^{fully}$ are indicator functions. They equal 1 when the image i is a partially labeled image or a fully labeled image, respectively. $\mathbb{I}_{ij}^{obj}$ indicates whether the proposal contains a foreground object. ω is introduced to down-weighting the background objects for robust learning.

The $\mathcal{L}_{bbox}$ contains the classification loss (the first two terms) and the box regression loss (the last term). For the classification loss, the model predicts a discrete probability distribution over C categories plus the background for each proposal, i.e., $\hat{\mathbf{o}}_{\mathbf{ijk}}, \forall k=1,\ldots,C+1$. For the regression loss, it is defined over the bounding-box coordinate offsets plus the angle, i.e., $\hat{v}_{iju}$, $\forall u=1,\ldots,5$. Further, the smooth L1 regularization ℓ_s is employed to stabilize the training.

E. The Overall Framework

We summarize our methods in Algorithm 1. Firstly, we use the labeled dataset to train the detector f, and the prediction results of f on the unlabeled dataset are filtered to form the candidate objects to be sampled (i.e., excluding the prediction objects that highly overlap with the ground truth object). Secondly, we employ the IGUS module to calculate the informativeness of the predicted objects. Thirdly, we utilize the DGCB module to calculate the category sampling distribution ζ_k^T, and initialize the annotation budges allocation vector a according to ζ_k^T. Finally, according to the annotation cost of each category shown in a, we annotate the candidate objects in the order of information content from high to low within the limit.

Algorithm 1 Our proposed MUS-CDB

Input: labeled data set D_L, unlabeled data set D_U, queried set D_P, annotation budget N, detection model f.

Train the detection model f with D_L.
$\mathcal{P}^U \leftarrow \{(X_i^U, f(X_i^U)\,|\,\forall i=1,...,N_U\}$.
Remove the objects in $\mathcal{P}^U$ which is highly overlap with the labeled object.
$\zeta \leftarrow \{\zeta_k\,|\,\forall k=1,\ldots,C\}$. ▷ by Equ. (11)
$\Phi \leftarrow \{\varphi_{ij}\,|\,\forall i=1,\ldots,N^U,\ \forall j=1,\ldots,\hat{n}_i\}$. ▷ By Equ. (13)
$\mathbf{a} \leftarrow \{a_k\,|\,a_k=N\cdot\zeta_k,\ \forall k=1,\ldots,C\}$. ▷ calculate the labeling budget for each class
$\Phi \leftarrow \mathrm{sort}(\Phi)$. ▷ sort Φ in descending order
while $N>0$ **do**
  $i,j \leftarrow \{i,j\,|\,\varphi_{ij}=\mathrm{pop}(\Phi)\}$.
  $c \leftarrow \mathrm{argmax}\{\hat{\mathbf{c}}_{ij}^U\}$.
  if $a_c>0$ **then**
    $D_P \leftarrow D_P \cup \mathrm{label}(\varphi_{ij})$. ▷ label the φ_{ij} instance according to the labeling rule
    $a_c \leftarrow a_c-1$.
    $N \leftarrow N-1$.
  end if
end while

Output: the updated queried set D_P

IV. Experiments

A. Experimental Settings

1) Dataset: DOTA is the largest oriented object detection dataset in the field of aerial imagery. It is also the common benchmark in the aerial object detection task. We conduct experiments on DOTA-v1.0 [11] and DOTA-v2.0 [12]. Specifically, DOTA-v1.0 contains 2,806 large-scale aerial images and 188,282 objects. DOTA-v2.0 collects more Google Earth, GF-2 Satellite and aerial images. There are 11,268 images and 1,793,658 objects in DOTA-v2.0. Each image is between 800 $\times$ 800 and 20000 $\times$ 20000 pixels in size. There are a total of 15 common classes in the DOTA-v1.0 containing plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP) and helicopter (HC). DOTA-v2.0 further adds the new categories, container crane (CC), airport (AP) and helipad (HP).

For DOTA-v1.0, the training and testing sets follows the setup of the previous work [19]. For DOTA-v2.0. we use the training set for training and the validation set for testing. We crop the original image to 1024 $\times$ 1024 patches with an overlap of 200. The random horizontal flip is employed in data augmentation.

2) Comparison Methods: We compare the proposed method with various types of active learning methods, including random selection (Random), uncertainty-based active learning methods (Entropy [25]), representation-based active learning methods (Coreset [39]), detection-specific active learning methods (Localization Stabiloty [3]) and object-

level active learning methods (Qbox [4]). In this paper, we introduce two baseline methods, image-based sampling and object-based query type. Both Qbox and our proposed method use an object-based query approach. Other methods use an image-based query approach. For Qbox, we take the model predicted bounding box as the candidate sampling object. The informativeness of each predicted bounding box is calculated using the inconsistency defined in Qbox, and send the predicted bounding box to the oracle for labeling in descending order. For entropy, we use average object entropy as image uncertainty.

