KTCN: Enhancing Open-World Object Detection with Knowledge Transfer and Class-Awareness Neutralization

Xing Xi, **Yangyang Huang**, **Jinhao Lin** and **Ronghua Luo***
South China University of Technology
xxyzll@yeah.net, huangyangy@whu.edu.cn, csljh_jasper@mail.scut.edu.cn, rhluo@scut.edu.cn,

Abstract

Open-World Object Detection (OWOD) has garnered widespread attention due to its ability to recall unannotated objects. Existing works generate pseudo-labels for the model using heuristic priors, which limits the model's performance. In this paper, we leverage the knowledge of the large-scale visual model to provide supervision for unknown categories. Specifically, we use the Segment Anything Model (SAM) to generate raw pseudo-labels for potential objects and refine them through Intersection over Union (IOU) and the shortest bounding box side length. Nevertheless, the abundance of pseudo-labels still exacerbates the competition issue in the one-to-many label assignment. To address this, we propose the **D**ual **M**atching **L**abel **A**ssignment (DMLA) strategy. Furthermore, we propose the **C**lass-**A**wareness **N**eutralizer (CAN) to reduce the model's bias towards known categories. Evaluation results on open-world object detection benchmarks, including MS COCO and Pascal VOC, show that our method achieves nearly **200%** the unknown recall rate of previous state-of-the-art (SOTA) methods, reaching **41.5** U-Recall. Additionally, our approach does not add any extra parameters, maintaining the inference speed advantage of Faster R-CNN, leading the SOTA methods based on deformable DETR at a speed of over **10** FPS. Our code is available at https://github.com/xxyzll/KTCN.

1 Introduction

Object detection, a fundamental task in computer vision, is applied across diverse real-world scenarios, including autonomous driving[Zou *et al.*, 2019; Solovyev *et al.*, 2021], medical diagnostics[Ozturk *et al.*, 2020; Loey *et al.*, 2021], and industrial safety[Zhou *et al.*, 2021; Guo *et al.*, 2022]. Traditional detection tasks[Zhang *et al.*, 2020; Tian *et al.*, 2019] commonly rely on training with closed datasets, anticipating the model to recognize objects already labeled within

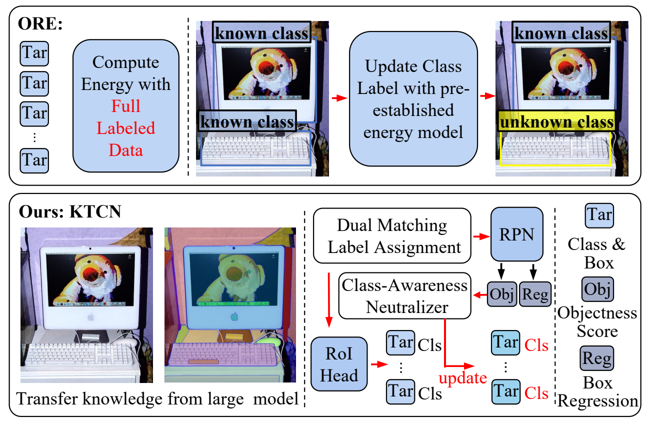

Figure 1: An outline of our method and ORE. The top section illustrates the energy in ORE and elucidates the process for refining category labels. At the bottom, a concise overview of the KTCN.

the dataset. However, it is impractical to annotate all objects in natural environments. Consequently, traditional detection models cannot identify unlabeled objects, limiting their application in real-world scenarios.

To tackle this, ORE[Joseph *et al.*, 2021] introduced the Open-World Object Detection (OWOD). The OWOD task consists of two main components: identifying potential objects and incremental learning. The former demands the model to recall as many latent objects as possible, while the latter expects the model to fine-tune existing knowledge at minimal cost to detect newly added objects. Unfortunately, the original implementation of ORE suffered from data leakage; the problem was also identified in OW-DETR[Gupta *et al.*, 2022], CAT[Ma *et al.*, 2023], and PROB[Zohar *et al.*, 2023].

ORE introduces an energy-based model to differentiate between known and unknown samples. Nevertheless, constructing the energy model necessitates access to fully annotated data, resulting in potential data leakage (Figure 1). Subsequent works recognized the issue and employed deformable DETR[Zhu *et al.*, 2020] as a baseline model, utilizing attention mechanisms[Gupta *et al.*, 2022], adaptive pseudo-labeling mechanisms[Ma *et al.*, 2023], and probabilistic objectness[Zohar *et al.*, 2023] to mine potential object samples. However, Faster R-CNN[Ren *et al.*, 2015] holds

*Corresponding Author

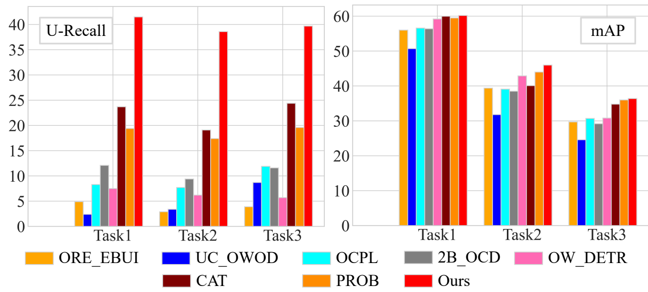

Figure 2: Unknown recall on standard OWOD of SOTA methods. The proposed method, KTCN, significantly surpasses all baselines and achieves comparable or even better performance on the known class.

an advantage[Dhamija *et al.*, 2020] in detecting unknown categories and has been adopted as a baseline network for open-vocabulary[Gu *et al.*, 2021; Zhong *et al.*, 2022], open-world[Wu *et al.*, 2022b; Zhao *et al.*, 2023] and novel category discovery[Zheng *et al.*, 2022; Fomenko *et al.*, 2022] object detection tasks. Therefore, we adopt Faster R-CNN as our baseline model in this paper.

