**Journal of Computer Science and Artificial Intelligence**
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 3, 2025

# Efficient Visual Region Recognition in the Open World Scenarios

**Jing Wang *, Yonghua Cao**

School of Software, Henan Polytechnic University, Jiaozuo, 454003, China

* **Corresponding author**: wjasmine@hpu.edu.cn

**Abstract:** Open-World Object Detection (OWOD) aims to address the core challenges that traditional detection models cannot dynamically adapt to unknown objects and continuously learn new categories. Most existing methods rely on visual similarity to discover new objects, yet they still face issues such as semantic disconnection and catastrophic forgetting in complex cross-domain scenarios, such as medical imaging and autonomous robotics. This paper proposes a causality-driven open-world object detection framework. By decoupling the functional causal features and the appearance features of objects, it realizes the discovery of unknown objects and incremental learning based on physical property reasoning. The core contributions include: constructing a feature enhancement model, which significantly improves the quality and expressiveness of image features; establishing a similarity calculation mechanism. By accurately calculating the similarity between text and images, it effectively avoids the over-confidence phenomenon in model prediction, thus improving the accuracy and reliability of detection; adopting a multi-granularity visual stream to conduct multi-dimensional and refined feature processing on image features, fully exploring the multi-level information in images. To comprehensively evaluate the performance of the proposed method, we conduct in-depth research on existing open-world benchmarks and extensively validate it on public benchmark datasets. The experimental results clearly demonstrate that this method achieves absolute performance gains in the detection tasks of unknown categories, fully demonstrating its strong generalization ability. This research result provides valuable reference and inspiration for the further development of the open-world object detection field, and is expected to promote technological progress and innovation in this field.

**Keywords:** Open world object detection, Multimodal visual detection, Incremental learning.

## 1. Introduction

With the rapid development of artificial intelligence technology, object detection, as one of the core tasks in the field of computer vision, plays a crucial role in many practical application scenarios, such as autonomous driving, intelligent security, and industrial inspection. However, most traditional object detection methods [1, 2] assume that the test data and training data are from the same distribution, and the model knows all possible target categories during training. This closed-world setting has significant limitations when facing the complex and ever-changing real world. The real world is an open environment where new object categories constantly emerge, and the diversity and complexity of scenes far exceed the coverage of training data. For example, in the autonomous driving scenario, new types of traffic signs, special vehicle types, or rare road conditions may be encountered. In security monitoring, previously unseen suspicious items or behavior patterns may appear. Therefore, how to enable object detection models to identify unknown objects and continuously learn new categories in the open world has become an important issue that urgently needs to be solved in the current field of computer vision.

At present, Open-World Object Detection (OWOD) has become a research hotspot in the field of computer vision. Existing OWOD methods [3] mainly focus on discovering new objects through visual similarity. [4] Some methods utilize Generative Adversarial Networks (GANs) to generate samples of unknown categories, expanding the training data to enhance the model's generalization ability to unknown categories. Others are dedicated to devising more effective feature extraction and classification strategies, attempting to accurately identify targets of unknown categories with limited training data of known categories.

Although existing methods have made some progress in open-world object detection, they still face severe challenges in complex cross-domain scenarios. Among them, the problems of semantic disconnection and catastrophic forgetting [5] are particularly prominent. Semantic disconnection means that when dealing with data from different domains, due to differences in data distribution and semantic representation, it is difficult for the model to establish a unified semantic understanding, resulting in a significant decline in the detection performance of unknown objects. For example, when shifting from natural image object detection to medical image object detection, the image modality, features, and semantics have changed dramatically, and existing methods can hardly adapt effectively.

Catastrophic forgetting refers to the phenomenon that when a model learns new categories, it easily forgets the knowledge learned previously, leading to a sharp decline in the detection performance of old categories. This makes it difficult for the model to achieve continuous and stable learning and detection in the open world. To address the above problems, this study proposes a causality-driven open-world object detection framework. By constructing a feature enhancement model, a precise similarity calculation mechanism, and adopting a multi-granularity visual stream, the decoupling of the functional causal features and the appearance features of objects is realized, thereby achieving the discovery of unknown objects and incremental learning based on physical property reasoning. The research results of this study not only help to break through the limitations of existing open-world object detection methods and improve the performance of the

model in complex cross-domain scenarios but also provide new ideas and methods for basic research and practical applications in the field of computer vision, which has important theoretical significance and application value.

At present, most of the research [6, 7] focuses on improving the detection accuracy of models for known-category targets, which undoubtedly serves as an important cornerstone for the development of the object-detection field. In line with this trend, we further explore the potential and methods of model detection for unknown-category targets in open-world scenarios. By innovatively introducing a causal-reasoning mechanism, we re-examine the relationship between object features and category determination, attempting to fundamentally solve the dilemma of inaccurate identification of unknown-category targets and open a new path for open-world object detection. A large amount of experimental data shows that the causality-driven open-world object-detection framework proposed by us outperforms most current open-world object-detection models in terms of identifying unknown objects and continuously learning new categories. Our contributions mainly cover the following four key aspects:

We significantly improve the quality and expressiveness of image features by introducing a feature enhancement module.

By establishing a similarity calculation mechanism to accurately compute the similarity between text and images, we effectively prevent the over-confidence phenomenon in model prediction, thus enhancing the accuracy and reliability of detection.

By adopting a multi-granularity visual stream, we conduct multi-dimensional and refined feature processing on image features, fully exploring the multi-level information within images.

We conducted extensive experiments on multiple popular benchmarks, demonstrating the effectiveness of our method. It outperforms many current open-world methods in the detection of unknown classes.