3) Performance Measurement: We report the performance of different AL methods on the DOTA-v1.0 and DOTA-v2.0 datasets by calculating the Average Precision of each class at an IoU threshold of 0.5 and a score threshold of 0.3. Since both datasets, especially DOTA-v2.0 are evaluated in low data regimes (i.e., several tail categories have only a few labels in the first few active learning iterations), choosing 0.5 as the IoU threshold compared to 0.75 or 0.5 : 0.95 thresholds can better distinguish performance.

4) Labeling Rules: Image-based sampling methods and object-based sampling methods have different labeling rules. First, object-based sampling methods send the predicted bounding box as a unit to the oracle for labeling, and if there is only one unlabeled ground-truth object in the query bounding box, the label of that object is returned. If there are multiple unlabeled ground-truth objects in the query bounding box, calculate the IOU of each ground-truth object and the query bounding box, and return the label of the object with the largest IOU. After the queried bounding box obtains the object annotations, the annotation cost (i.e., the number of annotated objects) is increased by one. According to [49]–[51], Oracle takes 7.8s to find a target, 25.5s to draw a box, and only 1.6s to identify whether there is a target in an area, So if the query bounding box received by oracle does not contain an object, we do nothing. The image-based sampling method sends the queried image to the oracle for exhaustive annotation, and the number of objects in the image is the annotation cost of the image. Note that the cost calculation of the image-based sampling method follows the experimental setup of Qbox [4].

5) Implementation Details: We adopt ReDet [19] architecture with ReResNet50 pretrained in ImageNet [52] as backbone. All experiments are conducted on GeForce RTX 2080 Ti (11G). For DOTA-v1.0, We randomly select 5% images (1052 images) from the training and validation sets as a small set of fully labeled images. The remaining images in the training and validation sets make up the unlabeled data set. In each round, 5000 objects are selected from the unlabeled set for querying. Then, the model with default pre-trained weights will be fine-tuned on the labeled datasets for 12 epochs with a batch size of 2, and the initial learning rate is set to 2.5 $\times$ 10^{-3}, using the SGD optimizer mentum of 0.9 and weight decay of 1 $\times$ 10^{-4}. The learning rate is reduced by 0.1 at 8 and 11 epochs. For DOTA-v2.0, we randomly selected 500 images from the training set as the initially labeled images. The remaining images in the training sets are considered as the unlabeled pool. In each round, 1000 objects are selected from the unlabeled data set for querying. The parameter settings of model training are consistent with DOTA-v1.0. For active learning, the parameter settings involved in our sampling method are as follows. For the adaptive weight μ in Equ. (15), we implement it as the background score predicted by the model for the negative proposals. In this way, we can effectively reduce the impact of the partially labeled data by down-weighting the loss of the false-negative objects. For the tradeoff parameters λ_{cls} and λ_{reg} in Equ. (14), we use the number of proposals and positive proposals in a mini-batch for normalization, respectively. Regarding the hyperparameter θ in Equ. 8, we set it to 0.08, and Table IV lists the performance change of our method under different values of this parameter.

B. Performance

1) Performance Comparisons on Benchmark Datasets: We report the performances of all AL methods on the DOTA-v1.0 and DOTA-v2.0 benchmarks to demonstrate that our method can better mine the most informative objects from the unlabeled data pool in remote sensing scenarios. The experimental results of the two datasets are shown in Table I.

For DOTA-v1.0, it is evident from Table I that MUS-CDB significantly outperforms the other methods at each iteration of active learning. Specifically, in terms of mAP, when 5000, 10000, 15000, and 20000 objects were sampled, the MUS-CDB surpassed the random method by 6.8, 7, 7.4, and 5.7, and the sub-optimal method by 0.7, 1.6, 2 and 2.7,

TABLE I: Mean average precision of different AL methods on DOTA-v1.0 [11] test set and DOTA-v2.0 [12] validation set with different numbers of queried objects. Each number is displayed in percentage, and numbers in bold are the best results per column. * denotes the AL method that follows the object-based query type.

