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 Dual Matching Label Assignment (DMLA) strategy. Furthermore, we propose the Class-Awareness Neutralizer (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 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
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 Dual Matching Label Assignment (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 Class-Awareness Neutralizer (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 Dropout Sample (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
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 Topk 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 \( \{X, Y\} \). Here, \( X \) is derived from the collection of images \( D_X \) within the dataset \( D \), while \( Y \) denotes the corresponding set of labels. The label set \( Y \) comprises \( n \) distinct labels: \( Y = \{y_1, y_2, \ldots, y_n\} \). Each label \( y_i \) includes a category \( l_i \) and its corresponding Bounding Box annotation \( b_i = [x_i, y_i, w_i, h_i] \), thus \( y_i = [l_i, b_i] \). Each category \( l_i \) belongs to the category set \( \mathcal{K} \), \( l_i \in \mathcal{K} \).

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 \( D_X^{T_i} \) for each task comes from \( D \), \( D_X^{T_i} \in D \). They have common parts but do not contain each other, i.e., \( D_X^{T_i} \cap D_X^{T_j} \neq \emptyset, D_X^{T_i} \cup D_X^{T_j} \neq D_X \) or \( D_X^{T_i}, i \neq j \) 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., \( \mathcal{K}_{T_i} \subseteq \mathcal{K}_{T_j} \). 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.

\( T_4 \) 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 topk 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...
and Rol 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 \( X \) to obtain class-agnostic masks:

\[
SAM(X) = \{i \in [1, n_{sam}] \mid (x_i^1, y_i^1), ..., (x_i^{n_{ms}}, y_i^{n_{ms}})\},
\]

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:

\[
b_i = \left( \min_{j \in [1, n_{ms}]} (x_j^1, y_j^1), \max_{j \in [1, n_{ms}]} (x_j^1, y_j^1) \right),
\]

\[
B_{SAM} = \{b_1, b_2, ..., b_{n_{sam}}\}.
\]

As shown in Figure 4, \( B_{SAM} \) 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 \( T_{IOU} \), we consider it as duplicate with gt instances:

\[
B_1 = \{b_i \in B_{SAM} \mid \max_{i \in [1, n_{sam}]} \text{IOU}(b_i, gt_i) < T_{IOU}\},
\]

where \( gt_i \) 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 \( T_{SZ} \), effectively filters out most noise. Thus, we compute the minimum of their lengths and widths. If its value is less than the threshold \( T_{SZ} \), we treat it as noise:

\[
B_2 = \{b_i \in B_1 \mid \min(\{x_1^1 - x_2^1, |y_1^1 - y_2^1|\}) \geq T_{SZ}\},
\]

where \( x_1^1, x_2^1, y_1^1, y_2^1 \) 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 \( P \):

\[
M_{Y_{gt}}^P = \text{IOU}(p_i, gt_j) \in \mathbb{R}^{n_P \times n_{gt}}, p_i \in P, gt_j \in Y_{gt},
\]

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

\[
\hat{M}_{Y_{gt}}^P = \max_{y_j \in Y_{gt}} \text{IOU}(p_i, gt_j) \in \mathbb{R}^{n_P}, p_i \in P.
\]

During RPN matching, if the IOU is greater than the threshold \( T_{Rpn} \), the prediction is treated as a positive sample. If it is less than the threshold \( T_{R} \), it is treated as a negative sample.

Predictions between \( T_{R} \) and \( T_{Rpn} \) are ignored:

\[
\hat{R}_{p_i} = R_{p_i}^+ \cup R_{p_i}^- \cup R_{p_i}^0, \quad \hat{R}_{p_i} = R_{p_i}^+ \cup R_{p_i}^- \cup R_{p_i}^0.
\]

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 \( \hat{R}_{roi} = R_{roi}^+ \cup R_{roi}^- \cup R_{roi}^0 \).

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 be object needed detection:

\[
S = \{S_i \in \{S_1, ..., S_{kn}, S_{an}\} \mid \sqrt{S_1 \cdot S_o} \cup \{S_{bg}\},
\]
where $S_1, ..., S_{kn}, S_{ub}$ and $S_{bg}$ denote known, unknown and background categories, respectively. $S_r \in [0, 1]$ 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 $2 \times$ 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$^1$).