Our exploration in OWOD centers on three critical challenges: (1) *how to identify potential unknown objects*, (2) *how to introduce supervision for these objects without reducing the model's performance on known categories*, and (3) *how to address the model's bias towards known objects*.

Addressing the initial concern, we utilize the zero-shot capability of the SAM (Segment Anything Model[Kirillov *et al.*, 2023]) to generate masks for all objects in the image, providing these masks as pseudo-labels to guide the model in identifying potential unknown objects. As shown in the bottom-left corner of Figure 1, we segment all objects in the image, obtaining class-agnostic masks for each object. Computing the maximum and minimum coordinates allows us to establish bounding boxes for the object. Following this, through a filtering process, we generate pseudo-labels for potential targets.

The second issue arises due to competition between known and unknown categories. To introduce supervision for potential objects, we utilize SAM to generate pseudo-labels. However, competition inevitably occurs during matching as the number of labels increases. Many predictions that could be matched to known instances are instead assigned to pseudo-labels, which affects the model's performance. To address this, we propose the **D**ual **M**atching **L**abel **A**ssignment (DMLA) method. Specifically, we consider that ground truth has fewer errors, and the model's performance on known categories entirely depends on the supervision of the available labels. Therefore, the matching of pseudo-labels must not affect the matching of ground truth labels. For the model's prediction, we prioritize matching the annotations of known categories. The unmatched portions are then assigned to pseudo-labels. Finally, we merge the results of both matchings. With this design, DMLA can resolve the competition between known and unknown categories and provide sufficient supervisory signals for training.

The issue of model bias is a significant concern, which is discussed in several works[Zhao *et al.*, 2023; Wang *et al.*, 2023b]. The OWOD task, the prediction is class-specific, leading to a bias towards known categories. To address this, we propose the **C**lass-**A**wareness **N**eutralizer (CAN), which utilizes the class-agnostic centerness scores from RPN to mitigate the model's bias towards known categories (bottom-right corner of Figure.1). Objects with higher centerness scores are likelier to be objects requiring detection, attributable to the RPN's class-agnostic predictions. Thus, we utilize the centerness score to update the confidence of predictions. A noteworthy point to consider is that updating confidence during training does not effectively address the issue of bias due to the model's reliance on class labels for optimization, which compromises its ability to handle unknown classes.

With the above methods, we achieve nearly **twice** the performance over the previous state-of-the-art (SOTA) (Figure 2). Our model constructs with the pure convolutional neural network without employing dense attention mechanisms. Consequently, it demonstrates a significant advantage in terms of inference speed. We summarize our contributions as follows:

- We strengthen the model's supervision for potential objects by transferring knowledge from larger model.
- We propose Dual Matching Label Assignment (DMLA) to address the competition between known and unknown categories.
- In order to address the bias issue within the model, we propose a method called Class-Awareness Neutralizer (CAN), which is designed to mitigate bias without the need for training participation.
- Evaluation on the OWOD task using Pascal VOC and MS COCO datasets shows that our method achieves **60.2** mAP and **41.5** Unknown Recall at **36.36** FPS, establishing a new SOTA.

2 Related Works

2.1 Open-Set Object Detection

The aim of Open-Set Object Detection (OSOD) is to identify unlabeled objects within the training dataset, which was initially introduced by [Dhamija *et al.*, 2020]. [Miller *et al.*, 2019] and [Miller *et al.*, 2018] utilize **D**ropout **S**ample (DS) to assess the uncertainty of the detector. [Han *et al.*, 2022] and [Miller *et al.*, 2021] propose the secondary decision boundary for potential targets. Additional research[Kim *et al.*, 2022; Wang *et al.*, 2022; Qi *et al.*, 2022] focuses on generating high-quality unknown candidates. The OWOD task is more challenging compared to the OSOD task, as it not only requires the model's ability to detect unlabeled objects within the training set but also evaluates the model's capacity for fine-tuning newly annotated instances.

2.2 Open-World Object Detection

The open-world object detection task was first introduced by ORE[Joseph *et al.*, 2021]. Current research primarily focuses on generating pseudo-labels for potential objects. ORE[Joseph *et al.*, 2021] designates predictions with high