# 2. Related Work

## 2.1. Out-of-Distribution Detection

Out-of-Distribution Detection (OOD for short) is a crucial issue in the field of machine learning. The core task of OOD detection is to accurately determine the consistency between the input data and the distribution of training data, to clarify whether the data belongs to the In-Distribution or comes from an undetermined Out-of -Distribution category. In many studies, MDS [8] uses the minimum Mahalanobis distance to measure the distance between class centroids. VIM [9] combines the norm of feature residuals, the principal space composed of training features, and the original logits to calculate OOD-Ness. [10] proposed an open-set object detection method named OpenDet. It expands the low-density latent regions through two learners, the Contrastive Feature Learner (CFL) and the Unknown Probability Learner (UPL), to achieve the identification of unknown objects. CFL expands the low-density regions by performing contrastive feature learning on known categories, and UPL learns the unknown probability of each instance as the threshold for dividing the low-density regions. [11] proposed an energy-based out-of-distribution detection for Graph Neural Networks (GNNs). It uses energy-based modeling and energy-based belief propagation for OOD detection in (semi -) supervised node classification. [12] proposed an innovative object detection method called Unknown Sniffer, which can simultaneously achieve accurate identification of known-category objects and effective detection of unknown-category objects. Its core innovations include the following two aspects: First, the authors designed a Generalized Objectness Confidence (GOC) scoring mechanism. This mechanism only relies on the labeled data of known categories for supervised learning, thus avoiding the problem of mis-suppression of unknown objects caused by over-reliance on the features of known categories in traditional methods. Second, in order to further reduce the interference of non-object samples in the background area, the authors proposed a Negative Energy Suppression Loss (NESL). By explicitly constraining the response intensity of the model to the background area, it significantly improves the detection sensitivity of unknown objects and the overall detection performance.

## 2.2. Open-World Object Detection

The concept of Open -World Object Detection was first proposed by Joseph et al. [13] in 2021. They designed a module named ORE, which serves as the core solution for distinguishing between known-category and unknown category objects. This module innovatively introduces the idea of an energy model, determining whether a sample belongs to an unknown category by calculating its energy value. Meanwhile, to further enhance the discriminative ability for unknown objects, ORE also incorporates the idea of contrastive clustering. By reducing the intra-class distance of known categories and increasing the distance between unknown categories and known categories in the feature space, it achieves effective identification and separation of unknown objects. This work has laid an important theoretical foundation for the field of open-world object detection. After that, OWSOL [14] constructed an evaluation benchmark for open-world weakly-supervised object localization, and proposed supervised and semantic-centroid -driven contrastive co-learning methods. RegionSpot [15] utilized pre-trained vision and vision-language foundation models to achieve region-level object recognition. [16] introduced an open-world object detection method named CapDet. This method unifies the dense captioning task and the open-world detection task into one framework, enabling the model to perform object detection given a category list and directly generate language descriptions for the predicted new-concept objects.

# 3. Proposed Method

This section will introduce the proposed framework in detail. In Section 3.1, the overall scheme of our proposed framework will be described. In Sections 3.2, 3.3, and 3.4, the detailed processes of our multi -granularity visual stream, feature enhancement module, and the construction of the similarity calculation mechanism will be presented respectively. Section 3.5 will describe the overall training and inference strategies.

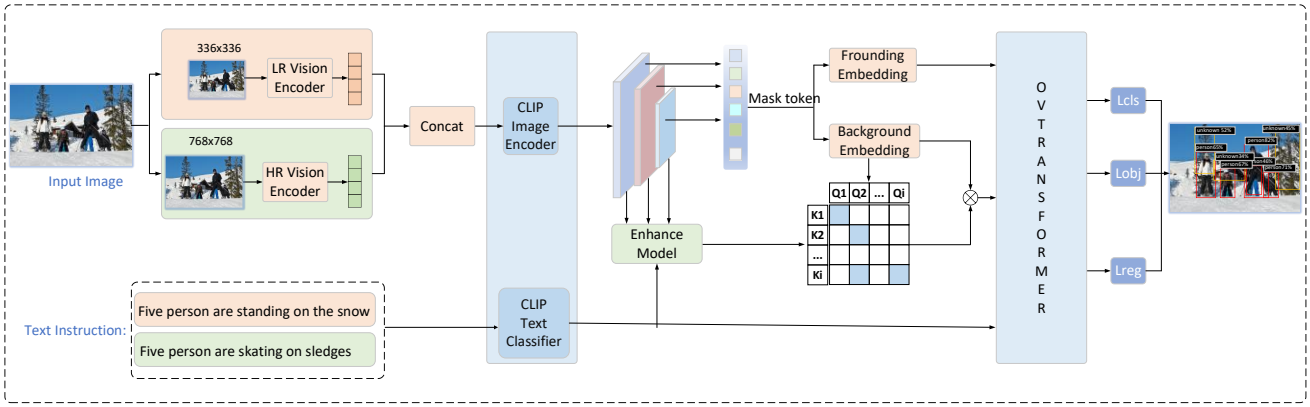


**Figure 1.** Overall Architecture.

The overall architecture is a multi-modal network. We introduce the multi-granularity visual stream module to process the detailed information of images. Meanwhile, to strengthen the correlation between images and text, an image enhancement module is proposed. Finally, to reduce the over-confidence problem, a similarity calculation between text and background is proposed to separate high-confidence instance objects from the background. The extracted image and text features are jointly input into the Transformer for detection.

## 3.1. Overall Architecture

As shown in Figure 1, first, we pass the image through the multi-granularity visual stream module. This module separates the input image into a low-resolution version and a high-resolution version, aiming to comprehensively capture the feature information of the image from different resolution perspectives. Images of different resolutions can reflect the macroscopic structure and microscopic details of the image, providing a rich data basis for subsequent feature extraction. Subsequently, the information of these two images with different resolutions is extracted respectively, and the extracted features are concatenated to obtain an image feature representation containing multi-scale features.

In terms of text processing, we use the text description of the image content as prompt information. This text information, together with the processed image features, enters the subsequent processing flow. This co-processing method of text and image can utilize the text information to supplement and guide the image features, enabling the model to better understand the semantic content in the image.

The image features processed and concatenated by the multi-granularity visual stream enter the feature pyramid structure. The feature pyramid [30] can further integrate image features of different scales and enhance the expressive ability of features. Then, we perform feature enhancement on the image information output by the feature pyramid and the text information. In this way, the correlation between the image features and the text information is strengthened, enabling the image and text to complement and confirm each other, thus improving the understanding and expression ability of the entire model for the image content. During the detection process, we divide the features into foreground features and background features. However, in actual detection, the model may be over-confident, resulting in the foreground instance information being wrongly included in the background. To solve this problem, we calculate the similarity between the enhanced features and the background features. By setting an appropriate threshold, the instance objects with high confidence are accurately separated from the background. Finally, we input the separated foreground information and the high-confidence instance object information extracted from the background into the Transformer [17] model. The powerful information interaction ability of the Transformer can deeply integrate and analyze these features, and explore the potential relationships between the features. After being processed by the Transformer, the model outputs the final detection result, providing an accurate judgment for object detection and recognition in the image.