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$^1$ by 0.7 mAP. For performance on unknown objects, we achieved the highest unknown recall rate of 41.5.

---

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.

<table>
<thead>
<tr>
<th>Task IDs (→)</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U-Recall</td>
<td>mAP(↑)</td>
<td>U-Recall</td>
<td>mAP(↑)</td>
</tr>
<tr>
<td>ORE [Joseph et al., 2021]</td>
<td>8.6</td>
<td>56.3</td>
<td>6.7</td>
<td>52.0</td>
</tr>
<tr>
<td>ORE-EBUI [Joseph et al., 2021]</td>
<td>4.9</td>
<td>56.0</td>
<td>2.9</td>
<td>52.7</td>
</tr>
<tr>
<td>UC-OWOD [Wu et al., 2022b]</td>
<td>0.4</td>
<td>50.7</td>
<td>3.4</td>
<td>33.1</td>
</tr>
<tr>
<td>OCPL [Yu et al., 2022]</td>
<td>8.3</td>
<td>56.6</td>
<td>7.7</td>
<td>50.6</td>
</tr>
<tr>
<td>2B-OCD [Wu et al., 2022a]</td>
<td>12.1</td>
<td>56.4</td>
<td>9.4</td>
<td>51.6</td>
</tr>
<tr>
<td>OW-DETR [Gupta et al., 2022]</td>
<td>7.5</td>
<td>59.2</td>
<td>6.2</td>
<td>53.6</td>
</tr>
<tr>
<td>CAT [Ma et al., 2023]</td>
<td>21.8</td>
<td>59.5</td>
<td>19.2</td>
<td>54.6</td>
</tr>
<tr>
<td>CAT [Ma et al., 2023]</td>
<td>23.7</td>
<td>60.0</td>
<td>19.1</td>
<td>55.5</td>
</tr>
<tr>
<td>PROB [Zohar et al., 2023]</td>
<td>19.4</td>
<td>59.5</td>
<td>17.4</td>
<td>55.2</td>
</tr>
<tr>
<td>Ours: KTCN</td>
<td>41.5</td>
<td>60.2</td>
<td>38.6</td>
<td>55.8</td>
</tr>
</tbody>
</table>

1466
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\(^\dagger\), the advantage reaches 90.4% (41.5 vs 21.8). For PROB, we achieved more than double performance 213.9%.

**Incremental learning** \((T_2 - T_1)\). In U-Recall, we maintain the advantage \((T_2: 101.1\% \text{ CAT}), (T_3: 62.7\% \text{ CAT})\). 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** \((\text{table 2a})\). 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 \((56.3 \text{ mAP and } 5.1 \text{ U-Recall})\).

**DMLA for RoI Head and Larger SAM Model** \((\text{table 2b})\). At this stage, we introduce a contrast with one-to-one label assignment \(\pi_{oto}\) as an alternative solution to boundary ambiguity\([Ge \textit{et al.}, 2021]\). However, compared to its one-to-many label assignment counterpart \(\pi_{otm}\), it still lags by 4.5 in the performance of the unseen category \((27.8 \text{ vs } 32.3)\). Therefore, we employ \(\pi_{otm}\) 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** \((\text{table 2c})\). \(\text{Tr}_{IOU}\) is utilized to filter out samples in pseudo-labels that overlap with ground truth labels. Setting \(\text{Tr}_{IOU}\) 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** \((\text{table 2d})\). We incorporate FPN and train the model exclusively with known category labels. We posit that randomly converting positive or negative samples into ignored samples\([Lin \textit{et al.}, 2017]\) 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 \((59.7 \text{ mAP})\).

**Second Filtering of Pseudo-Labels** \((\text{table 2e})\). The minimum size is used to filter small noises. In addition, we provide an alternative solution, mask filtering \(\text{(Inner)}\), which calculates 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 \(\text{Tr}_{SZ} = 5\) 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** \((\text{table 2f})\). 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** \((\text{table 2g})\). A larger threshold significantly reduces the number of pseudo-labels, which is beneficial for fine-tuning the model on known categories \((5.25\%): 59.9\%\).

---

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