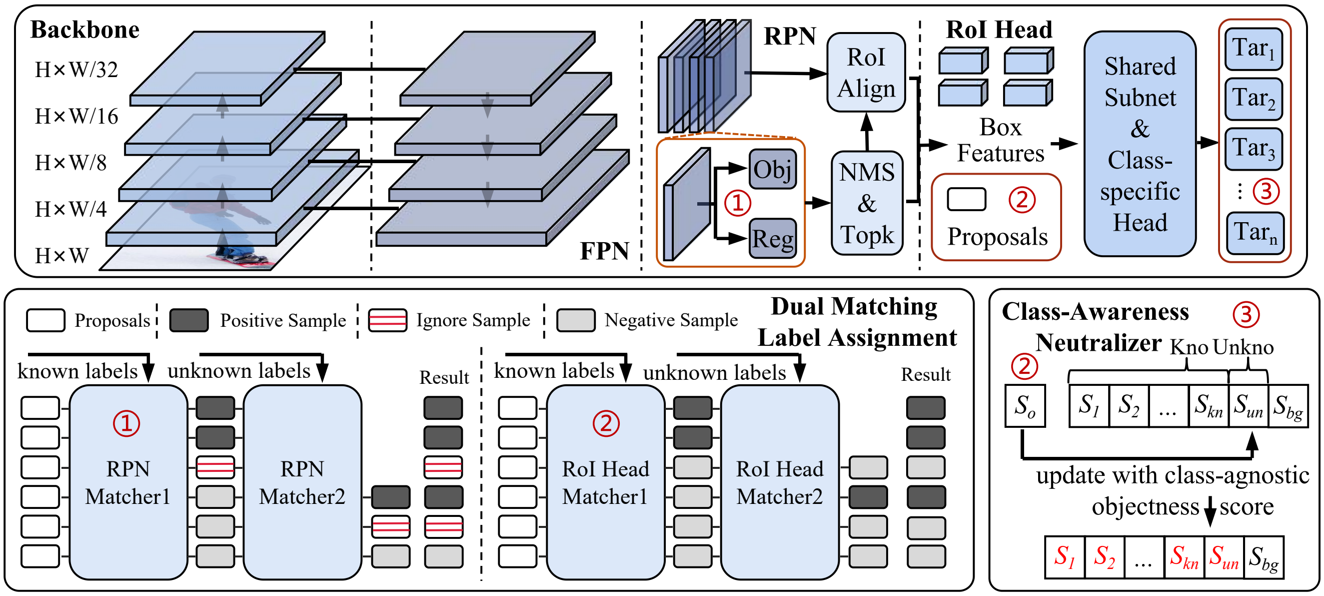

Figure 3: The overall structure of KTCN. Our method is based on the standard Faster R-CNN with FPN. During the training phase, we employ the proposed DMLA in both RPN and RoI Head. In the inference phase, the CAN is utilized to adjust the scores of predictions to neutralize the bias.

centrality scores and no overlap with annotated objects as potential candidate samples. UC-OWOD[Wu *et al.*, 2022b] identifies candidate samples that are not successfully matched in the RPN as potential candidates. OW-DETR[Gupta *et al.*, 2022] calculates objectness scores based on the attention mechanism and selects the Top$_k$ samples as pseudo labels. Both RE-OWOD[Zhao *et al.*, 2023] and CAT[Ma *et al.*, 2023] employ selective search[Uijlings *et al.*, 2013] to generate pseudo labels. Additionally, some methods attempt to estimate robust objectness. 2B-OCD[Wu *et al.*, 2022a] uses a second IOU-based objectness branch to estimate objectness and uses it to confirm detected known and unknown instances during inference. PROB[Zohar *et al.*, 2023] uses probabilistic objectness to adjust the model's confidence output. In this study, we propose the transfer of knowledge from large visual models to guide the model's learning of potential unknown objects.

3 Problem Formulation

Let's consider a model with inputs = $\{\mathbf{X}, \mathbf{Y}\}$. Here, $\mathbf{X}$ is derived from the collection of images $\mathcal{D}_{\mathbf{X}}$, within the dataset $\mathcal{D}$, while $\mathbf{Y}$ denotes the corresponding set of labels. The label set $\mathbf{Y}$ comprises n distinct labels: $\mathbf{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$. Each label $\boldsymbol{y}_i$ includes a category $\boldsymbol{l}_i$ and its corresponding Bounding Box annotation $\boldsymbol{b}_i = [x_i, y_i, w_i, h_i]$, thus $\boldsymbol{y}_i = [\boldsymbol{l}_i, \boldsymbol{b}_i]$. Each category $\boldsymbol{l}_i$ belongs to the category set $\mathcal{K}$, $\boldsymbol{l}_i \in \mathcal{K}$.

In a traditional object detection task, the training set and test set share the same category set $\mathcal{K}$. However, in the setting of OWOD, they are different. OWOD consists of four tasks $T = \{T_1, T_2, T_3, T_4\}$. The input image set $\mathcal{D}_{\mathbf{X}}^{T_i}$ for each task comes from $\mathcal{D}$, $\mathcal{D}_{\mathbf{X}}^{T_i} \in \mathcal{D}$. They have common parts but do not contain each other, i.e., $\mathcal{D}_{\mathbf{X}}^{T_i} \cap \mathcal{D}_{\mathbf{X}}^{T_j} \neq \emptyset$, $\mathcal{D}_{\mathbf{X}}^{T_i} \cup \mathcal{D}_{\mathbf{X}}^{T_j} \neq \mathcal{D}_{\mathbf{X}}^{T_i}$ or $\mathcal{D}_{\mathbf{X}}^{T_j}$, $i \neq j$ and $i, j \in [1, 2, 3, 4]$. The category set $\mathcal{K}_{T_i}$ for each task label satisfies that the previous task is a proper subset of the subsequent task, i.e. $\mathcal{K}_{T_{i-1}} \subset \mathcal{K}_{T_i}$. All tasks share the same test dataset. The category set of the test set includes the category sets of all training sets, $\mathcal{K}_{test} = \mathcal{K} = (\mathcal{K}_{T_1} \cup \mathcal{K}_{T_2} \cup \mathcal{K}_{T_3} \cup \mathcal{K}_{T_4})$. When the model is trained on task T_i, categories satisfying $l \in \mathcal{K}_{test}$ and $l \notin \mathcal{K}_i$ are considered unknown classes.

T_1 evaluates the recall ability of potential objects, i.e., the identifying potential objects. T_2-T_4 simulate incremental learning: when the annotator labels new categories, the ability to fine-tune existing knowledge.

4 Methodology

4.1 Overall Architecture

The overall structure is shown in Figure 3, which is built upon the standard Faster R-CNN with Feature Pyramid Network (FPN[Lin *et al.*, 2017a]). Initially, the input image is processed through the backbone network and FPN to generate multi-scale feature maps. The class-agnostic RPN then generates candidate bounding boxes. The top$_k$ candidate boxes, determined by the objectness scores produced by RPN, are selected and non-maximum suppression (NMS) is applied to eliminate redundant proposals. The remaining candidate boxes are sent to the RoI Head for fine-tuning and class-specific classification.

To provide supervision to potential objects, SAM is utilized to process the input image for generating raw boxes. These boxes are subsequently filtered using both IOU and the shortest side criteria (Section 4.2). To address the competition between ground truths and pseudo labels, the proposed Dual Matching Label Assignment (DMLA) is utilized in both RPN

Figure 4: An illustration of pseudo-label refinement. We employ both IOU and the shortest side to filter the generated candidates. IOU is utilized for filtering annotations of known categories, while the shortest side is used for noise reduction.

and RoI Head for label assignment (Section 4.3). In addition, during the inference stage, we use the Class-Awareness Neutralizer (CAN) proposed in this paper to adjust the prediction confidence. (Section 4.4).