Dataset	DOTA-v1.0					DOTA-V2.0				
AL cycle	Cycle-0	Cycle-1	Cycle-2	Cycle-3	Cycle-4	Cycle-0	Cycle-1	Cycle-2	Cycle-3	Cycle-4
Random	53.9	58.1	60.3	61.8	64.4	22.8	24.6	25.1	25.3	26.2
Coreset [39]	53.9	55.8	59.6	62.1	63.6	22.8	23.5	23.8	25.1	26.6
Localization Stability [3]	53.9	61.6	63.7	66.8	66.9	22.8	27.1	30.9	31.1	33.5
Entropy [25]	53.9	64.2	65.7	67.2	67.4	22.8	28.3	30.9	31.7	35.0
Qbox* [4]	53.9	58.7	61.4	64.1	65.2	22.8	24.1	27.8	28.8	31.2
IGUS*	53.9	64.6	66.3	68.3	69.1	22.8	33.0	35.8	38.0	39.5
DGCB*	53.9	**65.2**	66.9	**69.2**	69.2	22.8	33.2	36.1	39.1	40.6
MUS-CDB*	53.9	64.9	**67.3**	**69.2**	**70.1**	22.8	**34.1**	**37.2**	**40.2**	**40.9**

TABLE II: Average category-wise AP for each method on DOTA-v1.0 [11] test set after 5 iterations of active learning. Each number shown in the table is displayed in percentage. The optimal and second optimal results in each column are highlighted in bold, and the second optimal results are underlined to distinguish them. * denotes that the AL method follows the object-based query type.

	Common			Middle						Rare					
category	SV	LV	SH	PL	BR	TC	ST	HA	SP	BD	GTF	BC	SBF	RA	HC
Random	**72.0**	68.4	78.9	88.3	35.0	**90.8**	76.8	**54.6**	56.0	58.3	49.9	68.9	38.0	54.6	10.6
Coreset [39]	71.5	66.1	**81.0**	88.1	31.8	**90.8**	75.3	**54.2**	51.2	52.5	47.5	65.6	33.7	51.1	24.6
Localization Stability [3]	**72.0**	68.0	77.9	**88.8**	**42.4**	**90.8**	76.3	48.4	54.1	66.6	59.8	**73.9**	41.6	55.3	23.0
Entropy [25]	70.7	67.2	77.7	**88.6**	32.5	**90.9**	74.5	45.6	53.4	**71.5**	**62.9**	**76.7**	50.5	54.9	**37.6**
Qbox* [4]	**72.0**	**69.5**	78.1	87.7	32.3	90.7	74.6	46.6	52.9	63.2	58.1	67.6	42.8	50.7	23.3
IGUS*	**71.7**	**69.6**	81.1	88.3	37.2	90.7	77.4	53.1	54.9	68.9	62.1	72.4	**50.9**	55.7	32.8
DGCB*	71.6	69.2	**79.5**	88.5	36.8	90.7	**78.2**	54.0	**59.1**	69.4	62.6	72.1	50.7	**57.4**	**33.2**
MUS-CDB*	**71.7**	69.2	**79.5**	**88.8**	**38.8**	90.7	**78.5**	53.3	**57.3**	**70.8**	**63.4**	**73.9**	**51.2**	**55.9**	33.1

TABLE III: Average category-wise AP for each method on DOTA-v2.0 [12] validation set after 5 iterations of active learning. Each number shown in the table is displayed in percentage. The optimal and second optimal results in each column are highlighted in bold, and the second optimal results are underlined to distinguish them. * denotes that the AL method follows the object-based query type.

	Common			Middle						Rare							
category	SV	LV	SH	PL	BR	TC	ST	HA	SP	BD	GTF	BC	SBF	RA	HC	CC	AP
Random	29.5	37.5	59.2	72.5	2.4	74.4	47.5	11.7	23.5	18.3	19.2	5.0	8.3	35.3	0.0	0.0	2.1
Coreset [39]	29.1	36.3	63.3	71.9	3.9	74.4	47.9	**17.8**	20.4	8.5	15.5	2.2	9.1	38.2	0.0	0.0	0.0
Localization Stability [3]	28.8	33.3	59.2	**76.9**	**20.0**	75.8	48.8	13.2	21.2	33.6	29.7	16.2	13.3	41.5	4.3	0.0	7.5
Entropy [25]	29.6	37.3	58.7	**75.6**	6.2	78.6	47.9	8.9	20.8	36.9	29.9	28.1	12.7	37.5	**24.0**	0.0	2.6
Qbox* [4]	**31.7**	**44.3**	**64.6**	70.9	3.4	77.9	47.9	9.1	22.3	17.2	32.2	14.0	11.8	37.9	0.0	0.0	0.1
IGUS*	30.6	41.3	63.3	73.8	7.4	79.2	48.6	15.2	**25.8**	42.4	**50.2**	37.1	**34.5**	**45.3**	3.6	**0.3**	**10.3**
DGCB*	30.2	**41.4**	63.0	75.3	9.6	**81.4**	**49.1**	**17.7**	**25.8**	**45.2**	49.3	**42.2**	31.2	42.5	**8.1**	0.0	6.9
MUS-CDB*	**30.8**	40.9	**64.0**	74.9	**11.9**	**81.2**	**49.0**	16.3	**25.5**	**43.9**	**52.5**	**41.6**	**32.8**	**45.2**	6.9	**0.04**	**13.5**

respectively. These performance boosts prove that our method can accurately identify informative objects.

