Journal of Multimedia Information System VOL. 12, NO. 1, March 2025 (pp. 13-26): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2025.12.1.13

Open World Object Localization Based on Negative Sample Learning

Ling Yuan[1,2], Sanquan Wang[1,2], Lingke Zeng[1,2], Jun Chu[1,2*]

Abstract

In open world object localization tasks, existing methods place excessive emphasis on learning from high-quality object samples while neglecting low-quality samples and background samples, which are considered negative samples. This tendency not only hinders the model's ability to explore the features of object candidate regions but also limits further improvements in the quality of training samples. To address this issue, we propose an open world object localization method based on negative sample learning. We design a negative sample generation model to produce background samples or low-quality object samples with weaker features, aiming to balance the sample distribution and reduce the risk of the model learning incorrect features. By incorporating low-quality object samples as negative samples in the localization branch and training them alongside positive samples, we effectively mitigate the issue of sample imbalance and reduce the instability of sampling strategies. Meanwhile, the generated background samples are used as negative samples for the classification branch, minimizing interference from unlabeled object samples in background sampling. Additionally, in open world learning scenarios, the model is prone to interference from unknown class samples during the classification phase. To address this, we introduce an Out-of-Distribution (OOD) detector before the classification task. This detector can efficiently filter out unknown class samples, reducing their negative impact on the model's classification performance and thereby enhancing the classification accuracy for known categories. To validate the effectiveness of the proposed method, we conducted experiments on the COCO dataset. The results show that our method achieves significant improvements in both localization and classification accuracy compared to other methods.

Key Words: Open World Object Localization, Negative Sample Generation Module, Negative Sample Learning, Low-Quality Sample Learning.

I. INTRODUCTION

In the real world, humans often have the ability to discover and locate new objects that they have never seen before. This capability enables humans to adapt and learn in environments where they encounter new things. However, current state-of-the-art general object detection models do not perform well in recognizing and locating these new categories of objects. The underlying reason is that these models are designed based on a closed-world assumption. Simply put, their primary task is to identify and locate objects that have been annotated, while unannotated areas are defaulted to be treated as background by the model. Therefore, These models often struggle to recognize and locate objects from new categories [1].

The open world object localization task is dedicated to identifying and localizing all potential object regions in an image, encompassing both known and unknown object categories. In this task, generating high-quality object proposals is crucial, serving as the foundation for the model to learn in an open world environment.

To address this task, there are currently two categories of methods. The first category focuses on introducing samples of unknown categories into model training. By learning the characteristics of these unknown samples, the model can enhance its ability to locate and recognize unknown objects. The key to this approach lies in improving the model's generalization performance, enabling it to better adapt to various unknown situations that may arise in the real world. For instance, ORE [2] and OSODD [3] utilize unknown-aware RPNs to generate pseudo-labels for unknown objects, while Zhao [4] employs non-parametric selective search to improve the quality of pseudo labels. The second category of methods focuses on enhancing the model's capacity to extract features from object candidate regions or improving the quality of training samples. These methods mine common object features from existing training data, enabling the model to generate more accurate object proposals in an

Manuscript received February 23, 2025; Revised March 04, 2025; Accepted March 08, 2025. (ID No. JMIS-25M-02-002)
Corresponding Author (*): Jun Chu, +86-186-0791-1101, chuj@nchu.edu.cn
[1]School of Software, Nanchang Hangkong University, Nanchang, China, yling0673@163.com, 1057837578@qq.com, 1540187869 @qq.com, chuj@nchu.edu.cn
[2]Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang, China, yling0673@163.com, 1057837578 @qq.com, 1540187869 @qq.com, chuj@nchu.edu.cn

open world environment. The core of this approach is to enhance the model's learning ability, allowing it to more accurately capture the essential characteristics of objects. For example, Kim designed the OLN network [5] to mine object features by learning the associations between object candidate regions and ground truth boxes in terms of location and shape. Additionally, some work combines features extracted by the backbone network and those from regression branches in convolutional layers, enabling the model to mine higher-quality object features.

Existing open world object localization methods, whether through introducing unknown category samples, enhancing the model's ability to mine features from object candidate regions, or improving the quality of training samples, primarily focus on enhancing the model's ability to learn from high-quality positive samples. This tendency may lead to the model mistakenly selecting negative samples during the sampling process, as illustrated in Fig. 1, where background samples may contain highly object-like samples. Additionally, general object detection models perform poorly in open environments mainly because they encounter more unknown category samples during the classification phase, increasing the risk of unknown objects being misclassified as known categories.

Addressing the aforementioned issues, we propose a method. This method aims to enhance the localization accuracy of the model for unknown category objects without introducing samples from unknown categories. It achieves this by strengthening the model's ability to learn from negative and low-quality samples. Additionally, this method can also improve the classification accuracy of the model for known categories. Specifically, aiming at the problems existing in open world object localization tasks, especially the challenges faced by the model when processing samples with weak object features and background samples, as well as the impact of unknown category samples on general object detection models, this paper proposes the following improvements based on the FCOS [6] baseline network:

- To address the difficulty of effectively learning from low-quality samples with weak object features in open world localization tasks, we introduce a negative sample generation module to produce such samples for the localization task. This ensures that the model can encounter and learn from samples of varying quality, rather than being limited to high-quality object samples.
- To address the potential issue of unlabeled strong object-like samples in background sampling for object detection tasks, we utilize the negative sample generation module to provide the model with purer background samples, reducing the risk of interference from mislabeled background samples during training.
- To address the problem that the model cannot effectively distinguish between known and unknown category samples during the classification stage of object detection, we introduce an offline Out-Of-Distribution (OOD) detector. Before the model performs classification, it screens the input samples to filter out unknown category samples, thereby reducing their number in the classification stage.

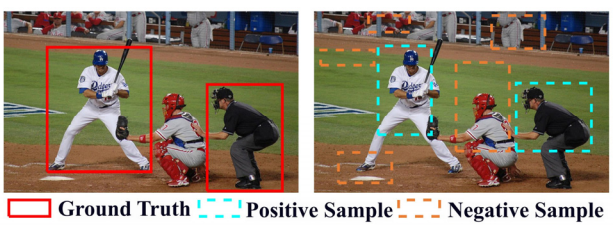


Fig. 1. Illustration of the shortcomings of open world localization methods in negative sample sampling.