4.2 Transference of Knowledge in Large Model

The primary challenge in OWOD lies in identifying potential targets and subsequently providing effective supervision. Given the significant advantages that SAM demonstrates in zero-shot tasks[Kirillov *et al.*, 2023], we employ it to generate pseudo labels for objects that lack annotations. Initially, SAM is utilized for the segmentation of the input image $\mathbf{X}$ to obtain class-agnostic masks:

$$SAM(\mathbf{X}) = \{i \in [1, n_{sam}] | \left((x_1^i, y_1^i), ..., (x_{n_{ms}}^i, y_{n_{ms}}^i)\right)\}, \tag{1}$$

where n_{sam} and n_{ms} are the number of masks and the number of pixels in the current mask, respectively. Subsequently, the maximum coordinates are calculated for each mask to generate the collection of bounding boxes:

$$\boldsymbol{b}_i = \left(\min_{j \in [1, n_{ms}]} (x_j^i, y_j^i), \max_{j \in [1, n_{ms}]} (x_j^i, y_j^i) \right), \tag{2}$$

$$\mathcal{B}_{SAM} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{n_{sam}}\}. \tag{3}$$

As show in Figure 4, $\mathcal{B}_{SAM}$ may contain ground truths or noises. To avoid this, we utilize the IOU and shortest side of boxes to filter.

For each pseudo-label, we calculate its IOU with all known annotations and determine the maximum value. If the IOU is greater than the threshold $\mathbf{Tr}_{\mathrm{IOU}}$, we consider it as duplicate with gt instances:

$$\mathcal{B}_1 = \{\boldsymbol{b}_j \in \mathcal{B}_{SAM} | \left(\max_{i \in [1, n_{gt}]} \mathbf{IOU}(\boldsymbol{b}_j, \boldsymbol{gt}_i) \right) < \mathbf{Tr}_{\mathrm{IOU}}\}, \tag{4}$$

where $\boldsymbol{gt}_i$ is the instance of known categories. Despite the zero-shot capabilities of SAM, it still generates considerable noise. Upon inspection, we observed that SAM often predicts new masks for certain internal pixels of already segmented objects. these noisy masks are smaller relative to their encompassing objects; thus, employing an appropriate threshold, $\mathbf{Tr}_{\mathrm{SZ}}$, effectively filters out most noise. Thus, we compute the minimum of their lengths and widths. If its value is less than the threshold $\mathbf{Tr}_{\mathrm{sz}}$, we treat it as noise:

$$\mathcal{B}_2 = \{\boldsymbol{b}_j \in \mathcal{B}_1 | \min(\left|x_1^i - x_2^i\right|, \left|y_1^i - y_2^i\right|) \geq \mathbf{Tr}_{\mathrm{SZ}}\}, \tag{5}$$

where $x_1^i, x_2^i, y_1^i, y_2^i$ are bounding box coordinates generated in equation (2).

4.3 Dual Matching Label Assignment (DMLA)

Label assignment is a crucial step in object detection, which is used to determine the relationship between the model's predictions P and labels Y. To provide supervision for potential objects, SAM is utilized to generate pseudo labels. Many samples could be matched to known annotations, but due to those labels, are assigned to pseudo-labels. This competition impairs the model's performance.

To address this, we propose DMLA as shown in the bottom left of Figure 3. To avoid the ambiguous[Ge *et al.*, 2021], we adopt the Max IOU[Lin *et al.*, 2017b] for known categories. We first calculate the IOU between ground truth labels Y_{gt} and all predictions P:

$$\mathbf{M}_{\mathrm{Y}_{gt}}^{\mathrm{P}} = \mathbf{IOU}\left(\boldsymbol{p}_i, \boldsymbol{gt}_j\right) \in \mathbb{R}^{n_{gt} \times n_p}, \boldsymbol{p}_i \in P, \boldsymbol{gt}_j \in Y_{gt}, \tag{6}$$

where n_p is the number of P. Then, we compute the maximum value for each prediction associated with known labels:

$$\tilde{\mathbf{M}}_{\mathrm{Y}_{gt}}^{\mathrm{P}} = \max_{y_j \in \mathrm{Y}_{gt}} \mathbf{IOU}\left(\boldsymbol{p}_i, \boldsymbol{gt}_j\right) \in \mathbb{R}^{n_p}, \boldsymbol{p}_i \in P. \tag{7}$$

During RPN matching, if the IOU is greater than the threshold $\mathbf{Tr}_{\mathrm{h}}$, the prediction is treated as a positive sample. If it is less than the threshold $\mathbf{Tr}_{\mathrm{l}}$, it is treated as a negative sample. Predictions between $\mathbf{Tr}_{\mathrm{l}}$ and $\mathbf{Tr}_{\mathrm{h}}$ are ignored.

$$\mathbf{R}_{\pi_1} = \mathbf{R}_{\pi_1^-} \cup \mathbf{R}_{\pi_1^\sim} \cup \mathbf{R}_{\pi_1^+}, \quad [\underbrace{0, ...,}_{\mathbf{R}_{\pi_1^-}} \underbrace{\mathbf{Tr}_{\mathrm{l}}, ...,}_{\mathbf{R}_{\pi_1^\sim}} \underbrace{\mathbf{Tr}_{\mathrm{h}}, ..., 1}_{\mathbf{R}_{\pi_1^+}}], \tag{8}$$

where $\mathbf{R}_{\pi_1^+}$, $\mathbf{R}_{\pi_1^\sim}$ and $\mathbf{R}_{\pi_1^-}$ denote the positive, ignored and negative sample set generated in first matching $\mathbf{R}_{\pi_1}$, respectively. Then, we match pseudo labels $\mathcal{B}_2$ with $\mathbf{R}_{\pi_1^-}$ to produce the corresponding positive, ignored, and negative sample sets $\mathbf{R}_{\pi_2^+}$, $\mathbf{R}_{\pi_2^-}$ and $\mathbf{R}_{\pi_2^\sim}$. Ultimately, the matching results generated by DMLA is $\mathbf{R}_{rpn} = \mathbf{R}_{\pi_1^+} \cup \mathbf{R}_{\pi_1^\sim} \cup \mathbf{R}_{\pi_2^+} \cup \mathbf{R}_{\pi_2^\sim} \cup \mathbf{R}_{\pi_2^-}$. In the RoI Head, DMLA executes a similar matching process. The difference is that both the first and second matches do not generate an ignored set. Therefore, the set produced by DMLA in the RoI Head is $\mathbf{R}_{roi} = \mathbf{R}_{\pi_1^+} \cup \mathbf{R}_{\pi_2^+} \cup \mathbf{R}_{\pi_2^-}$.