For DOTA-v2.0, we can observe that MUS-CDB shows more impressive improvements over image-based and object-based sampling methods. Specifically, the results in Table I demonstrate that in terms of mAP, when 1000, 2000, 3000, and 4000 objects are sampled, the MUS-CDB surpassed the random method by 9.5, 12.1, 14.9, and 14.7, and the sub-optimal method by 5.8, 6.3, 8.5 and 5.9, respectively. The performance improvement on DOTA-v2.0 is more significant than that on DOTA-v1.0, which shows our approach can better guide low-performing models to sample highly informative objects based on only a small number of training samples. In addition, through Table I, we can also observe that our method only queries 1000 unlabeled objects in DOTA-v2.0 and can reach or even surpass the performance of other comparative methods sampling 4000 unlabeled objects. The results show that our proposed method MUS-CDB can save more than 75% of the labeling cost to reach the same performance compared to the baselines and state-of-the-art active object detection methods.

According to Table I, Entropy and localization stability perform well in DOTA-v1.0 and DOTA-v2.0. On the contrary, coreset and random perform poorly as expected, which explains that the uncertainty-based methods are more effective than representation-based ones in chaotic scenes. The object-based sampling method Qbox does not perform well on both datasets. Because Qbox is originally designed for ordinary scenes, which is very different from remote sensing scenes, therefore, the sampling strategy of Qbox cannot identify the informativeness of objects in remote sensing images accurately, which also shows the necessity of designing novel active learning methods for aerial object detection tasks. The ablation method DGCB also performs well in both datasets. This result reveals the importance of considering class balancing in active selection for remote sensing images.

2) Category-wise AP: Table III and II show the average category-wise AP of all the compared methods on DOTA-v1.0 and DOTA-v2.0, respectively. The optimal and second-best results in each column are highlighted in boldface, while the second-best result is additionally underlined.

From Table III and II, we can observe that our method MUS-CDB and its variants can usually achieve the best performance in most categories. Especially some categories that are not well-learned. This result indicates that our proposed query strategy can impartially improve the performance of each category. On the other side, Qbox and uncertainty-based active learning methods excessively query specific categories while ignoring the overall performance improvement. Specifically, Qbox focuses on the performance of bridge, tennis court, and storage tank. However, the category's performance in the rare categories has been neglected. Similarly, entropy only achieves

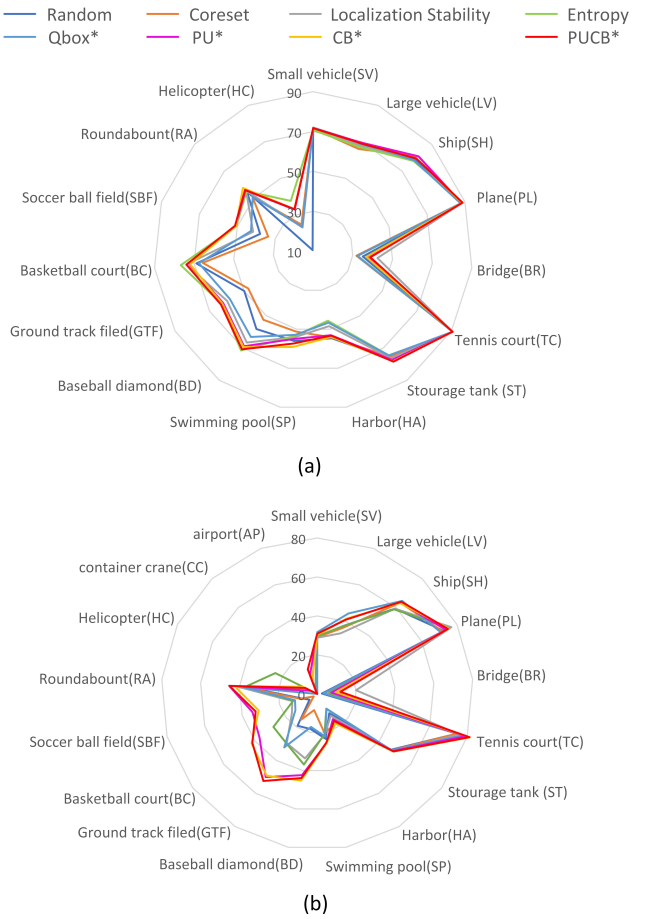

Fig. 5: Radar charts for each category of objects in different datasets. (a) DOTA-v1.0 [11]. (b) DOTA-v2.0 [12]. Different colored lines represent different active learning methods. The larger area enclosed by the outer line, the better performance of the corresponding method. The value in this graph denotes the average mAP over five iterations of active learning.

the performance improvement of helicopter and container crane and pays little attention to other categories in the Middle and Rare categories (e.g., Plane, Swimming pool, and Ground track field). The performance superiority of our method over other methods in each category is visualized in Figure 5.