II. RELATED WORK

This section will introduce the development of open world object detection [7-12]. Unlike general object detection, open world object detection focuses on identifying and processing objects of unknown categories in open and uncertain environments. Open world learning mimics human learning abilities, enabling models to gradually incorporate newly discovered categories into the set of known categories, thereby recognizing and detecting unknown objects. Early research [5] primarily focused on generating candidate regions for new objects without involving category recognition and classification, which is a crucial prerequisite for open world object detection. Against this background, ORE [2] was the first to propose an open world object detection framework, which labels unknown categories as "unknown" and learns from them after obtaining labels, allowing the model to flexibly adapt to new categories. Next, we will separately introduce the research status of open set object recognition and detection, open world object recognition and detection, as well as open world object localization, which is closely related to this paper.

2.1. Open Set Object Recognition and Detection

Open set object recognition and detection is a significant research direction in the field of computer vision, aiming to address the diversity and uncertainty of object categories in dynamic and complex environments. Research in this area is crucial for enhancing the robustness and adaptability of models, especially in practical application scenarios such as autonomous driving and intelligent surveillance.

Open set object recognition requires the model to accurately identify known objects and correctly label unknown objects after being trained solely on samples of known classes. Scheirer et al. [13] developed an open set classifier under the one-vs-rest framework, improving the classification

performance for known samples and reducing the risk of misidentifying unknown samples. The OpenMax [14] classifier further advanced this field by recognizing unknown classes in deep feature spaces, enhancing the accuracy of discrimination against unknown categories. Liu et al. [15] introduced the long-tail recognition setting and developed a metric learning framework that effectively identifies unseen categories as unknown. Other approaches, such as out-of-distribution sample detection [16] and novel class recognition [17], also enhance the accuracy of open set recognition. The introduction of self-supervised learning [18] and unsupervised learning [19] has brought new perspectives to this field.

Open set object detection is more complex, as it requires the model to precisely locate unknown objects during training and address the challenge of distinguishing between known and unknown objects. Dhamija et al. [1] were the first to propose the open set object detection protocol, guiding subsequent research efforts. Miller et al. [20] enhanced the accuracy of object detection by extracting uncertainty information from labels. Subsequent studies utilized spatial and semantic uncertainty metrics of object detectors to exclude unknown objects [21]. Miller et al. [22] also found that affinity clustering combinations can improve classification, uncertainty estimation, and object detection performance. However, these methods lack adaptability in dynamic environments, requiring retraining or adjustment when encountering new unknown objects to cope with changes in data distribution.

2.2. Open World Object Recognition and Detection

Open World Recognition and Detection differs significantly from Open Set Object Recognition in that it uses dynamic datasets, capable of continuously incorporating new known categories in a manner akin to continuous learning. Bendale and Boult [23] first introduced the concept of the Open World, revolutionizing the traditional static classification approach based on training with a fixed set of categories. The proposed setting requires models to recognize both known and unknown categories simultaneously. Additionally, Pernici et al. [24] investigated the problem of face identity learning in the open world, while Wu et al. [25] proposed a method using visible category sets to match new samples.

In summary, the open world recognition and detection work can better adapt to the changing environment by using dynamic data sets and flexible settings, and make corresponding updates when encountering new categories.

Joseph et al. [2] first introduced the concept of Open World Object Detection (OWOD), breaking the limitation of traditional frameworks that only recognize known categories, enabling models to process new categories. In OWOD, models need to recognize both known and unknown objects, and label unknown objects to update the model and adapt to changing environments. This poses numerous challenges: generating accurate bounding boxes for known categories and high-quality candidate regions for unknown objects, distinguishing between known objects, backgrounds, and unknown objects, and detecting objects of different sizes while effectively constructing contextual associations between them.

To address the challenges in OWOD, Joseph et al. [2] proposed the ORE algorithm based on Faster R-CNN. ORE utilizes the Region Proposal Network (RPN) to generate category-independent candidate regions, labels candidate regions that do not overlap with known object ground truth boxes and exhibit high “objectness” scores as unknown, and separates categories using latent space clustering techniques. At the same time, an energy-based binary classifier is used to further distinguish between unknown and known classes. Despite progress, ORE still struggles to adapt to the open world. OSODD [3] enhances performance through domain-agnostic augmentation, contrastive learning, and semi-supervised clustering. Zhao et al. [4] improves the quality of pseudo-labels for unknown objects by using non-parametric Selective Search [26] instead of the unknown-aware RPN. OpenDet [27] distinguishes between known and unknown objects by separating high-density and low-density regions in latent space.

2.3. Open World Object Localization

The task of open world object localization is inherently similar to the localization component of object detection. Object detection encompasses both localization and recognition tasks, with many previous studies implementing object detection through these two stages. Firstly, the localization stage identifies bounding boxes for potential objects in the image. Subsequently, these bounding boxes are passed to the classification stage for categorizing each object. Taking Faster R-CNN [28] as an example, its classification stage relies on bounding boxes generated by the RPN, a method that has proven effective across multiple datasets. However, due to the close dependence of the classification stage on the results of the localization stage, the object proposals generated in the localization stage often overly conform to known categories in the training data, leading to poor performance when facing unknown categories, and prone to false positives or missed detections.

In early object detection methods, techniques such as selective search [26] and edge detection [29] were used to generate category-agnostic proposals, which provided training samples for the subsequent classification tasks in object detection. These methods also provided inspiration for open world object localization. With the development of deep learning, RPN [28] and its improved versions [30-31]

learned to identify regions with a high probability of containing objects. However, RPN tends to overfit to known categories. To address this issue, the Object Localization Network (OLN) [5] replaced the category-agnostic classification head of Faster R-CNN with a localization quality prediction head, thereby avoiding treating unknown objects as background. Zhao et al. [4] proposed an auxiliary module that uses a non-selective search method to generate category agnostic proposals, assisting the RPN in generating object proposals. The aim was to reduce the probability of RPN overfitting to known categories. Saito et al. [32] employed background-erasing augmentation and multi-domain training strategies to reduce the bias of classification-based proposal networks. Gao [33] leveraged the idea of semi-supervised learning, using pseudo-labels to enable the model to learn unknown object categories.