4.4 Class-Awareness Neutralizer (CAN)

The bias problem in the model manifests as the higher classification probability for certain categories. For the OWOD, model does not encounter annotations of the unknown category, thus the bias problem still exists. Although we introduced SAM to generate pseudo labels for potential categories, these labels have lower accuracy compared to the manual annotations of known categories.

To address this, we propose the CAN, which uses the class-agnostic ability of RPN to adjust the confidence (Figure 3 bottom right). Specifically, since the prediction of RPN is in a class-agnostic form, we believe that those with higher objectness in RPN are more likely to object needed detection:

$$S = \{S_i \in \{S_1, ..., S_{kn}, S_{un}\} | \sqrt{S_i \cdot S_o}\} \cup \{S_{bg}\}, \tag{9}$$

Task IDs (→)	Task 1		Task 2				Task3				Task 4		
	U-Recall	mAP (↑)	U-Recall	mAP (↑)			U-Recall	mAP(↑)			mAP (↑)		
	(↑)	Current known	(↑)	Previously known	Current known	Both	(↑)	Previously known	Current known	Both	Previously known	Current known	Both
ORE[Joseph *et al.*, 2021]	8.6	56.3	6.7	52.0	25.2	38.6	6.9	37.9	12.4	29.4	30.1	13.1	25.8
ORE-EBUI[Joseph *et al.*, 2021]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
UC-OWOD[Wu *et al.*, 2022b]	2.4	50.7	3.4	33.1	30.5	31.8	8.7	28.8	16.3	24.6	25.6	15.9	23.2
OCPL[Yu *et al.*, 2022]	8.3	56.6	7.7	50.6	27.5	39.1	11.9	38.7	14.7	30.7	30.7	14.4	26.7
2B-OCD[Wu *et al.*, 2022a]	12.1	56.4	9.4	51.6	25.3	38.5	11.6	37.2	13.2	29.2	30.0	13.3	25.8
OW-DETR[Gupta *et al.*, 2022]	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
CAT†[Ma *et al.*, 2023]	21.8	59.5	19.2	54.6	32.6	43.6	24.4	42.3	18.9	34.5	34.4	16.6	29.9
CAT[Ma *et al.*, 2023]	23.7	60.0	19.1	55.5	32.7	40.1	24.4	42.8	18.7	34.8	34.4	16.6	29.9
PROB[Zohar *et al.*, 2023]	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	**22.2**	36.0	**35.7**	**18.9**	**31.5**
Ours: KTCN	**41.5**	**60.2**	**38.6**	**55.8**	**36.3**	**46.0**	**39.7**	**43.5**	22.1	**36.4**	35.1	16.2	30.4

Table 1: **State-of-the-art comparison for Open-World Object Detection**. U-Recall signifies the recall rate for unknown objects, while mAP represents Mean Average Precision, a widely accepted evaluation metric in object detection. Previously and Current known denote the categories seen in previous tasks and the newly added in the current task, respectively. Both refers to all categories seen so far in the current task. The previous best result is underlined.

where $S_1, ..., S_{kn}$, S_{un} and S_{bg} denote known, unknown and background categories, respectively. $S_o \in [0, 1]$ represents corresponding centerness. The product will reduce confidence, but at the same time, it will improve the absolute ranking of predictions with high centerness, making it more likely to be retained in non-maximum suppression. The principal square root can increase the scores of these predictions, preventing them from being filtered out due to lower confidence.

5 Experiments

5.1 Datasets

Consistent with previous works[Ma *et al.*, 2023; Zohar *et al.*, 2023], we use the VOC[Everingham *et al.*, 2010] and COCO[Lin *et al.*, 2014] to construct a hybrid dataset. Then, the dataset is divided into four tasks, with the respective numbers of training images and instances for each task being: T_1(16551, 47223), T_2(45520, 113741), T_3(39402, 114452), T_4(40260, 138996). T_1 includes the categories from VOC[Everingham *et al.*, 2010], and each subsequent task adds 20 new categories based on the previous tasks for the incremental experiment. The test set includes all categories with 10246 images and 61707 test instances. During evaluation, all previously seen categories are considered known categories (including newly added ones), while the remaining ones are considered unknown categories.

5.2 Metric

We adhere to the standard evaluation metrics for the OWOD task[Joseph *et al.*, 2021; Gupta *et al.*, 2022]. The performance of the model on known categories is assessed by the widely accepted object detection metric, Mean Average Precision (mAP). For unknown categories, model's performance on objects without annotations is evaluated using the recall rate (U-Recall). In incremental learning tasks (T2-T4), the performance on known categories is divided into two aspects: performance on categories previously encountered, and performance on categories introduced in the current task. The former is used to assess the impact of the introduction of new categories on the model's existing knowledge during incremental learning tasks, while the latter evaluates the model's adaptability and performance on newly introduced categories.

5.3 Details

Our experiments are based on Detectron2[Wu *et al.*, 2019] with 2 × NVIDIA GeForce RTX 4090 (total 48G). We set the batch size to 8 and the initial learning rate to 0.005. The maximum number of iterations was set at 120k. When the number of iterations reached 60k and 100k, we reduced the model's current learning rate to one-tenth of its original value. In multi-scale training, the shortest side of the image was scaled between 220 and 1088, with the longest side not exceeding 1333. For incremental learning, we set the learning rate to 0.001 ($T_2 - T_3$) and 0.0005 (T_4), the number of warm-up iterations to 500, and the maximum number of iterations to 80k, without reducing the learning rate. Apart from T_1, all tasks employed replay to review knowledge[Joseph *et al.*, 2021]. For SAM, we chose SAM_H to generate pseudo-labels.