3) Statistics of Queried Class Distribution: Figure 6 shows the class frequency of the DOTA-v1.0 dataset and the average class frequency of training samples selected by different methods in each cycle of AL. As expected, we see that the actual distribution of each class in the DOTA-v1.0 is exactly long-tailed. In the face of this problem mentioned above, our method (gray column) will select many common categories without the DGCB module. With the addition of the DGCB module (blue column), our method can allocate more annotation budgets to some medium and rare categories. Especially baseball diamond, bridge, ground track field, basketball court, soccer-ball field, roundabout, harbor, swimming pool, and helicopter categories. The category distribution of the random sampling method tends to be consistent with the dataset's original category distribution; The yellow curve is the growth curve formed by our method and random method on different categories. Our method outperforms the random method in all categories except the small vehicle category. The performance improvement of our method is most evident in middle and rare classes (i.e., baseball diamond, bridge, ground track field, soccer-ball field, and helicopter). These data confirmed the diversity and balance of samples selected by our method.

TABLE IV: Performance comparison on DOTA-v2.0 [12] with different score thresholds in Equ. 8 of image-guided uncertainty sampling module (IGUS).

θ	Cycle-0	Cycle-1	Cycle-2	Cycle-3	Cycle-4
0.05	22.8	30.8	36.7	39.4	40.6
0.10	22.8	**34.1**	**37.2**	**40.2**	**40.9**
0.15	22.8	32.7	37.1	39.4	40.4

4) Score Threshold θ: θ is the parameter of module IGUS in Equ. (8), which is used to impose preference on the images with distinct information in the active selection phase. According to the results in Table IV, it can be known that the optimal value of θ is obtained around 0.08. Therefore, we use θ = 0.10 for all our experiments. This can be explained by the fact that when θ is close to the lower bound of 0.05, more noise is involved in calculating the image uncertainty score. Therefore the final object information calculation is inaccurate. Conversely, if θ becomes larger, the performance degrades rapidly. This is because predicted bounding boxes containing rare patterns in the image are filtered because of low predicted scores and cannot participate in calculating of image information.

5) Visualization of the Queried Boxes: In order to demonstrate whether MUS-CDB selects the expected objects for sampling, we visualize the top-ranked objects sampled by the object-based sampling method MUS-CDB versus Qbox and the top-ranked images sampled by the image-based sampling methods random, entropy, coreset, and localization stability. The visualization results are listed in Figure 7. There is a long tail phenomenon in the DOTA-v1.0 dataset, where small vehicle, large vehicle, and ship belong to the head category, and baseball diamond, ground track field, basketball court, soccer-ball field, roundabout, and helicopter belong to the tail category. According to Figure 7, we can see that our method MUS-CDB can preferentially sample informative examples of tail categories, such as helicopter, baseball diamond, ground track field, and soccer-ball field. However, the sampling results of the random method are consistent with the original category distribution of the DOTA-v2.0 dataset, manifested in over-sampling the low-informative head categories and intermediate categories such as plane, ship, small vehicle, and large vehicle. Although the collected images have rich land types, such as airports, ports, and cities, the distribution of the queried samples is too similar to the labeled dataset, which does not improve the model performance very much. As an object-based sampling method, Qbox oversamples head categories, such as small vehicle, large vehicle, and ship, and fails to sample tail categories. Although informative samples in the head category are collected, these samples do not improve the

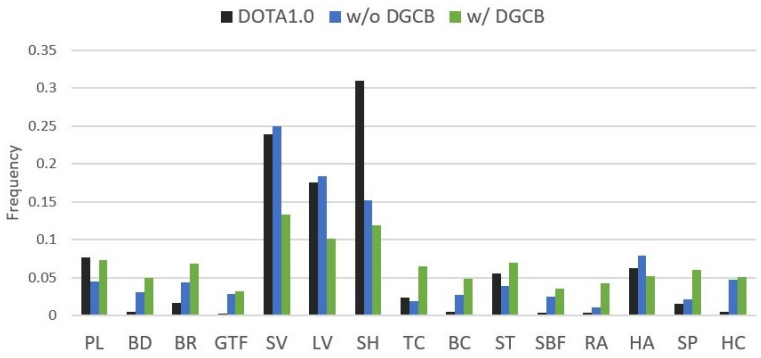

Fig. 6: The average category frequency(%) of DOTA-v1.0 [11] sampled by our AL methods in each round. Our proposed distribution-guided class balancing module (DGCB) can sample more middle and rare categories, and fewer common categories.