III. METHODS

This paper employs FCOS [6] as the backbone network and addresses three challenges encountered during model training. Firstly, when localizing objects of unknown categories, it faces the scarcity of low-quality object samples. Secondly, in the process of object classification, background samples are susceptible to contamination from unlabeled samples that possess prominent object features. Thirdly, the model struggles to effectively differentiate between known and unknown samples, leading to misclassification of unknown samples during the classification phase. To tackle these challenges, this section will delve into the potential causes and propose corresponding solutions.

3.1. Network Framework

The overall framework of the network, as illustrated in Fig. 2, primarily comprises two crucial components: a negative sample generation module and an unknown class sample filtering module based on the OOD detector [34]. The

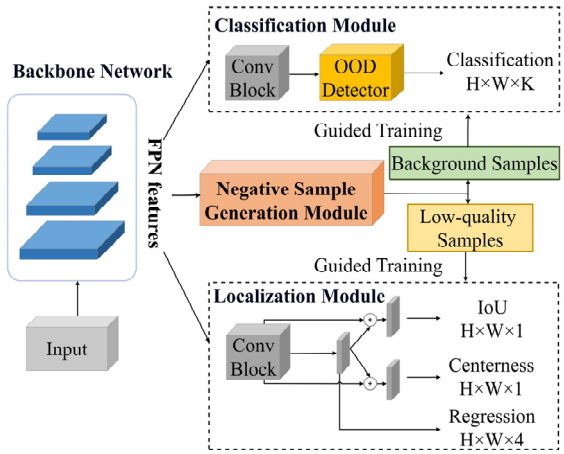


Fig. 2. Illustration of overall network framework.

method presented in this chapter is an improvement upon FCOS [6]. In the localization module, the model incorporates existing enhancements tailored for FCOS. Specifically, FCOS utilizes the centerness metric to estimate the objectness of candidate regions. However, research has shown that Intersection over Union (IoU) is also an effective method for measuring objectness. To better extract object features and improve the accuracy of bounding box predictions, some work integrates an IoU branch into the localization module, which works in conjunction with the centerness branch to jointly assess the quality of candidate bounding boxes. This enables the model to generate more precise and higher-quality object candidate boxes. The formulas for calculating IoU and centerness are provided below:

$$I = \left(min(l, l^*) + min(r, r^*)\right) \times \left(min(b, b^*) + min(t, t^*)\right), \tag{1}$$

$$U = (l^* + r^*) \times (t^* + b^*) + (l + r) \times (t + b) - I, \tag{2}$$

$$IOU = \frac{I}{U}, \tag{3}$$

$$centerness = \sqrt{\frac{min(l^*, r^*)}{max(l^*, r^*)} \times \frac{min(t^*, b^*)}{max(t^*, b^*)}}, \tag{4}$$

where $[l^*,\ r^*,\ t^*,\ b^*]$ denotes the ground truth bounding box corresponding to the object region, while $[l,\ r,\ t,\ b]$ represents the distances from the center point of this region to the left, right, top, and bottom sides of its ground truth bounding box, respectively.

To further enhance performance, some advanced models combine features extracted from the backbone network with those from the regression branch, which are then processed through convolutional layers to produce corresponding outputs. In this chapter, this structure is referred to as the Conditional Head. This paper adopts this structure and additionally introduces an OOD detector [34] in the classification module, aiming to filter out unknown class samples.

During the overall training process, the negative sample generation module is first trained to provide low-quality samples or background samples for subsequent model training. In the formal training phase, the features generated by the feature extraction network are fed into the negative sample generation module, as well as the classification and regression branches of the model's detection head. The classification branch utilizes the background samples generated by the negative sample generation module, combined with its original positive samples, for training. This reduces the likelihood of high-object unannotated samples appearing in the background samples. The localization branch is trained using samples from center point sampling and the lower-quality object samples generated by the negative sample generation module. This improves the balance between

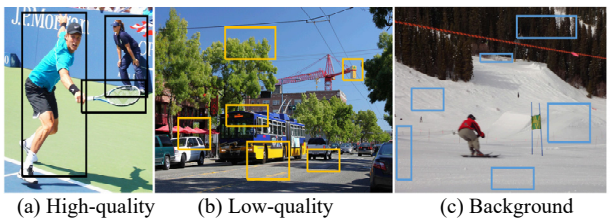
high and low localization quality samples, preventing the model from focusing solely on highly object-like samples and reducing the risk of misclassifying unknown category samples as low-quality samples. Simultaneously, the model utilizes the OOD detector [34] to filter out unknown class samples, which are then passed to the classification branch, enhancing the model's classification accuracy for known class samples.

3.2. Negative Sample Generation Module

In open world learning tasks, existing research has primarily focused on screening high-quality positive samples, with less attention given to negative samples. This has led to models becoming overly reliant on feature-prominent object samples while neglecting the learning of low-quality and background samples, thereby limiting model performance. To address this, this section proposes a negative sample generation module designed to produce background samples or low-quality samples with weaker features. This module not only enhances the diversity of training samples and maintains sample balance but also reduces the risk of the model learning erroneous features, thus improving the quality of object candidate regions.

OLN [5] as an open world object localization network, possesses the ability to generate high quality object candidate boxes that are not limited to known categories. It abandons traditional dependence on category information and instead utilizes geometric cues to accurately estimate the localization quality of candidate regions, thereby effectively identifying objects. The localization quality score accurately reflects the likelihood that a candidate region belongs to an object, enabling OLN to easily distinguish between samples with prominent object features, samples with less prominent features, and purely background regions, as shown in Fig. 3.

Therefore, we utilize the OLN model [34] as the negative sample generation model and append a sample classifier at its tail. This classifier divides candidate samples into three categories based on their scores: high-quality samples with prominent object features, low-quality samples with less prominent object features, and background samples. Low-quality and background samples are collectively defined as



(a) High-quality (b) Low-quality (c) Background

Fig. 3. Visual comparison of high-quality samples, low-quality samples, and background samples.