5.4 Comparison With State-of-the-art Models

We conducted comparative experiments with other SOTA methods, and the results are shown in Table 1. ORE established an energy model that updates detected known categories to unknown categories, which may lead to data leakage as it uses fully annotated data. Following [Zohar *et al.*, 2023; Ma *et al.*, 2023], we removed the energy model in ORE and evaluated its performance (ORE-EBUI). Although CAT achieved higher results by introducing the cascaded structure, it brought a large amount of computational consumption. Therefore, we added a version without the cascaded structure (CAT†).

Identifying Potential Objects (T_1). In T_1, our method achieved comprehensive leadership, whether in unknown recall rate or mAP on categories already seen. In terms of performance on known categories, we exceed PROB by 0.7 mAP and surpass CAT† by 0.7 mAP. For performance on unknown objects, we achieved the highest unknown recall rate of 41.5.

Set_{RPN}	mAP_{CK}	U-Recall
Full	54.8	3.5
$\mathbf{P}_\sim$	54.2	3.6
$\mathbf{P}_-$	**56.3**	**5.1**
$\mathbf{P}_\sim \cup \mathbf{P}_-$	54.8	3.4

(a) **DMLA for RPN**. Set_{RPN} denotes the prediction set employed during the second assignment. Full: one-time matching of all labels. $\mathbf{P}_\sim$: ignored set. $\mathbf{P}_-$: negative set.

Set_{ROI}	mAP_{CK}	U-Recall
Full	54.2	23.6
π_{oto}	**56.1**	27.8
π_{otm}	55.4	32.3
SAM_H	55.3	**34.7**

(b) **DMLA for RoI Head and Larger SAM Model**. Full: one-time matching of all labels. π_{otm}: the one-to-many label assignment. π_{oto}: the one-to-one assignment.

$\mathbf{Tr}_{\text{IOU}}$	mAP_{CK}	U-Recall
0.1	55.4	30.7
0.3	55.5	33.6
0.7	55.5	**35.1**
0.9	**55.6**	**35.1**

(c) **IOU Threshold for Initial Filtering**. The threshold, $\mathbf{Tr}_{\text{IOU}}$, is used to control the filtering of pseudo-labels by comparing them with ground truth labels.

RPN	RoI	mAP_{CK}
	✔	57.6
✔		**59.7**
✔	✔	58.3
		59.1

(d) **Match Sampling**. ✔ indicates that we randomly select ordered pairs from positive/negative set and convert them into ignored set.

$\mathbf{Tr}_{\text{SZ}}$	mAP_{CK}	U-Recall
0	58.8	40.2
5	59.9	**41.2**
25	**60.3**	36.5
Inner	60.1	39.3

(e) **Second Filtering of Pseudo Labels**. Number refers to the length of shortest side filtered during the second filtering process, while Inner denotes the use of calculated overlap between masks for filtering purposes.

CAN	mAP_{CK}	U-Recall
Wo	59.9	41.2
Full	**60.2**	41.5
Unknown	59.4	**41.8**
Training	59.8	41.0

(f) **Class-Awareness Neutralizer**. The table compares CAN usage: non-utilization (Wo), application to both known and unknown (Full), application to unknown (Unknown) and application to train (Training).

T_2	mAP_{Pre}	mAP_{CK}	mAP_{Bo}	U-Recall
5_0	55.7	34.9	45.3	**39.2**
5_5	55.8	36.0	46.0	38.6
5_25	**55.9**	**36.3**	**46.1**	34.6
Mv_Prev	54.5	35.0	44.8	34.2

(g) **Incremental Learning**. In the notation 5_0, the first numeral signifies the value of $\mathbf{Tr}_{\text{SZ}}$ during T_1, while the subsequent numeral reflects its value in T_2. Mv_Prev indicates the addition of categories previously encountered to the known categories set in the initial IOU filtering, thereby filtering out potential annotations of known categories.

Table 2: **Ablation Study**. We conducted a comparison of all configurations within the experiment and explored additional potential methodologies. In the table, mAP_{Pre}, mAP_{CK} and mAP_{Bo} represent the mAP for previously encountered categories, newly introduced categories in the current task, and for all categories encountered, respectively.

The closest result to this is CAT, but we still lead significantly with a recall rate advantage of 75.1% (41.5 vs 23.7). For CAT†, the advantage reaches 90.4% (41.5 vs 21.8). For PROB, we achieved more than double performance 213.9%.

Incremental learning ($T_2 - T_4$). In U-Recall, we maintain the advantage (T_2: 101.1% CAT), (T_3: 62.7% CAT). With the increase in known categories, our model gradually falls behind PROB in known categories. Such a result is acceptable since KTCN recalls more unknown targets, which reduces the proportion of seen categories in the top 100 predictions.

5.5 Ablation Study

In this section, we demonstrate the step-by-step construction of our KTCN. We start with Faster R-CNN.

DMLA for RPN (table 2a). Initially, we investigate which category of sets is most conducive for matching pseudo labels in RPN. The result indicates that performing the secondary match within the negative sample set yields the most favorable results (56.3 mAP and 5.1 U-Recall).

DMLA for RoI Head and Larger SAM Model (table 2b). At this stage, we introduce a contrast with one-to-one label assignment π_{oto} as an alternative solution to boundary ambiguity[Ge *et al.*, 2021]. However, compared to its one-to-many label assignment counterpart π_{otm}, it still lags by 4.5 in the performance of the unseen category (27.8 vs 32.3). Therefore, we employ π_{otm} and match in the negative sample for unknown categories. Building upon this, we introduce the SAM_H model with higher precision, which yields a 2.4 improvement in U-Recall, achieving 34.7.