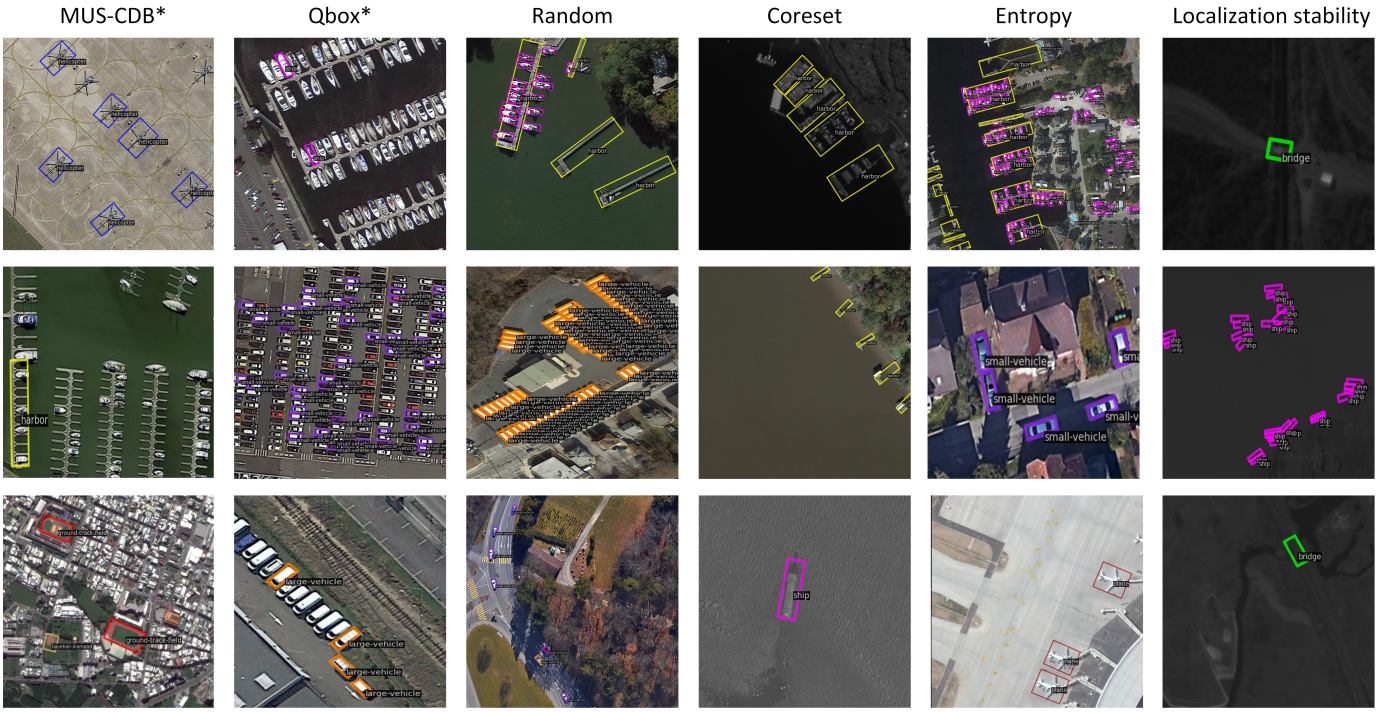

Fig. 7: Examples of the top-rated objects or images queried by MUS-CDB, Qbox [4], Random, Coreset [39], Entropy [25] and Localization Stability [3] on the DOTA-v1.0 [11]. * denotes that the AL method follows the object-based query type. The visual comparison results in the first column and the second column show that our proposed method MUS-CDB can sample more valuable middle and rare classes than another object-based sampling method Qbox. In addition, compared with other image-based sampling methods, it can be found that MUS-CDB can effectively save sampling costs through object-based sampling, and then sampling more diverse objects.

model performance much. Entropy can effectively find uncertain samples, but the image-based sampling method results in that most of the queried samples are in the same pattern, which leads to information redundancy. The images sampled by the coreset and localization stability methods have a single land type. The coreset method preferentially samples images in the ocean background, and the localization stability method preferentially samples images in the dark scenes. Although the sample distribution in the images collected by the two methods is relatively sparse, which can effectively alleviate the problem of redundant annotation. However, the two methods pay too little attention to the tail category, resulting in a low-performance improvement of the final model. These results demonstrate that our proposed method can achieve cost-effective querying in remote sensing scenarios.

C. Ablation Study

To further investigate the effectiveness of each component of MUS-CDB, we conduct ablation experiments on the DOTA-v2.0 task. As shown in Table V, satisfactory and consistent gains from the Random baseline to our whole method demonstrate the validity of each module. From Table V we can see that both IGUS and DGCB module has achieved significant performance improvements compared to the Random baseline. This shows that the module of IGUS and DGCB are helpful

TABLE V: Ablation studies on strategies of Image-guided Uncertainty Sampling module (IGUS), Distribution-guided Class Balancing module (DGCB) and partial training loss $\mathcal{L}_{partial}$.