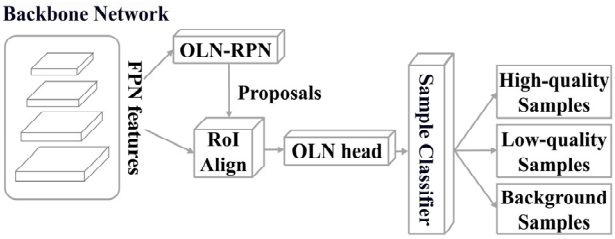


Fig. 4. Framework diagram of the negative sample generation module.

negative samples and are applied to the localization and classification branches of the model, respectively. This design allows both branches to effectively learn relevant sample features, significantly improving model accuracy. The network framework is illustrated in Fig. 4.

With this setup, the OLN model classifies candidate boxes to extract samples with less prominent features and background samples. The model's predictions range from [0,1] and are categorized into three intervals based on the localization quality score: above 0.7, between 0.3 and 0.7, and below 0.3. Specifically, samples with a localization quality score above 0.7 are highly likely to contain objects and typically exhibit distinct object features. Samples scoring between 0.3 and 0.7 are identified as low-quality samples with relatively less prominent object features. Meanwhile, samples with scores below 0.3 are considered background regions, lacking clear object features. The classification method is detailed below:

$$S = \begin{cases} S_h, & (score > 0.7), \\ S_l, & (0.3 < score < 0.7), \\ S_{bg}, & (score < 0.3), \end{cases} \tag{5}$$

where S denotes the samples generated by the negative sample generation module, "score" is the localization quality score of sample S, S_h represents high-quality samples, S_l represents low-quality samples, and S_{bg} represents background samples.

3.3. Low-Quality Sample Generation Based on the Negative Sample Generation Module

Despite previous open world object localization methods enhancing the model's ability to mine sample features by combining the centerness branch and the IoU branch, there are still deficiencies in the sample sampling process. Firstly, since both centerness sampling and IoU sampling select samples based on the positional and shape relationships between the samples and the ground truth boxes, the high-quality samples selected by the two methods may overlap in terms of features. This means that during training, the model may repeatedly learn certain specific

features.

Secondly, the IoU sampling method relies on the ground truth boxes of known categories to screen low-quality samples, which has its limitations. It may mistakenly classify candidate regions with obvious object features as low-quality samples. This is because IoU calculation is primarily based on the overlap between the candidate box and the ground truth box, rather than the objectiveness of the candidate box itself. Therefore, when the IoU value between a candidate box and a ground truth box is low, even if the candidate box actually contains a clear object, it may be misclassified as a low-quality sample. This misclassification can lead the model to learn incorrect features, thereby affecting the accuracy of its object localization.

To address the above issues, this chapter abandons IoU sampling based on existing work. Instead, it only adopts the centerness sampling method to collect high-quality object samples for the model and utilizes samples with localization quality scores between 0.3 and 0.7 from the negative sample generation module as low-quality object samples. With this approach, the localization module can not only learn the features of high-quality samples with prominent object features but also learn the features of low-quality samples more accurately. This capability enables the localization module to better distinguish samples with different features. Such a strategy enhances the model's ability to recognize and locate samples of different qualities.

3.4. Background Sample Generation Based on the Negative Sample Generation Module

Ensuring the model's localization accuracy in an open world while improving its ability to accurately recognize known categories is a key challenge in open world learning. To address this issue, a method called Unknown Object Masking has been proposed. This method, for the first time, optimizes background samples by reducing the interference of unlabeled object samples on the model, thereby enhancing the model's ability to distinguish unlabeled object samples from background samples.

The core of this method lies in hiding background regions that contain potential object features to prevent these regions from being mistakenly sampled as background samples. However, this masking strategy cannot fully cover all unlabeled objects in the image, making it impossible to form a completely continuous masked area. This can lead to unlabeled objects still being mistakenly collected as background samples during background sample collection. Additionally, discontinuous masked areas can hinder the model's learning of background information.

We use the negative sample generation module to solve this problem. This module selects samples with localization quality scores less than 0.3 as background samples, replacing the background samples originally collected through the Unknown Object Masking method. In this way, the model can obtain more accurate and pure background samples, thereby better learning background features and improving overall object detection performance.

3.5. Unknown Class Sample Filtering Based on the OOD Detector

Open world learning tasks introduce a large number of candidate bounding boxes belonging to unknown categories. Due to the lack of effective training to distinguish between known and unknown category objects, this may increase the risk of unknown category objects being mistakenly identified as known categories. To reduce the risk posed by unknown category samples during the classification stage and ensure the detection accuracy of the model for known categories, we adopt an offline OOD detector to identify and eliminate unknown classes from the samples. The main advantage of this approach lies in its flexibility and compatibility, as the offline OOD detector can be integrated into different object detection models as an independent module without requiring any specific modifications to the model itself. Additionally, this offline framework design also addresses potential objective conflicts between object detection and OOD detection tasks.

Specifically, this method employs the offline OOD detector proposed by Liu et al. [34], which trains and fine-tunes the DINO model using data from known categories. In this process, the OOD detector is configured as a classifier with K+1 output categories, where K represents the number of known object categories. The additional category is dedicated to identifying samples that do not belong to the known categories, including both unknown category samples and potential new category samples. Unlike previous research that requires a large amount of additional OOD data for training, this method only uses labeled data from known categories for training. It categorizes all samples that do not belong to known categories (including unknown category samples and background samples) as OOD category samples. During training, the IoU between the ground truth boxes and candidate regions is compared, and regions with lower IoU values are classified as OOD category samples, as these regions are likely to contain unknown category or background samples.

Although treating background samples as OOD is not the most ideal choice, this method provides an effective mechanism for the model to distinguish between known and unknown category samples during the learning process, thereby reducing the risk of unknown categories appearing in subsequent classification stages. To distinguish whether a object sample belongs to a known category or the OOD category, this method introduces the Mahalanobis Distance

to calculate the threshold γ_{ood} for the OOD score, with the specific formula as follows:

$$\gamma_{ood} = \max_{k} -(f(x)-\tilde{\mu}_k)^{\top}\tilde{\Sigma}^{-1}(f(x)-\tilde{\mu}_k), \quad (6)$$

where $f(x)$ represents the intermediate feature vectors extracted from the candidate regions in the image by the DINO model, and $\tilde{\mu}_k$ denotes the estimated mean vector of the category, which reflects the average manifestation of that category in the feature space. $\tilde{\Sigma}$ as the estimated variance matrix, describes the correlation and variation among the dimensions of the feature vectors. Before classifying the candidate bounding boxes, this method calculates the OOD score for all samples and compares it with a preset threshold γ_{ood}. If a sample's score exceeds γ_{ood}, it is classified as a known category and sorted into the corresponding classification. If a sample's OOD score falls below γ_{ood}, it is identified as an OOD sample and removed from the dataset.