IOU Threshold for Initial Filtering (table 2c). $\mathbf{Tr}_{\text{IOU}}$ is utilized to filter out samples in pseudo-labels that overlap with ground truth labels. Setting $\mathbf{Tr}_{\text{IOU}}$ to 0.9 allows the model to reach its peak performance in both known and unknown positions, achieving 55.6 mAP and 35.1 U-Recall.

Match Sampling (table 2d). We incorporate FPN and train the model exclusively with known category labels. We posit that randomly converting positive or negative samples into ignored samples[Lin *et al.*, 2017b] could detrimentally impact the model's performance. The findings reveal that sampling set during the RPN and refraining from sampling during the RoI Head leads to the best outcome (59.7 mAP).

Second Filtering of Pseudo-Labels (table 2e). The minimum side is used to filter small noises. In addition, we provide an alternative solution, mask filtering (Inner), which calculates quantifies the overlap by computing the ratio of overlapping pixels to the total pixels in the smaller mask. If the overlap rate is more than 0.9, we treat it as noise. With $\mathbf{Tr}_{\text{SZ}}$ = 5, the model achieves the highest U-Recall of 41.2. Although the mask filtering lags behind by 2.1 points in U-Recall, it demonstrates a marginal enhancement in mAP by 0.2. We maintain the use of minimum side filtering to attain an elevated recall rate.

Class-Awareness Neutralizer (table 2f). In this section, we investigate the impact of CAN on predictions for known and unknown categories separately. The results indicate that employing CAN within all categories yields the highest performance in both mAP and U-Recall. When CAN is utilized in training, the performance declines by 0.4 mAP and 0.5 U-Recall since it hurts class-agnostic ability of RPN.

Incremental Learning (table 2g). A larger threshold significantly reduces the number of pseudo-labels, which is beneficial for fine-tuning the model on known categories (5_25:

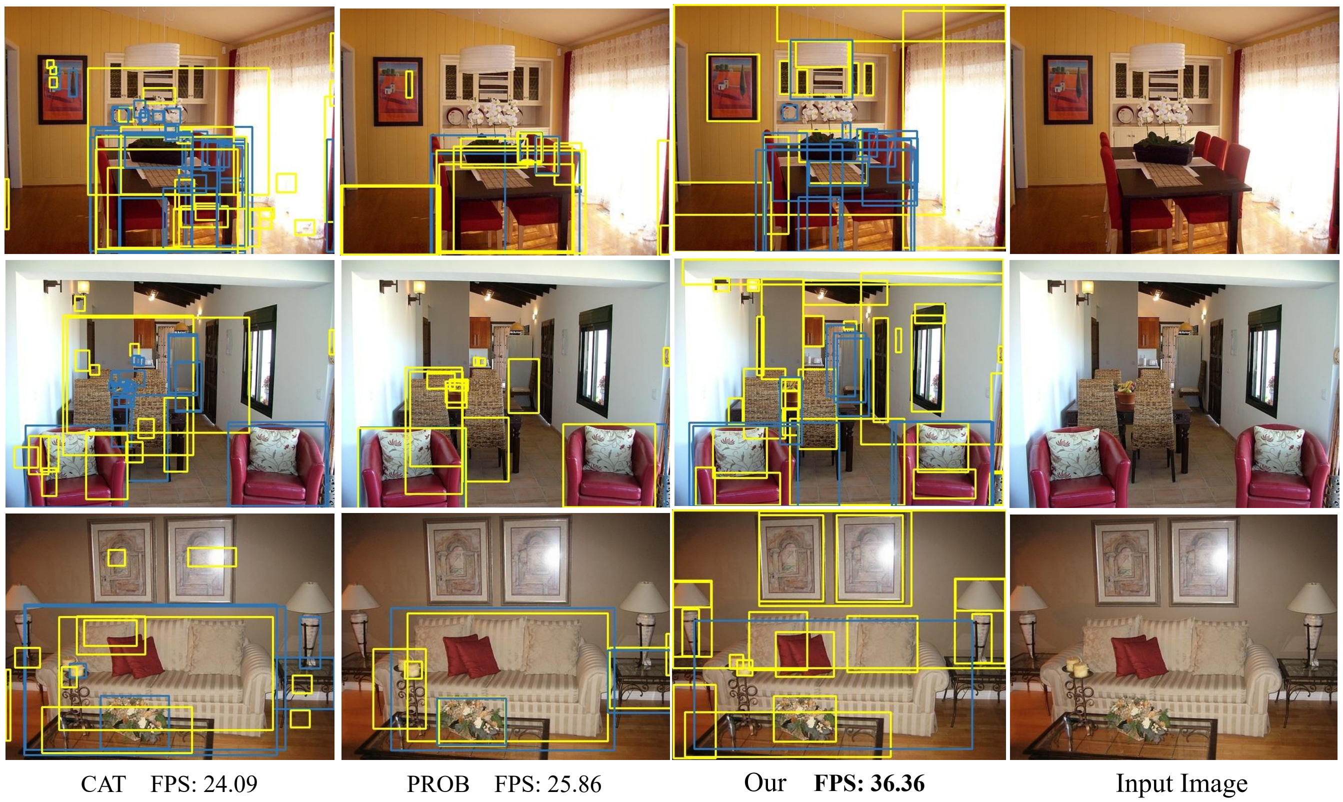

Figure 5: **Qualitative Analysis and Inference Speed Comparison**. We benchmarked against the latest networks, CAT and PROB. The Frames Per Second (FPS) were evaluated on a single 3090. The box of blue denotes known categories detected, while yellow represents the prediction of potential objects.