Method	IGUS	DGCB	$\mathcal{L}_{partial}$	Cycle-0	Cycle-1	Cycle-2	Cycle-3	Cycle-4
Random				22.8	24.6	25.1	25.3	26.2
(a)	✓			22.8	33.0	36.2	38.1	39.9
(b)		✓		22.8	32.9	37.0	37.3	40.5
(c)	✓	✓		22.8	32.4	36.3	38.3	39.9
(d)	✓	✓	✓	22.8	**32.9**	**38.3**	**40.9**	**41.2**

for sampling highly informative predicted bounding boxes. In addition, we can see that (d) is obviously better than (c), which suggests that the newly designed partial training loss effectively reduces the interference of background noise, making the informativeness of the object predicted by the model more reliable.

V. Conclusion

In this paper, we propose an object-based active learning method MUS-CDB to alleviate the huge burden of data annotation of aerial object detection. On the one hand, we devise an Image-guided uncertainty sampling selection criterion in active querying to identify the most informative instances. On the other hand, we consider the long-tailed problem of the remote sensing image dataset, and impose class preference during active sampling to promote the diversity of selected objects. An effective training method for partially labeled data is also proposed to fully utilize the queried knowledge. Extensive experiments on the DOTA-v1.0 and DOTA-v2.0 benchmarks demonstrate the superiority of the proposed MUS-CDB method. In the future, we will further study the effectiveness of taking into account the angle information in the active selection for aerial object detection.

VI. References Section

References

[1] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4558–4572, 2020.

[3] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu, "Localization-aware active learning for object detection," in *Asian Conference on Computer Vision*, pp. 506–522, Springer, 2018.

[4] Y.-P. Tang, X.-S. Wei, B. Zhao, and S.-J. Huang, "Qbox: Partial transfer learning with active querying for object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[5] S. V. Desai and V. N. Balasubramanian, "Towards fine-grained sampling for active learning in object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 924–925, 2020.

[6] S. Roy, A. Unmesh, and V. P. Namboodiri, "Deep active learning for object detection.," in *BMVC*, p. 91, 2018.

[7] F. Li, S. Li, C. Zhu, X. Lan, and H. Chang, "Cost-effective class-imbalance aware cnn for vehicle localization and categorization in high resolution aerial images," *Remote Sensing*, vol. 9, no. 5, p. 494, 2017.

[8] K.-P. Ning, X. Zhao, Y. Li, and S.-J. Huang, "Active learning for open-set annotation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–49, 2022.

[9] N. Xu, C. Huo, J. Guo, Y. Liu, J. Wang, and C. Pan, "Adaptive remote sensing image attribute learning for active object detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 111–118, IEEE, 2021.

[10] G. Xu, X. Zhu, and N. Tapper, "Using convolutional neural networks incorporating hierarchical active learning for target-searching in large-scale remote sensing images," *International Journal of Remote Sensing*, vol. 41, no. 11, pp. 4057–4079, 2020.

[11] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

[12] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7778–7796, 2021.

[13] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[14] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 11, pp. 1745–1749, 2018.

[15] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, 2019.

[16] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, and C. Xu, "Dynamic refinement network for oriented and densely packed object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11207–11216, 2020.

[17] Z. Wei, D. Liang, D. Zhang, L. Zhang, Q. Geng, M. Wei, and H. Zhou, "Learning calibrated-guidance for object detection in aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2721–2733, 2022.

[18] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *European Conference on Computer Vision*, pp. 677–694, Springer, 2020.

[19] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2786–2795, 2021.

[20] H. Qiu, H. Li, Q. Wu, F. Meng, K. N. Ngan, and H. Shi, "A2rmnet: Adaptively aspect ratio multi-scale network for object detection in remote sensing images," *Remote Sensing*, vol. 11, no. 13, p. 1594, 2019.

[21] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 197–212, 2019.

[22] Z. Zheng, Y. Zhong, A. Ma, X. Han, J. Zhao, Y. Liu, and L. Zhang, "Hynet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 1–14, 2020.

[23] D. Liang, Q. Geng, Z. Wei, D. A. Vorontsov, E. L. Kim, M. Wei, and H. Zhou, "Anchor retouching via model interaction for robust object detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[24] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "Abnet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[25] B. Settles, "Active learning literature survey," 2009.

[26] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

[27] S. Ebrahimi, W. Gan, D. Chen, G. Biamby, K. Salahi, M. Laielli, S. Zhu, and T. Darrell, "Minimax active learning," *arXiv preprint arXiv:2012.10467*, 2020.