## 3.2. Multi-granularity Visual Stream Module

This approach of multi-granularity visual stream not only effectively balances the relationship between details and the overall view but also enables precise feature extraction and understanding in complex visual scenes. For instance, in object detection tasks, the low-granularity feature stream can accurately locate the contours and boundaries of objects, while the high-granularity feature stream can provide object category information and contextual semantics, thus helping the model identify and classify targets more accurately. In image segmentation tasks, the combination of multi-granularity feature streams can better handle complex structures in images. For example, in medical image analysis, low -granularity features can capture the fine structures of cells, and high-granularity features can assist in identifying the types of tissues and diseased areas. Furthermore, the architecture of the multi-granularity visual stream is highly adaptable. It can dynamically adjust the granularity of feature extraction according to different visual tasks. For example, when processing high-resolution images, it can increase the depth of low-granularity feature extraction to capture more detailed information. When dealing with low-resolution images, it can focus on the extraction of high-granularity features to quickly grasp the overall semantics of the image. This flexibility enables the multi-granularity visual stream to perform well in a variety of visual tasks, providing strong support for fields such as object recognition, scene understanding, and image generation. With the continuous development of deep -learning technology, the implementation methods of multi-granularity visual streams are also constantly being optimized [18]. For example, by introducing the attention mechanism, the model can automatically focus on the most informative parts of different-granularity features, further improving the efficiency and accuracy of feature extraction. At the same time, multi-scale feature fusion techniques are also evolving. Through the design of more reasonable feature fusion strategies, such as weighted fusion and feature interaction, it can better integrate feature information of different granularities, providing a more comprehensive and high-

quality feature representation for subsequent visual tasks. Therefore, as an advanced visual information processing framework, the multi -granularity visual stream can effectively capture rich features from details to the overall view by decomposing and processing image data at multiple scales and levels. This multi-level and multi-scale feature extraction method not only provides strong support for complex visual tasks but also offers new ideas and directions for the future development of visual technology.

The initial step of the multi-granularity visual stream is to decompose the input image into feature maps of different resolutions, to obtain features of different granularities. Suppose the input image is $I \in \mathbb{R}^{H\times W\times C}$, where H represents the image height, $W$ represents the image width, and Crepresents the number of image channels. We use a series of convolutional kernels $K^l \in \mathbb{R}^{k^l\times k^l\times C\times C^l}$ to extract the feature map of the $l$-th layer. Here, $K^l$ is the size of the convolutional kernel in the $l$-th layer, and $C^l$ is the number of channels of the feature map in the $l$ -th layer. The convolutional operation can be expressed as:

$$F^l = \mathrm{Conv}(I, K^l) + b^l \tag{1}$$

Where $\mathrm{Conv}(\cdot)$ represents the convolution operation, and $b^l \in \mathbb{R}^{C^l}$ is the bias term of the $l$-th layer. Usually, after convolution, an activation function $\sigma(\cdot)$ (such as ReLU) is applied to introduce non- linearity:

$$\hat{F}^l = \sigma(F^l) \tag{2}$$

The feature maps $\hat{F}^l$ of different layers correspond to features of different granularities. The lower the layer number, the higher the resolution of the feature map, which contains low-granularity detailed information. The higher the layer number, the lower the resolution of the feature map, which contains high-granularity overall information.

The core of the multi-granularity visual stream is to fuse features of different granularities to obtain more comprehensive information. In this paper, the adopted fusion method is to perform element – by-element addition on feature maps of different resolutions. That is, when fusing the feature maps of the i-th layer and the j-th layer, the fused feature map is:

$$F^1_{merge} = \alpha_i \cdot U^i + \alpha_j \cdot D^j \tag{3}$$

Where $\alpha_i$ and $\alpha_j$ are learnable weights, which are used to adjust the contributions of different feature maps.

After feature fusion, the fused feature map is usually further processed through one or more convolutional layers to obtain the final output features. Suppose we use a convolutional kernel $K^{out} \in \mathbb{R}^{k^{out}\times k^{out}\times C_{merge}\times C_{out}}$ to perform a convolution operation on the fused feature map. The final output features are:

$$O = \mathrm{Conv}\left(F_{merge}, K^{out}\right) + b^{out} \tag{4}$$

Among them, $b^{out} \in \mathbb{R}^{C_{out}}$ is the bias term, $C_{merge}$ is the number of channels of the fused feature map, and $C_{out}$ is the number of channels of the final output features. Similarly, an activation function can also be applied to process the final output:

$$\hat{O} = \sigma(O) \tag{5}$$

Through the above steps, including feature extraction, down-sampling, up-sampling, feature fusion, and final processing, effective processing of the image at multiple scales and levels is achieved.

The effective multi-scale and multi-level image processing method described above has significant advantages. It greatly enhances the model's efficiency in utilizing image information. Traditional image processing methods often only focus on features at a single scale, easily overlooking important information at other scales. In contrast, multi-scale processing technology can comprehensively utilize features at different scales, enabling the model to understand the image content more comprehensively and in-depth. In practical applications, image data is often affected by various noises and interferences, such as changes in lighting, blurring, and occlusion. Multi-scale processing can analyze and process images at different scales, making the model more resistant to these noises and interferences. Even when part of the image is occluded or noisy, the model can still accurately identify the target through features at other scales.

## 3.3. Feature Enhancement Module

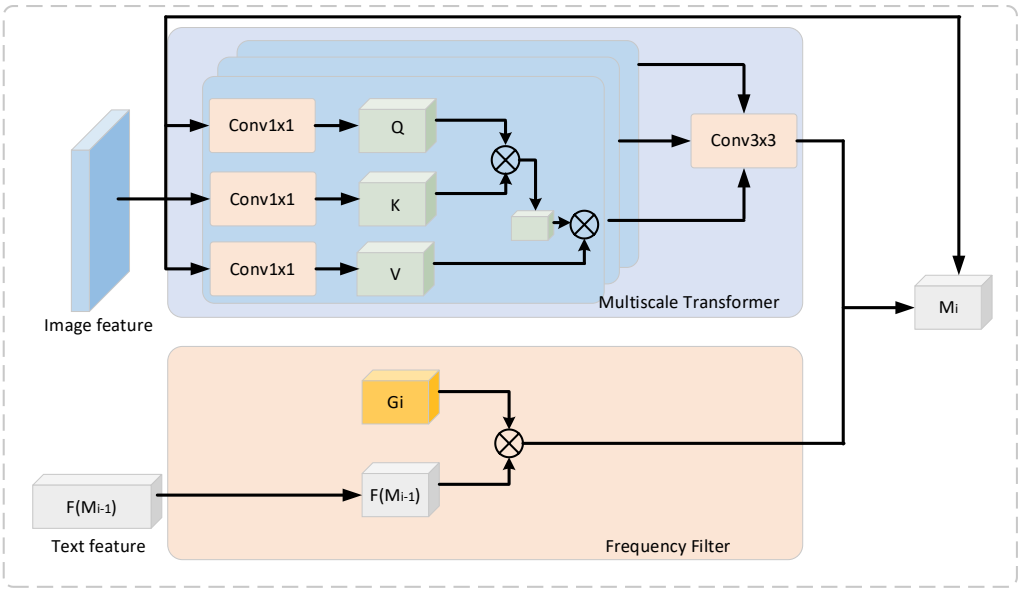


**Figure 2.** Feature Enhancement Module

The cross-modal feature enhancement module proposed in this paper constructs a hierarchical feature enhancement