46.1 vs 5_0: 45.3). Conversely, it demonstrates a disadvantage in U-Recall (5_25: 34.6 vs 5_0: 39.2). Thus, we chose $\mathbf{Tr}_{\mathrm{SZ}} = 5$ as the threshold in incremental learning for balance in known and unknown. Building upon this, we explored whether applying IOU filtering to all previously encountered categories would yield better performance, given that pseudo-labels may include categories previously seen. However, as indicated in table 2g, it underperformed others in both known and unknown categories. Therefore, during $T_2 - T_4$, we report our results using a consistent $\mathbf{Tr}_{\mathrm{SZ}} = 5$ and applying IOU filtering solely to the newly introduced categories.

5.6 Visualization

In Figure 5, we present the qualitative analysis of our model and a comparison of inference speeds with the latest works. Our approach demonstrates superior unknown object recall capabilities, as evidenced by its ability to identify paintings, light fixtures (see the first row), throw pillows, windows (see the second row), table lamps, and red pillows (see the third row), while other methods classify them to the background. In inference speed, CAT's cascading structure results in the lowest speed. Compared to the original Faster R-CNN, our method does not introduce any additional parameters, thus offering a significant inference advantage. Specific, Our method leads CAT 12.27 (36.36 vs 24.09) and PROB 10.5 (36.36 vs 25.86), showcasing not only the efficiency of KTCN but also its potential for applications.

6 Conclusions

In this study, we introduced the large visual model, the Segment Anything Model (SAM), to generate pseudo labels for potential objects. To leverage the knowledge of SAM, we proposed the Dual Matching Label Assignment (DMLA) to address the competition in one-to-many label assignment. Furthermore, we developed the Class-Awareness Neutralizer (CAN) to eliminate the model's bias toward seen categories. With these methods, we achieved new state-of-the-art (SOTA) in recall rates of the unknown category and inference speed. We expect the methods proposed in this paper to advance the application of open-world object detection tasks in the real-world environment.

Acknowledgements

This work was also partially supported by Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou)

References

[Dhamija *et al.*, 2020] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of* *the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1021–1030, 2020.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[Fomenko *et al.*, 2022] Vladimir Fomenko, Ismail Elezi, Deva Ramanan, Laura Leal-Taixé, and Aljosa Osep. Learning to discover and detect objects. *Advances in Neural Information Processing Systems*, 35:8746–8759, 2022.

[Ge *et al.*, 2021] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021.

[Gu *et al.*, 2021] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[Guo *et al.*, 2022] Zexuan Guo, Chensheng Wang, Guang Yang, Zeyuan Huang, and Guo Li. Msft-yolo: Improved yolov5 based on transformer for detecting defects of steel surface. *Sensors*, 22(9):3467, 2022.

[Gupta *et al.*, 2022] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9235–9244, 2022.

[Han *et al.*, 2022] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9591–9600, 2022.

[Huang *et al.*, 2021] Li Huang, Qiaobo Fu, Meiling He, Du Jiang, and Zhiqiang Hao. Detection algorithm of safety helmet wearing based on deep learning. *Concurrency and Computation: Practice and Experience*, 33(13):e6234, 2021.

[Joseph *et al.*, 2021] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021.

[Kim and Lee, 2020] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 355–371. Springer, 2020.

[Kim *et al.*, 2022] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2):5453–5460, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2019] Yang Liu, Zhuo Ma, Ximeng Liu, Siqi Ma, and Kui Ren. Privacy-preserving object detection for medical images with faster r-cnn. *IEEE Transactions on Information Forensics and Security*, 17:69–84, 2019.

[Loey *et al.*, 2021] Mohamed Loey, Gunasekaran Manogaran, Mohamed Hamed N Taha, and Nour Eldeen M Khalifa. Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection. *Sustainable cities and society*, 65:102600, 2021.

[Ma *et al.*, 2023] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19681–19690, 2023.

[Miller *et al.*, 2018] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018.

[Miller *et al.*, 2019] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 international conference on robotics and automation (icra)*, pages 2348–2354. IEEE, 2019.

[Miller *et al.*, 2021] Dimity Miller, Niko Sünderhauf, Michael Milford, and Feras Dayoub. Uncertainty for identifying open-set errors in visual object detection. *IEEE Robotics and Automation Letters*, 7(1):215–222, 2021.

[Ozturk *et al.*, 2020] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121:103792, 2020.

[Qi *et al.*, 2022] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Solovyev *et al.*, 2021] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

[Tian *et al.*, 2019] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[Uijlings *et al.*, 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104:154–171, 2013.

[Wang *et al.*, 2022] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4422–4432, 2022.

[Wang *et al.*, 2023a] Hai Wang, Yansong Xu, Zining Wang, Yingfeng Cai, Long Chen, and Yicheng Li. Centernet-auto: A multi-object visual detection algorithm for autonomous driving scenes based on improved centernet. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[Wang *et al.*, 2023b] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023.

[Wu *et al.*, 2019] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[Wu *et al.*, 2022a] Yan Wu, Xiaowei Zhao, Yuqing Ma, Duorui Wang, and Xianglong Liu. Two-branch objectness-centric open world detection. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, pages 35–40, 2022.

[Wu *et al.*, 2022b] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *European Conference on Computer Vision*, pages 193–210. Springer, 2022.

[Yu *et al.*, 2022] Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class prototype learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 626–630. IEEE, 2022.

[Zhang *et al.*, 2020] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020.

[Zhao *et al.*, 2023] Xiaowei Zhao, Yuqing Ma, Duorui Wang, Yifan Shen, Yixuan Qiao, and Xianglong Liu. Revisiting open world object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Zheng *et al.*, 2022] Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2022.

[Zhong *et al.*, 2022] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[Zhou *et al.*, 2021] Xiaokang Zhou, Xuesong Xu, Wei Liang, Zhi Zeng, Shohei Shimizu, Laurence T Yang, and Qun Jin. Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 18(2):1377–1386, 2021.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[Zohar *et al.*, 2023] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2023.

[Zou *et al.*, 2019] Qin Zou, Hanwen Jiang, Qiyu Dai, Yuanhao Yue, Long Chen, and Qian Wang. Robust lane detection from continuous driving scenes using deep neural networks. *IEEE transactions on vehicular technology*, 69(1):41–54, 2019.