This method effectively reduces the number of unk nown category samples at the classification stage, ther eby decreasing the likelihood of misidentifying unkno wn categories as known ones. Such processing enhanc es the model's performance in detecting known catego ries.

IV. EXPERIMENTS

To verify the effectiveness of the proposed method, we conducted the following experiments. First, we performed ablation studies to analyze the contribution of each module. Second, for the open world object localization task, we compared the impact of IoU scores, centerness scores, and different sampling methods on improving candidate box quality. Additionally, we conducted a visual analysis of different negative sample generation modules to validate the reliability of OLN. Finally, we compared our model with mainstream methods on the COCO dataset to evaluate its detection performance in open world learning tasks.

4.1. Experimental Setup

Experimental Data: In this study, we primarily used the MS COCO dataset [35] for experiments and evaluation. To ensure fairness in the evaluation, the COCO dataset was divided into 20 known categories, which are present in PASCAL VOC [36], and 60 unknown categories, which do not appear in PASCAL VOC. During the training phase, the model was trained exclusively on the 20 known categories, while during the evaluation phase, it was tested on the 60 unknown categories. To ensure that the evaluation focuses on the model's ability to detect unseen categories, annotations of known categories were ignored during the evaluation process.

Evaluation Metrics: In the open world object localization task, the model's performance is evaluated using the Average Recall (AR) metric. This metric assesses the model's performance by calculating the recall of unknown categories within the top N generated candidate boxes. Here, N refers to the number of candidate boxes considered during the evaluation. For the open world object classification task, performance is evaluated using the Average Precision (AP) metric. This metric is calculated by first computing the average precision for each category in the known category set, and then averaging these average values.

Experimental Details: The algorithm in this paper uses FCOS as the baseline network. The experiments were conducted on an Ubuntu 18.04.5 operating system, with an Intel Xeon Silver processor, Python version 3.7, and the deep learning framework PyTorch 1.7.1. The GPU used is Nvidia GeForce RTX 3090, with two cards providing a total of 48 GB of VRAM. The initial learning rate for all experiments was set to 0.001.

4.2. Ablation Study

To validate the effectiveness of the method proposed in this paper, we conducted experiments on the low-quality sample substitution method, the background sample substitution method, and the unknown-category sample filtering method based on an OOD detector, within the frameworks of open world localization tasks and open world classification tasks, respectively.

Table 1 shows the experimental results of replacing the traditional IoU sampling method with the low-quality sample replacement method in the open world localization task. Here, OWP refers to the model that combines the centerness branch with the IoU branch, using the Conditional Head and joint sampling of centerness and IoU. NS refers to the low-quality sample replacement method. According to the data in Table 1, it can be seen that the low-quality sample replacement method improves the model's localization accuracy for unknown categories, resulting in a 1.27% and 1.07% increase in AR10 and AR100, respectively. This validates the effectiveness of the low-quality sample replacement method in open world object localization tasks.

Table 2 presents the experimental results of using the

Table 1. Ablation study for open world object localization task.

Method	$COCO^{base}$		$COCO^{novel}$	
	AR10	AR100	AR10	AR100
FCOS	55.33	59.23	-	-
OWP	34.52	54.10	14.51	31.26
OWP+NS	35.67	55.45	15.78	32.33

Table 2. Ablation study for open world classification task.

Method	$COCO^{base}$			$COCO^{novel}$	
	AP	AR10	AR100	AR10	AR100
FCOS	44.21	55.33	59.23	-	-
Joint (OWP+classification)	46.01	58.24	63.45	11.12	26.89
Joint+unknown	46.21	60.42	64.85	11.32	27.82
Joint+**BS**	46.83	60.91	65.24	11.71	28.36
Joint+**BS**+**OOD**	**47.72**	**61.76**	**66.73**	**12.44**	**28.97**

background sample replacement method and the OOD-based unknown class sample filtering method in the open world classification task. These methods are designed to improve the model's recognition performance for known category objects. In Table 2, the background sample replacement method replaces the unknown object masking method. The background sample replacement method provides the model with cleaner background samples, which helps the model better distinguish between foreground and background. Additionally, by introducing the OOD-based unknown class sample filtering method, samples from unknown categories can be removed from the classification branch, allowing the model to focus more on classifying known categories.

In Table 2, "joint" refers to training with both open world localization and classification branches jointly; "unknown" refers to the addition of the unknown object masking method; "BS" stands for the background sample replacement method; and "OOD" indicates the introduction of the OOD-based unknown class sample filtering method.

Overall, the experimental results in Table 2 demonstrate the effectiveness of the background sample replacement method and the OOD-based unknown class sample filtering method in the open world classification task. Both methods help improve the model's performance when handling known categories.

4.3. Comparative Experiment on the Impact of IoU Score, Centerness Score, and Different Samplin g Methods on Candidate Box Quality

This section compares several methods for improving the quality of object candidate box predictions. Specifically, this experiment incorporates three localization quality metrics: centerness, IoU, and the square root of the product of centerness and IoU ($\sqrt{\text{centerness} \times \text{IOU}}$). Two sampling strategies were adopted: center sampling, which selects samples based on the centerness of the candidate boxes, and IoU sampling, which selects samples based on the IoU of the candidate boxes. Additionally, a low-quality sample substitution method (NS) was introduced to provide low-quality samples with less prominent features.

The experimental results are shown in Table 3. In this table, "conditional" refers to the method that combines features extracted by the backbone network and features from the regression branch in the convolutional layer to predict the localization quality score. This approach more effectively utilizes information in the network, thereby improving the quality of candidate boxes. "CS" refers to the centerness sampling method, and "IS" refers to the IoU sampling method.

By analyzing Table 3, it can be observed that using Conditional IoU-CS-NS produces the highest quality object candidate boxes.

Table 3. Comparison of factors affecting object candidate box quality.