framework by fusing spatial domain and frequency domain features. As shown in Figure 2, the module first maps the image features to a query vector $Q \in \mathbb{R}^{H\times W\times C'}$ through an independent convolutional layer, and at the same time encodes the frequency domain features into a key vector K and a value vector $V$ respectively. By calculating the similarity matrix $QK^T$ and applying a scaling factor $\sqrt{d_k}$, the module generates a cross-modal attention weight map $M \in \mathbb{R}^{(HW)\times(HW)}$, which dynamically adjusts the contribution of frequency domain features to spatial features through a soft attention mechanism. The specific enhancement process can be described as follows:

1) Cross-modal feature calibration: The directional injection of frequency domain features into the spatial domain is achieved through matrix multiplication $Z = MV$. The weight coefficient of high-frequency components in the edge region is twice that in the background region, significantly enhancing the expressive ability of detailed features.

2) Residual feature fusion [31]: A skip connection is used to linearly superimpose the enhanced feature $Z$ with the original image features, which not only preserves the integrity of the low-level features but also avoids the gradient vanishing problem.

3) Nonlinear representation optimization: A convolutional layer containing the LeakyReLU [32] activation function is used to perform a nonlinear transformation on the fused features, effectively improving the compactness of the feature distribution.

To enhance the image features, we introduce a multi-scale Transformer, which can operate on the input image features. It takes the output of the previous state's fusion block $M_{i-1}$ as input, divides it into spatial image patches of different sizes, and computes block-wise self-attention in different heads. Specifically, we first extract image patches of shape $r_h \times r_h \times C$ from $M_{i-1}$, and reshape them into one-dimensional vectors for the $h$- th head. Then, we use a fully-connected layer to embed the flattened vectors into the query embedding $Q_i^h \in \mathbb{R}^{N\times C_h}$, where $N = (H/4r_h) \times (W/4r_h)$ and $C_h = r_h \times r_h \times C$. We obtain the key embedding $K_i^h$ and value embedding $V_i^h$ through similar operations. Then, we calculate the attention matrix through the following processes respectively:

$$A_i^h = softmax\left(\frac{Q_i^h(K_i^h)^T}{C_h}\right)V_i^h \tag{6}$$

$A_i^h$ represents the result that is resized to the original spatial resolution. Finally, the features from different heads are concatenated and further passed through a residual block to obtain the output $T_i \in \mathbb{R}^{(H/4)\times(W/4)\times C}$.

For text, we adopt frequency-domain filtering, mainly because there are many words in large-scale text datasets. Among them, some high-frequency words may be functioning words without actual semantics (such as function words like “of”, “is”, “at” in English equivalent sense), and some low-frequency words may be spelling mistakes, rare words, or words that only appear once or twice in a specific context. Through frequency filtering, we can remove these high-frequency and low-frequency words that have little impact on the overall semantics, thereby significantly reducing the number of words that need to be processed. For example, when processing a news text dataset, after filtering out function words and low-frequency rare technical terms, the size of the vocabulary may be reduced by several times. This greatly reduces the computational complexity and time cost of subsequent processing (such as text classification and clustering). After filtering out high-frequency meaningless words and low-frequency noise words, the remaining words are often more crucial for text semantic expression. These words can more accurately reflect the theme and core content of the text. Moreover, an excessive number of noise words and meaningless words may cause the model to learn some unnecessary features, thus increasing the risk of overfitting. Frequency filtering can remove these noises, enabling the model to focus more on meaningful features and improving the model's generalization ability. Additionally, through frequency filtering, we can reduce the uncertainty and volatility in the data, allowing the model to maintain relatively stable performance when facing different datasets or data distribution changes.

## 3.4. Similarity Calculation Mechanism

To mitigate the problem of over-confidence, a similarity calculation method between text and background is proposed to separate instance objects with high confidence from the background. The similarity calculation between text and background is based on the deep fusion of multi-modal information. First, for the input image, advanced feature extraction networks such as Convolutional Neural Networks (CNN) [19] are used to obtain the visual features of the image. These features contain information about the color, texture, and shape of each region in the image, comprehensively describing the visual content of the image. Meanwhile, for the text information describing the image content, natural language processing techniques such as pre-trained language models (e.g., BERT [20]) are employed to convert the text into semantic vectors. These semantic vectors can capture the semantic information in the text, reflecting the themes and concepts expressed by the text. Next, to calculate the similarity between the text and the image background, a method based on the attention mechanism is adopted. The attention mechanism can automatically focus on the relevant parts in the text and the image background and assign different weights to each part. Specifically, by calculating the correlation between the text semantic vector and the image background feature vector, an attention map is obtained. This attention map represents the degree of correlation between each word in the text and each region in the image background. Then, according to the attention map, the image background features are weighted and summed to obtain a background feature vector that integrates the text information. Finally, by calculating the similarity between this fused background feature vector and the feature vectors of each instance object in the image, the matching degree between each instance object and the text description is determined. Multiple methods can be used for similarity calculation, such as cosine similarity and Euclidean distance. Cosine similarity can measure the directional similarity between two vectors, while Euclidean distance can measure the spatial distance between two vectors. Based on the calculated similarity values, a confidence score can be assigned to each instance object. In this paper, the first method is selected.