[28] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

[29] H. H. Aghdam, A. Gonzalez-Garcia, J. v. d. Weijer, and A. M. López, "Active learning for deep detection neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3672–3680, 2019.

[30] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 93–102, 2019.

[31] L. Cai, X. Xu, J. H. Liew, and C. S. Foo, "Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10988–10997, 2021.

[32] A. Casanova, P. O. Pinheiro, N. Rostamzadeh, and C. J. Pal, "Reinforced active learning for image segmentation," *arXiv preprint arXiv:2002.06583*, 2020.

[33] T. Kasarla, G. Nagendar, G. M. Hegde, V. Balasubramanian, and C. Jawahar, "Region-based active learning for efficient labeling in semantic segmentation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1109–1117, IEEE, 2019.

[34] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 580–588, Springer, 2018.

[35] C. Mayer and R. Timofte, "Adversarial sampling for active learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3071–3079, 2020.

[36] J.-J. Zhu and J. Bento, "Generative adversarial active learning," *arXiv preprint arXiv:1702.07956*, 2017.

[37] Y. Yan and S.-J. Huang, "Cost-effective active learning for hierarchical multi-label classification.," in *IJCAI*, pp. 2962–2968, 2018.

[38] A. Kirsch, J. Van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," *Advances in neural information processing systems*, vol. 32, 2019.

[39] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.

[40] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *Advances in neural information processing systems*, vol. 23, 2010.

[41] Y.-P. Tang and S.-J. Huang, "Self-paced active learning: Query the right thing at the right time," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 5117–5124, 2019.

[42] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," *arXiv preprint arXiv:1906.03671*, 2019.

[43] S. Agarwal, H. Arora, S. Anand, and C. Arora, "Contextual diversity for active learning," in *European Conference on Computer Vision*, pp. 137–153, Springer, 2020.

[44] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1605–1613, 2018.

[45] M. Laielli, G. Biamby, D. Chen, A. Loeffler, P. D. Nguyen, R. Luo, T. Darrell, and S. Ebrahimi, "Region-level active learning for cluttered scenes," *arXiv preprint arXiv:2108.09186*, 2021.

[46] B. Xie, L. Yuan, S. Li, C. H. Liu, and X. Cheng, "Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8068–8078, 2022.

[47] Z. Liang, X. Xu, S. Deng, L. Cai, T. Jiang, and K. Jia, "Exploring diversity-based active learning for 3d object detection in autonomous driving," *arXiv preprint arXiv:2205.07708*, 2022.

[48] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3060–3069, 2021.

[49] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[50] A. L. Chandra, S. V. Desai, V. N. Balasubramanian, S. Ninomiya, and W. Guo, "Active learning with point supervision for cost-effective panicle detection in cereal crops," *Plant Methods*, vol. 16, no. 1, pp. 1–16, 2020.

[51] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "We don't need no bounding-boxes: Training object class detectors using only human verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 854–863, 2016.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

ACKNOWLEDGMENT

The authors would like to thank Professor Gui-Song Xia and Dr. Jiang Ding from Wuhan University for their helpful discussion and for solving the problem when using DOTA 2.0 Dataset. This work was supported by the Natural Science Foundation of China (62272229).

BIOGRAPHY SECTION

Dong Liang received the B.S. degree in Telecommunication Engineering and the M.S. degree in Circuits and Systems from Lanzhou University, China, in 2008 and 2011, respectively. In 2015, he received Ph.D. at Graduate School of IST, Hokkaido University, Japan. He is currently an associate professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include pattern recognition and image processing. He was awarded the Excellence Research Award from Hokkaido University in 2013. He has published several research papers including in IEEE TIP/TNNLS/TGRS/TCSVT, Pattern Recognition, and AAAI.

Jing-Wei Zhang received the B.S. degree in computer science and technology from Jiangsu University, Zhenjiang, China, in 2021. She is currently pursuing the master's degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing. Her research interests include active learning and object detection.

Ying-Peng Tang received the BSc degree from the Nanjing University of Aeronautics and Astronautics, China, in 2020. He is currently pursuing the Ph.D. degree in computer science and technology with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include active learning and semi-supervised learning. He has been awarded for China National Scholarship in 2022, and the Excellent Master thesis in Jiangsu Province in 2021.

Sheng-Jun Huang received the BSc and PhD degrees in computer science from Nanjing University, China, in 2008 and 2014, respectively. He is now a professor in the College of Computer Science and Technology at Nanjing University of Aeronautics and Astronautics. His main research interests include machine learning and data mining. He has been selected to the Young Elite Scientists Sponsorship Program by CAST in 2016, and won the China Computer Federation Outstanding Doctoral Dissertation Award in 2015, the KDD Best Poster Award at the in 2012, and the Microsoft Fellowship Award in 2011. He is a Junior Associate Editor of Frontiers of Computer Science.