Method	$COCO^{base}$		$COCO^{novel}$	
	AR10	AR100	AR10	AR100
Centerness	22.71	44.09	9.90	25.48
IOU	32.67	53.26	13.43	28.31
$\sqrt{\text{centerness} \times \text{IOU}}$	26.20	48.66	11.60	28.10
IOU-CS-IS	32.86	53.59	13.08	30.19
Conditional IOU-CS-IS	34.52	54.10	14.51	31.26
Conditional $\sqrt{\text{centerness} \times \text{IOU}}$-CS-IS	32.91	53.23	12.37	29.45
Conditional IOU-CS-**NS**	**35.67**	**55.45**	**15.78**	**32.33**
Conditional $\sqrt{\text{centerness} \times \text{IOU}}$-CS-**NS**	33.54	53.38	13.82	30.19

4.4. Visualization Analysis of Different Negative Sample Generation Modules

In open world object localization tasks, learning negative samples is crucial for the model's performance. As shown in Fig. 5, there are significant differences in localization performance across different negative sample generation modules. Traditional methods often fail to accurately distinguish between high-quality object samples, low-quality object samples, and actual background, and this lack of distinction can lead to the model learning misleading information, which in turn affects the experimental accuracy. The method in this paper references the open world object localization model (OLN), which can more reliably provide low-quality or background samples.

Fig. 5 shows the negative samples generated by different models for the same image. It is clearly observed that the OLN method demonstrates higher precision in generating negative samples. In contrast, the RPN model fails to avoid unmarked objects in the generated negative samples, indicating its relatively weak ability to filter negative samples. This experiment demonstrates that the OLN method can provide more accurate negative samples for model training, reducing the interference of erroneous information, and thus helping to improve the accuracy and robustness of the model in open world object localization tasks.

However, during the experimental process, we observed that the model's performance would be affected to a certain degree when it was exposed to low-light scenarios. This is primarily because low-light conditions limit the clarity and contrast of images, posing greater challenges for the model to distinguish between targets and backgrounds. Specifically, in such environments, some of the negative samples generated by the model may contain larger object regions, as illustrated in Fig. 6. Consequently, further refinement of the model is required in the future to enhance its accuracy in complex scenarios.

4.5. Comparison with Mainstream Algorithms

In this section, to validate the effectiveness of the proposed method, a comparison is made with other existing methods. These comparisons are divided into two main parts: the first part focuses on the open world object localization task, and the second part focuses on the open world object classification task. The specific results are shown in

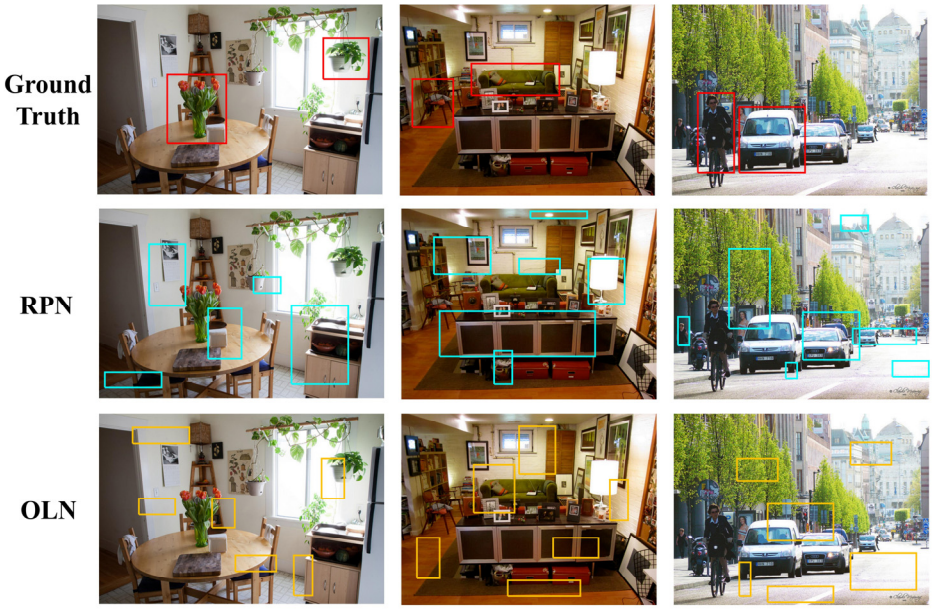


Fig. 5. Comparison of different negative sample generation modules.

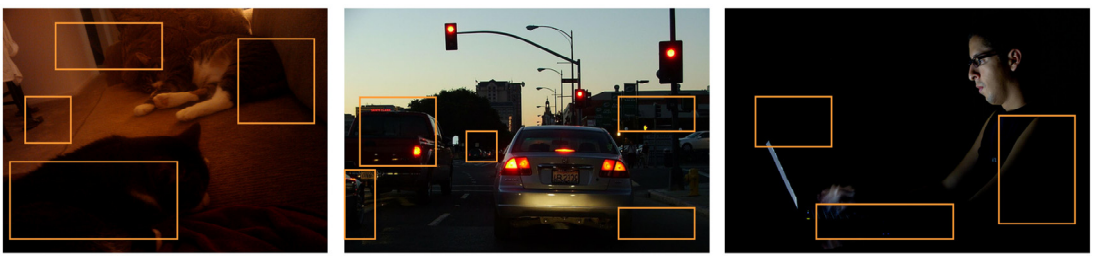


Fig. 6. Illustration of failure cases.

Table 4. Comparison with mainstream algorithms.

	Method	$COCO^{base}$			$COCO^{novel}$	
		AP	AR10	AR100	AR10	AR100
Localization	FCOS (baseline)	-	-	-	10.53	24.40
	RPN	-	-	-	7.5	20.10
	IoU-aware	-	-	-	9.78	22.23
	OLN-RPN	-	-	-	11.70	27.40
	Cascade RPN	-	-	-	12.62	27.69
	OWP	-	34.52	54.10	14.51	31.26
	OWP+NS (**Ours**)	-	**35.67**	55.45	**15.78**	**32.33**
Classification	FCOS (baseline)	44.21	55.33	59.23	-	-
	Joint (OWP+classification)	46.01	58.24	63.45	11.12	26.89
	Joint+unknown	46.21	60.42	64.85	11.32	27.82
	Class-Aware OLN	-	-	-	13.30	26.50
	Joint+BS+OOD (ours)	**47.72**	**61.76**	66.73	**12.44**	**28.97**

Table 4.