Based on the confidence scores obtained from the similarity calculation between text and background, we can separate instance objects with high confidence from the background. The specific steps are as follows: First, set a confidence threshold $T$. This threshold can be adjusted according to the specific application scenario and task

requirements. For tasks with high accuracy requirements, such as medical image diagnosis and autonomous driving, a relatively high threshold can be set. For tasks with high recall requirements, such as information retrieval and image annotation, a relatively low threshold can be set. Then, traverse all instance objects in the image, and mark the instance objects with confidence scores higher than the threshold as high-confidence instance objects. These instance objects are highly matched with the text description and have high reliability. To further improve the accuracy of the separation, some post-processing methods can be adopted. For example, use the spatial relationships and contextual information among instance objects for screening. If two high-confidence instance objects are too close in space and there are conflicts in their semantic information, we can make a judgment based on the contextual information and select the object that better matches the text description.

First, we map the input features to a common semantic space.

$$Q = W_q \cdot X_{query} \in \mathbb{R}^{B\times L_q\times D} \tag{7}$$

$$K = W_k \cdot X_{key} \in \mathbb{R}^{B\times L_k\times D} \tag{8}$$

Among them, $W_q, W_k \in \mathbb{R}^{D\times D}$ are learnable parameter matrices, and then the dot-product similarity is calculated.

$$\mathrm{Sim}(Q,K) = \frac{QK^T}{\sqrt{d_k}} \in \mathbb{R}^{B\times L_q\times L_k} \tag{9}$$

Among them, $d_k$ is the dimension of the key vector. Then, a Softmax function with temperature scaling (set to 0.1) is used.

$$A_{ij} = \mathrm{Softmax}\left(\frac{\mathrm{Sim}(Q_i,K_j)}{\tau}\right) = \frac{\exp\left(\frac{\mathrm{Sim}(Q_i,K_j)}{\tau}\right)}{\sum_{k=1}^{L_k}\exp\left(\frac{\mathrm{Sim}(Q_i,K_k)}{\tau}\right)} \tag{10}$$

$$Confidence_{ij} = A_{ij} \tag{11}$$

Among them, $\tau$ is the temperature hyperparameter. Next, we will filter out the high-confidence instance objects. For the $i$- th query (related to the text) and the j- th key (related to the instance object), if $Confidence_{ij} > T$ then this instance object is marked as a high-confidence instance object. Otherwise, it is treated as the background. Through the above steps, the model can separate the instance objects with high-confidence scores from the background, avoiding prediction errors caused by the model's over-confidence.

## 3.5. Training and Inference Strategies

In this paper, the following joint loss function is used for end-to-end training:

$$L = \alpha L_{cls} + \beta L_{reg} + \gamma L_{sim} \tag{12}$$

Among them, $L_{cls}$, $L_{reg}$, and $L_{sim}$ represent the classification, regression, and similarity loss functions respectively. $L_{cls}$ uses the focal loss [33] to address the class imbalance problem. $L_{reg}$ uses the smooth L1 regression loss, and $L_{sim}$ is the similarity loss function that measures the degree of difference between the features of instance objects and background features. $\alpha$, $\beta$ and $\gamma$ represent the corresponding coefficients respectively. Through the above joint training, unknown object instances can be detected more effectively.

# 4. Experiments

## 4.1. Datasets and Metrics

In this section, we conduct many experiments on two datasets widely used in current open-world object detection to verify the effectiveness of the proposed method. We mainly verify the effectiveness of our method on the S-OWODB and M-OWODB datasets [21, 22]. The specific division methods are shown in Table 1 and Table 2. Under the standard detection evaluation system, we select the mAP (mean Average Precision) as the metric to measure the detection performance of known classes. mAP can comprehensively reflect the accuracy of the model in the known-class object detection task. For the evaluation of the detection ability of unknown classes, we use the Unknown Object Recall (U-Recall) as the core metric. U-Recall intuitively reflects the model's ability to recognize unknown objects. The higher its value, the better the model's performance in discovering unknown objects. In addition, we also introduce the metric WI [23] to evaluate the model. WI is used to measure the degree to which unknown objects are wrongly confused with known objects. According to relevant research and the unified settings of this experiment, the value of WI is fixed at 0.8 in all experiments.

$$\mathrm{WI} = \frac{P_{\mathcal{K}}}{P_{\mathcal{K}\cup\mathcal{U}}} - 1 \tag{13}$$

Among them, $P_{\mathcal{K}}$ is the prediction for known classes, and $P_{\mathcal{K}\cup\mathcal{U}}$ is the prediction for both known and unknown classes.

**Table 1.** Division method of the M-OWODB dataset

| | Task1 | Task2 | Task3 | Task4 |
|---|---|---|---|---|
| Semantic split | Animals, Person | Appliances, Outdoor | Sports, Food | Indoor, Kitchen |
| Training images | 89,490 | 55,870 | 39,402 | 38,903 |
| Train instances | 421,243 | 163,512 | 114,452 | 160,794 |
| Test images | 4,952 | | | |
| Test instances | 36,781 | | | |

**Table 2.** Division method of the S-OWODB dataset

| | Task1 | Task2 | Task3 | Task4 |
|---|---|---|---|---|
| Sementic split | VOC Classes | Outdoor, Truck | Sports, Food | Indoor, Kitchen |
| Train images | 16,551 | 45,520 | 39,402 | 40,260 |
| Train instances | 47,223 | 113,741 | 114,452 | 138,996 |
| Test images | 4,952 | 1,914 | 1,642 | 1,738 |
| Test instances | 14,976 | 4,966 | 4,826 | 6,039 |

## 4.2. Implementation Details

The method in this paper uses CLIP based on VIT-L/14 as the backbone network. In the image encoding stage, the adopted architecture consists of 6 encoder layers and 6 decoder layers. Meanwhile, VLM is selected as the text encoder. The entire experimental process includes 4 incremental learning tasks. Regarding the setting of training parameters, we use 6 GPUs for training. The batch size of each GPU is set to 8. The AdamW optimizer is chosen, and its learning rate is set to 1e-6.

## 4.3. Main Reaults

**Table 3.** Shows the comparison with the current state-of-the -art Open-World Object Detection (OWOD) models on the M-OWODB dataset.

| Task IDS → | Task 1 | | | Task 2 | | | | | Task 3 | | | | | Task 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| metrics → | WI (↓) | U-Recall (↑) | K-mAP (↑) | WI (↓) | U-Recall (↑) | K- mAP (↑) | | | WI (↓) | U-Recall (↑) | K- mAP (↑) | | | K- mAP (↑) | | |
| | | | Curr | | | Prev | Curr | Both | | | Prev | Curr | Both | Prev | Curr | Both |
| ORE-EBUI [13] | 0.062 | 4.9 | 56 | 0.028 | 2.9 | 52.7 | 26 | 39.4 | 0.021 | 3.9 | 38.2 | 12.7 | 29.7 | 29.6 | 12.4 | 25.3 |
| OW-DETR [24] | 0.057 | 7.5 | 59.2 | 0.028 | 6.2 | 53.6 | 33.5 | 42.9 | 0.015 | 5.7 | 38.3 | 15.8 | 30.8 | 31.4 | 17.1 | 27.8 |
| 2B-OCD [25] | 0.048 | 12.1 | 56.4 | 0.016 | 9.4 | 51.4 | 25.3 | 38.5 | 0.014 | 11.7 | 37.2 | 13.2 | 29.2 | 30.1 | 13.3 | 25.8 |
| PROB [26] | 0.057 | 19.4 | 59.5 | 0.034 | 17.4 | 55.7 | 32.2 | 44 | 0.015 | 19.6 | 43 | 22.2 | 36 | 35.7 | 18.9 | 31.5 |
| CAT [27] | 0.066 | 23.7 | 60 | 0.032 | 19.1 | 55.5 | 32.7 | 44.1 | 0.020 | 24.4 | 42.8 | 18.7 | 34.8 | 34.4 | 16.6 | 29.9 |
| MEPU-FS [28] | 0.056 | 31.6 | 60.2 | 0.023 | 30.9 | 57.3 | 33.3 | 44.8 | 0.016 | 30.1 | 42.6 | 35.4 | 35.4 | 34.8 | 19.1 | 30.9 |
| MEPU-SS [28] | 0.057 | 30.3 | 60 | 0.023 | 30.6 | 57 | 33.1 | 44.5 | 0.016 | 30.0 | 42.2 | 35 | 35 | 34.3 | 18.9 | 30.4 |
| Ours | 0.029 | 35.1 | 63.0 | 0.004 | 33.3 | 56.4 | 35.2 | 45.5 | 0.015 | 33.1 | 43.5 | 34.8 | 37.5 | 36.3 | 21.2 | 32.5 |

To evaluate the detection accuracy of the algorithm for known classes, we select K-mAP as the evaluation metric for known classes. Among them, “Prev” is used to measure the detection accuracy of the model for known classes in previous tasks; “Curr” reflects the accuracy of the model in detecting known classes in the current task. For the detection effect of unknown classes, we use the U-Recall metric to measure the recall rate, that is, the model's ability to detect unknown classes. It should be noted that in Task 4, all the original unknown classes have been learned and classified as known classes, so the U-Recall metric is not involved in the evaluation of Task 4.

**Table 4.** Shows the comparison with the current state-of-the-art Open-World Object Detection (OWOD) models on the S-OWODB dataset.

| Task IDS → | Task 1 | | | Task 2 | | | | | Task 3 | | | | | Task 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| metrics → | WI (↓) | U-Recall (↑) | K-mAP (↑) | WI (↓) | U-Recall (↑) | K- mAP (↑) | | | WI (↓) | U-Recall (↑) | K- mAP (↑) | | | K- mAP (↑) | | |
| | | | Curr | | | Prev | Curr | Both | | | Prev | Curr | Both | Prev | Curr | Both |
| ORE-EBUI [13] | 0.024 | 1.5 | 61.4 | 0.040 | 3.9 | 56.5 | 26.1 | 40.6 | 0.026 | 3.6 | 38.7 | 23.7 | 33.7 | 33.6 | 26.3 | 31.8 |
| OW-DETR [24] | 0.029 | 5.7 | 71.5 | 0.041 | 6.2 | 62.8 | 27.5 | 43.8 | 0.025 | 6.9 | 45.2 | 24.9 | 38.5 | 38.2 | 28.1 | 33.1 |
| RandBox [29] | 0.009 | 2.4 | 53.2 | 0.005 | 1.8 | 45.3 | 35.1 | 40 | 0.004 | 5.5 | 42.5 | 25.2 | 36.7 | 36.4 | 39.5 | 37.2 |
| Ours | 0.020 | 7.1 | 70.5 | 0.004 | 8.8 | 58.8 | 38.6 | 45.3 | 0.025 | 11.3 | 46.7 | 27.3 | 40.2 | 40.2 | 42.3 | 39.5 |

#### 4.3.1. OWOD Split

Table 3 presents a comprehensive comparison between the method proposed in this paper and other current methods on the COCO benchmark. Thanks to the advanced techniques proposed, the method in this paper outperforms most existing models in terms of the performance of detecting unknown objects. In Tasks 1, 2, and 3, the unknown object recalls rate (U-Recall) of the latest method MEPU [23] is 31.6%, 30.9%, and 30.1% respectively. In contrast, the U-Recall of the model in this paper reaches 35.1%, 33.3%, and 38.1% respectively. On the key indicator of U-Recall, which measures the ability to detect unknown objects, the model in this paper achieves an improvement of up to 3% compared with MEPU [23]. This result fully demonstrates that our model could effectively detect known objects in previous tasks, showing strong advantages in the open-world object detection scenario.

#### 4.3.2. MS-COCO Split

Compared with the M-OWODB dataset, the detection difficulty of the MS-COCO Split increases. The model faces greater challenges and thus has higher requirements. Table 4 shows the comparison between our method and other methods. The unknown object recall rates of the latest method RandBox are 2.4%, 1.8%, and 5.5% respectively. In contrast, our method achieves 7.1%, 8.8%, and 11.3% for the detection of unknown classes. Compared with RandBox, it has an effective improvement of 7%, which proves that our method has strong generalization ability and robustness.

## 4.4. Ablation Study

To verify the effectiveness of the proposed algorithm, we conducted extensive ablation experiments for analysis and verification.