In the localization task, using the binary classification FCOS as the baseline network, it can be observed that the proposed method significantly outperforms one-stage open world object localization methods, such as RPN and IoU-aware. Additionally, the proposed method also outperforms some two-stage open world object localization methods, such as Cascade RPN. For intuitive comparison, we have visualized the localization results obtained by methods including FCOS, RPN, Cascade RPN, and OWP+ NS. As illustrated in Fig. 7, our method demonstrates significantly better localization performance, detecting more objects than the others.

In the classification task, FCOS is also used as the baseline network. In Table 4, "joint" refers to training with both the OWP and classification branches combined, "unknown" refers to the addition of the unknown object masking method, "Class-Aware OLN" refers to the addition of OLN to the classification branch, "BS" refer to the background sample replacement method, and "OOD" refers to the introduction of the Out-of-Distribution detector-based unknown class sample filtering method. Through comparison, it is found that the proposed method achieves the best performance in both the detection accuracy for known categories and the localization accuracy for unknown categories. We have visualized the detection results for known classes us-

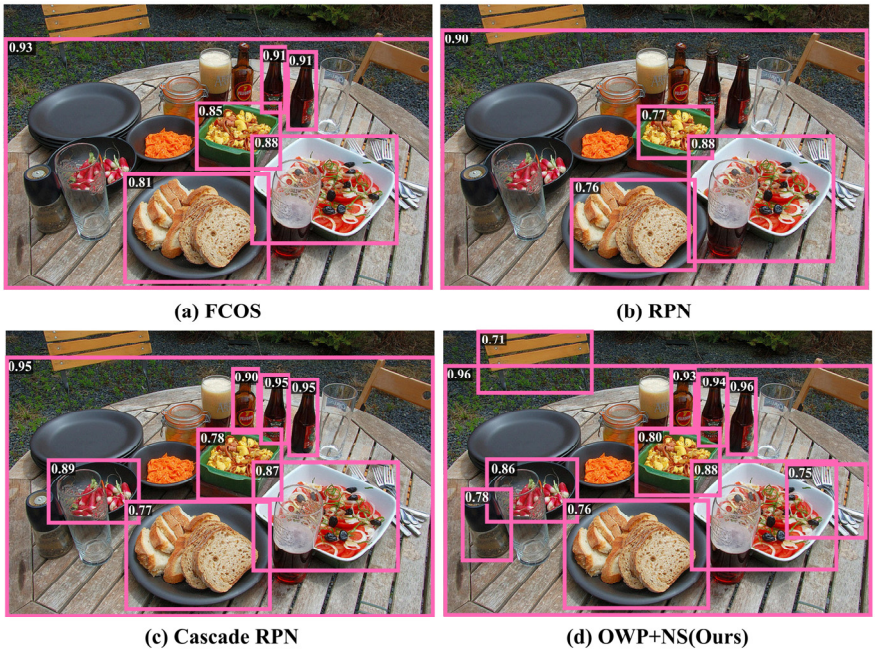


Fig. 7. Comparison of effects for open world object localization methods.

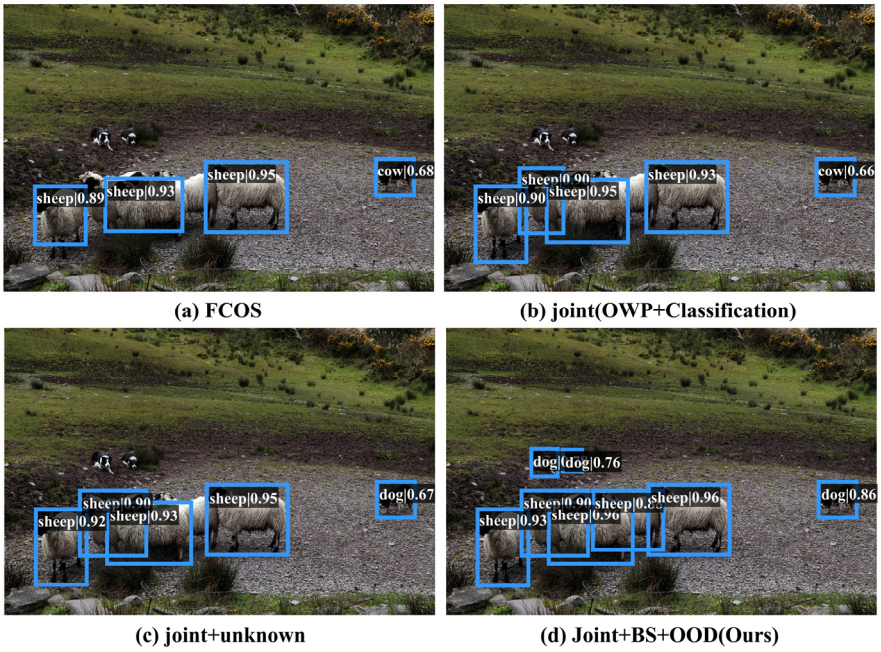


Fig. 8. Comparison of effects for open world object classification methods.

ing methods such as FCOS, joint (OWP+Classification), joint+unknown, and Joint+BS+OOD. As illustrated in Fig.8, our detection results are more accurate. Specifically, our method detects more instances of sheep and correctly identifies dogs. In summary, the proposed method demonstrates excellent performance in both open world object localization and classification tasks, effectively improving the model's detection capability for known categories and localization accuracy for unknown categories.

V. CONCLUSION

Addressing issues such as sample imbalance, unstable sampling strategies, and decreased recognition accuracy for known categories in open world object localization tasks, this paper proposes an improved method based on negative sample learning. By designing a negative sample generation model, we produce high-quality negative samples and low-quality object samples, effectively mitigating the sample imbalance problem and reducing the risk of the model learning incorrect features. Additionally, an offline OOD detector is used to filter unknown samples before classification, minimizing interference with known categories' accuracy. Experimental results demonstrate that the proposed method significantly enhances the localization and classification accuracy of the model on the COCO dataset, validating its effectiveness. The work presented in this paper offers a novel solution for open world object localization tasks, enabling better handling of unknown category objects while maintaining high recognition accuracy for known categories. Future research can further explore how to integrate incremental learning to enable the model to progressively expand its recognition capabilities for unknown categories.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant 62162045.