**Table 5.** Component ablation experiments

| Task IDS → | | | Task 1 | | Task 2 | | | | Task 3 | | | | Task 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MGVS | FEM | SCM | U-Recall (↑) | K-mAP(↑) Curr | U-Recall (↑) | K-mAP (↑) Prev | Curr | Both | U-Recall (↑) | K-mAP (↑) Prev | Curr | Both | K-mAP (↑) Prev | Curr | Both | |
| × | × | × | 29.6 | 56.1 | 29.6 | 51.1 | 32.7 | 42.9 | 31.1 | 38.9 | 30.6 | 34.1 | 33.9 | 18.5 | 29.3 | |
| ✓ | × | × | 31.8 | 61.7 | 30.1 | 52.3 | 33.3 | 43.8 | 31.1 | 41.1 | 31.6 | 35.1 | 35.9 | 19.2 | 30.1 | |
| ✓ | ✓ | × | 32.9 | 62.4 | 31.2 | 53.1 | 33.8 | 44.3 | 32.1 | 42.1 | 32.4 | 35.6 | 36.9 | 20.1 | 30.9 | |
| ✓ | × | ✓ | 33.5 | 62.8 | 32.1 | 53.9 | 34.3 | 44.6 | 32.4 | 42.9 | 33.4 | 36.9 | 36.1 | 20.4 | 31.8 | |
| ✓ | ✓ | ✓ | 35.1 | 63.0 | 33.3 | 56.4 | 35.2 | 45.5 | 33.1 | 43.5 | 34.8 | 37.5 | 36.3 | 21.2 | 32.5 | |

#### 4.4.1. Ablating Components

As shown in Table 5, to verify the function of each module, we designed a set of control experiments. First, we divided the entire system into multiple independent modules and clarified the function and expected output of each module. For each module, we set up an experimental group and a control group. The experimental group included the complete implementation of the module, while the control group removed the module, with other conditions remaining the same. When we added both MGVS and FEM, the detection accuracy of the model for both known and unknown objects improved. This proves that it is effective for us to extract image features at different resolutions. It shows that the high-granularity features obtained at low resolution can capture the overall structure and semantic information of the image, providing the model with a macroscopic understanding of the scene. Meanwhile, the low-granularity features at high resolution focus on image details such as edges and textures, which helps the model accurately identify the specific features of the target. FEM further strengthens these multi-resolution features, enabling better cooperation and complementation between features of different granularities. When we combined MGVS and SCM, there was also an improvement compared to the Baseline. When we used all three modules simultaneously, the performance of the model reached the optimal level. The experimental data fully proves the effectiveness and superiority of the method proposed in this paper.

**Table 6.** Feature layer ablation experiments.

| $l$ | mAP (↑) | U-Recall (↑) | WI (↓) |
|---|---|---|---|
| 2.0 | 62.3 | 34.5 | 0.034 |
| 4.0 | 62.7 | 34.8 | 0.033 |
| 6.0 | 63.0 | 35.1 | 0.029 |
| 8.0 | 62.8 | 34.9 | 0.031 |

Ablation experiments on the number of feature layers in the multi-granularity visual stream were conducted on the M-OWODB dataset.

**Table 7.** Scaling factor ablation experiments.

| $d_k$ | mAP (↑) | U-Recall (↑) | WI (↓) |
|---|---|---|---|
| 36 | 62.6 | 34.3 | 0.033 |
| 49 | 62.8 | 34.7 | 0.032 |
| 64 | 63.0 | 35.1 | 0.029 |
| 81 | 62.5 | 34.4 | 0.035 |

Ablation experiments on the scaling factor $d_k$ in the feature enhancement module were conducted on the M-OWODB dataset.

**Table 8.** Settings of the temperature hyperparameter $\tau$.

| $\tau$ | mAP (↑) | U-Recall (↑) | WI (↓) |
|---|---|---|---|
| 0.01 | 62.6 | 34.3 | 0.033 |
| 0.10 | 63.0 | 35.1 | 0.029 |
| 0.15 | 62.8 | 35.0 | 0.028 |
| 0.20 | 62.5 | 34.4 | 0.035 |

#### 4.4.2. Hyperparameter Ablation

In Table 6, we verified the feature layers involved in the multi-granularity visual stream. Different numbers of feature layers lead to differences in image feature extraction. We conducted comparative experiments with different numbers of layers. Through the experiments, we found that the best effect was achieved when the number of feature layers was 6, which proves that feature extraction is more sufficient at this time. However, as the number of layers increases, the effect declines, indicating that the model may have overfitted. We also conducted an ablation experiment on the scaling factor in the feature enhancement model, as shown in Table 7. This is mainly because different dimensions of the key vector may represent different feature information. When the dimension is high, if no scaling is performed, the features in certain dimensions may have an excessive impact on the attention scores, thus masking the information in other dimensions. The introduction of the scaling factor can balance the features of each dimension, enabling each dimension of features to play a relatively reasonable role in calculating the attention scores. This can prevent the model from over-focusing on the features of certain dimensions and ignoring other important information, thereby improving the model's comprehensive utilization ability of features and enhancing the feature enhancement effect. In Table 8, we continuously changed the temperature hyperparameter $\tau$. When $\tau$ is relatively large, the probability distribution output by the Softmax function becomes smoother, the probability values of each category are closer, and the model's prediction results are more "uncertain". When the value of $\tau$ is relatively small, the probability distribution output by the Softmax function becomes sharper, the model has a higher confidence in a certain category, and the prediction results are more "certain". By continuously adjusting the value of $\tau$, we found that the best effect was achieved when $\tau = 0.10$.

## 4.5. Qualitative Results

Figures 3 and 4 show the comparison of the visualization results between our method and other methods on the M-OWODB and S-OWODB datasets. In the visual comparison, it can be clearly seen that for the target detection and segmentation tasks of known classes, some other methods can accurately identify and label some common targets. However, previous methods often have the problem of inaccurate boundary positioning. There is a certain deviation between

the marked bounding boxes and the actual edges of the targets, and the details are handled rather roughly. In sharp contrast, our method performs much better in dealing with known-class targets. It can not only accurately identify the targets but also draw the bounding boxes that highly match the actual shapes of the targets. When it comes to unknown-class targets, the limitations of other methods become more prominent. Most traditional methods either misclassify unknown-class targets as known classes or directly ignore these targets when encountering them, resulting in many misjudgments and omissions in the visualization results. Thanks to its unique design and strong generalization ability, our method can effectively detect unknown-class targets and mark and distinguish them in a reasonable way. In the visualization results, our method can clearly circle the approximate range of unknown targets. Although it may not be able to accurately give their specific classes for the time being, it can provide valuable clues for subsequent analysis and research. This is of great significance in practical applications, as it can help users quickly discover newly-emerged target types and lay the foundation for further exploration and identification.