REFERENCES

[1] R. Dhamija, M. Günther, J. Ventura, and T. E. Boult, "The overlooked elephant of object detection: Open set," in *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, 2020, pp. 1010-1019.

[2] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, 2021, pp. 5826-5836.

[3] J. Zheng, W. Li, J. Hong, L. Petersson, and N. Barnes, "Towards open-set object detection and discovery," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, 2022, pp. 3960-3969.

[4] X. Zhao, Y. Ma, D. Wang, Y. Shen, Y. Qiao, andX. Liu, "Revisiting open world object detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 5, pp. 3496-3509, May 2024.

[5] D. Kim, T. Y. Lin, A. Angelova, I. S. Kweon, and W.

Kuo, "Learning open-world object proposals without learning to classify," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5453-5460, Apr. 2022.

[6] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1922-1933, 2020.

[7] J. Chu, W. Lin, and P. Xu, "Reviews of feature mismatch in object detection," *Journal of Nanchang Hangkong University (Natural Science Edition)*, vol. 35, no. 3, pp. 1-8, Mar. 2021.

[8] H. Zheng and J. Chu, "Feature fusion method for object detection," *Journal of Nanchang Hangkong University (Natural Science Edition)*, vol. 36, no. 4, pp. 59-67, 2022.

[9] Y. Zhang, J. Chu, and L. Wang, "Bird detection combined with the DPM model aggregate channel features in natural scene," *Journal of Nanchang Hangkong University (Natural Science Edition)*, vol. 32, no. 1, pp. 76-82, p. 89, 2018.

[10] Y. W. Lee and B. G. Kim, "Attention-based scale sequence network for small object detection," *Journal of Heliyon (Cell Press)*, vol. 10, no. 12, 2024.

[11] H. J. Park, Y. J. Choi, Y. W. Lee, and B. G. Kim, "ssFPN: scale sequence (S^2) feature based-feature pyramid network for object detection," *Journal of Sensors (MDPI)*, vol. 23, no. 9, p. 4432, 2023.

[12] H. L. Lee, Y. J. Kim, and B. G. Kim, "A survey for 3D object detection algorithms from images," *Journal of Multimedia Information System (KMMS)*, vol. 9, no. 3, pp. 183-190, Sep. 2022.

[13] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757-1772, Jul. 2013.

[14] Bendale and T. E. Boult, "Towards open set deep networks, " in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 1563-1572.

[15] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 2532-2541.

[16] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv Prep. arXiv:1706.02690*, 2020.

[17] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, vol. 31, 2018, pp. 6823-6834.

[18] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigginton, and V. Ordonez, "Generative-discriminative feature representations for open-set recognition," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020, pp. 11811-11820.

[19] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 4011-4020.

[20] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, pp. 3243-3249.

[21] D. Hall, F. Dayoub, J. Skinner, H. Zhang, D. Miller, and P. Corke, "Probabilistic object detection: Definition and evaluation," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, 2020, pp. 1020-1029.

[22] D. Miller, F. Dayoub, M. Milford, and N. Sünderhauf, "Evaluating merging strategies for sampling-based uncertainty techniques in object detection," in *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, 2019, pp. 2348-2354.

[23] Bendale and T. Boult, "Towards open world recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, 2015, pp. 1893-1902.

[24] F. Pernici, F. Bartoli, M. Bruni, and A. D. Bimbo, "Memory based online learning of deep representations from video streams," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* Salt Lake City, UT, 2018, pp. 2324-2334.

[25] H. Xu, B. Liu, L. Shu, and P. Yu, "Open-world learning and application to product classification," in *The World Wide Web Conference*, 2019, pp. 3413-3419.

[26] J. R. R. Uijlings, K. E. A. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, pp. 154-171, 2013.

[27] J. Han, Y. Ren, J. Ding, X. Pan, K. Yan, and G. Xia, "Expanding low-density latent regions for open-set object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9591-9600.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region pro-

posal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[29] L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *Computer Vision-ECCV 2014: 13th European Conference*, Zurich, Switzerland, Sep. 2014, pp. 391-405.

[30] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," in *Neural Information Processing Systems*, 2019, p. 32.

[31] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 2960-2969.

[32] K. Saito, P. Hu, T. Darrell, and K. Saenko, "Learning to detect everything in an open world," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2022, pp. 268-284.

[33] Gao, J. Hao, and Y. Guo, "OSDet: Towards open-set object detection," in *2023 International Joint Conference on Neural Networks (IJCNN),* Gold Coast, Australia, 2023, pp. 1-8.

[34] Y. C. Liu, C. Y. Ma, X. Dai, J. Tian, P. Vajda, and Z. He, et al., "Open-set semi-supervised object detection," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2022, pp. 143-159.

[35] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, and J. Hays, et al., "Microsoft coco: Common objects in context," in *Computer Vision-ECCV 2014: 13th European Conference*, Zurich, Switzerland, Sep. 2014, pp. 740-755.

[36] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge", *International Journal of Computer Vision*, vol. 88, pp. 303-338, 2010.

AUTHORS

Ling Yuan received the B.S. degree in 2022 in Software Engineering from the Nanchang Hangkong University, Jiangxi, China, where she is currently a M.S. student in Software Engineering. Her research interests include computer vision, image processing and deep learning.

Sanquan Wang received the B.S. degree in Software Engineering from Nanchang Hangkong University in 2021 and the M.S. degree in Software Engineering from the same university in 2024. His research interests include computer vision, image processing and deep learning.

Lingke Zeng received the B.S. degree in 2022 in Software Engineering from Baicheng Normal University, Jilin, China. He is currently a M.S. student at Nanchang Hangkong University, Jiangxi, China. His research interests include computer vision, image processing, and deep learning.

Jun Chu received the Ph.D. degree from Northwestern Polytechnic University, Xi'an, China, in 2005. She was a Post-Doctoral Researcher with the Exploration Center of Lunar and Deep Space, National Astronomical Observatory of Chinese Academy of Sciences. She was a Visiting Scholar with the University of California at Merced, Merced, CA, USA. She is currently the Director of the Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition and a Full Professor with the School of Software, Nanchang Hangkong University. Her research interests include computer vision and pattern recognition.