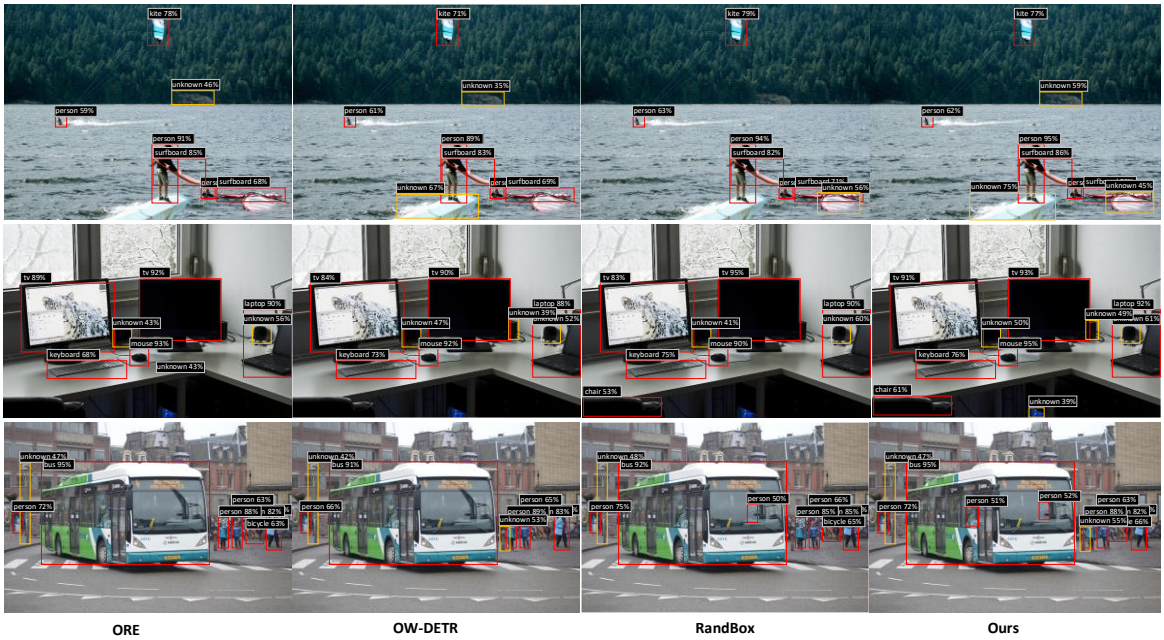


**Figure 3.** Demonstration of qualitative results on the M-OWODB dataset.


**Figure 4.** Demonstration of qualitative results on the S-OWODB dataset.

## 5. Conclusion

This paper focuses on the challenges in open-world object detection, where traditional models struggle to adapt to unknown objects and continuously learn new classes. A causality-driven open-world object detection framework is proposed. Through the design of modules, we construct a feature enhancement model to improve the quality and expressiveness of image features. A similarity calculation mechanism is built to avoid over-confidence during model prediction, enhancing the accuracy and reliability of detection. A multi-granularity visual stream is employed to process image features in multiple dimensions and at a refined level, excavating multi-level information. Experimental results on

benchmark datasets show that this method achieves significant performance gains in unknown-class detection tasks, demonstrating strong generalization ability. The research findings in this paper provide valuable references for the development of the open-world object detection field and are expected to promote technological innovation and progress in this area.

## Acknowledgments

This work was supported by Key Science and Technology Program of Henan Province (No.252102210091 and 252102220120), University Young Backbone Teachers Program of Henan Province (2023GGJS053), Fundamental Research Funds for the Universities of Henan Province (NSFRF220414), Excellent Young Teachers Program of Henan Polytechnic University (No.2019XQG - 02).

## References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[2] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Bharadwaj R, Naseer M, Khan S, et al. Enhancing Novel Object Detection via Cooperative Foundational Models [J]. arxiv preprint arxiv: 2311.12068, 2023.

[4] Qi H, Huang Z, ** B, et al. SAM-GAN: An improved DCGAN for rice seed viability determination using near-infrared hyperspectral imaging [J]. Computers and Electronics in Agriculture, 2024, 216: 108473.

[5] Heng A, Soh H. Selective amnesia: A continual learning approach to forgetting in deep generative models [J]. Advances in Neural Information Processing Systems, 2024, 36.

[6] Wu X, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection [J]. Neurocomputing, 2020, 396: 39-64.

[7] Pu Y, Liang W, Hao Y, et al. Rank-DETR for high quality object detection [J]. Advances in Neural Information Processing Systems, 2024, 36.

[8] Lee, Kimin, et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks." Advances in neural information processing systems 31 (2018).

[9] Wang, Haoqi, et al. "Vim: Out-of-distribution with virtual-logit matching." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[10] Han, Jiaming, et al. "Expanding low-density latent regions for open-set object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[11] Liu, Weitang, et al. "Energy-based out-of-distribution detection." Advances in neural information processing systems 33 (2020): 21464-21475.

[12] Liang, Wenteng, et al. "Unknown Sniffer for Object Detection: Don't Turn a Blind Eye to Unknown Objects." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[13] Joseph, K. J., et al. "Towards open world object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[14] Xie, Jinheng, et al. "Open-World Weakly-Supervised Object Localization." arXiv preprint arXiv:2304.08271 (2023).

[15] Yang, Haosen, et al. "Recognize any regions." arXiv preprint arXiv:2311.01373 (2023).

[16] Long, Yanxin, et al. "Capdet: Unifying dense captioning and open-world detection pretraining." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[17] Waswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//NIPS. 2017.

[18] Zhao X, Li X, Duan H, et al. Mg-llava: Towards multi-granularity visual instruction tuning [J]. arxiv preprint arxiv:2406.17770, 2024.

[19] Chen Y. Convolutional neural network for sentence classification [J]. 2015.

[20] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arxiv preprint arxiv:1810.04805, 2018.

[21] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[22] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge [J]. International journal of computer vision, 2010, 88: 303-338.

[23] Dhamija, Akshay, et al. "The overlooked elephant of object detection: Open set." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.

[24] Gupta, Akshita, et al. "Ow-detr: Open-world detection transformer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[25] Wu, Yan, et al. "Two-branch objectness-centric open world detection." Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis. 2022.

[26] Zohar Orr, Kuan-Chieh Wang, and Serena Yeung. "Prob: Probabilistic objectness for open world object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[27] Ma, Shuailei, et al. "Cat: Localization and identification cascade detection transformer for open-world object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[28] Fang, Ruohuan, et al. "Unsupervised Recognition of Unknown Objects for Open-World Object Detection." arXiv preprint arXiv:2308.16527 (2023).

[29] Wang, Yanghao, et al. "Random boxes are open-world object detectors." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[30] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[31] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[32] Xu B, Wang N, Chen T, et al. Empirical evaluation of rectified activations in convolutional network (2015) [J]. arxiv preprint

[33] Ross T Y, Dollár G. Focal loss for dense object detection[C]//proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2980-2988.