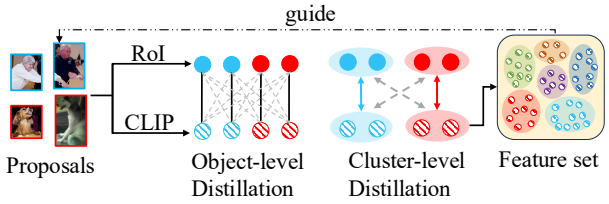
CAKE: Category Aware Knowledge Extraction for Open-Vocabulary Object Detection

Shiyuan Ma[1]*, Donglin Qian[1]*, Kai Ye[1], Shengchuan Zhang[1]†

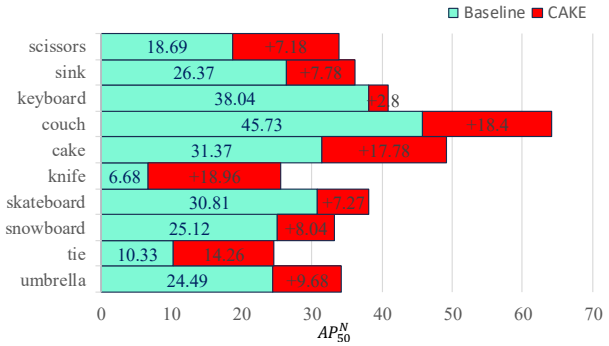
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China
Xiamen University, 361005, P.R. China.
{mashiyuan, qiandl, yekai}@stu.xmu.edu.cn, zsc_2016@xmu.edu.cn

Abstract

Open vocabulary object detection (OVOD) task aims to detect objects in novel categories beyond the base categories in the training set. To this end, the detector needs to access image-text pairs containing rich semantic information or the visual language pre-trained model (VLM) learned on them. Recent OVOD methods rely on knowledge distillation from VLMs. However, there are two main problems in current methods: (1) Current knowledge distillation frameworks fail to take advantage of the global category information of VLMs and thus fail to learn category-specific knowledge. (2) Due to the overfitting phenomenon of base categories during training, current OVOD networks generally have the problem of suppressing novel categories as background. To address these two problems, we propose a Category Aware Knowledge Extraction framework (CAKE), which consists of a Category-Specific Knowledge Distillation branch (CSKD) and a Category Generalization Region Proposal Network (CG-RPN). CSKD can more fully extract category-strong related information through category-specific distillation, and it is also conducive to filtering the exclusion problem between individuals of the same category; in this process, the model constructs a category-specific feature set to maintain high-quality category features. CG-RPN leverages the guidance of the feature set to adjust the confidence scores of region proposals, thereby mining proposals that potentially contain novel categories of objects. Extensive experiments show that our method can plug and play well with many existing methods and significantly improve their detection performance. Moreover, our CAKE framework can reach state-of-the-art performance on OV-COCO and OV-LVIS datasets.



(a) An overview of our proposed CAKE.



(b) Enhancement on challenging novel categories.

Figure 1: (a) We introduce extra cluster-level distillation to learn category-specific knowledge. Its output is also used to build a feature set to guide proposal generation, which helps alleviate overfitting. (b)We present the results of CAKE on novel categories where the AP on the baseline is below 50, showing that CAKE can effectively improve performance on novel categories.

Introduction

Traditional object detection methods follow a close set setting, which all focuses on identifying objects from categories present in the training set, corresponding detection networks can perform excellently in recognizing these categories(Ren et al. 2015; Mi et al. 2022; Xie et al. 2021; Chen et al. 2024, 2022). However, real-world images often contain a vast array of object categories, thus constructing a dataset that covers all possibilities is impractical. Additionally, due to the long-tail distribution of data, detector performance can vary significantly across different categories.

Open Vocabulary Object Detection (OVOD) addresses this challenge by enabling the detection of both known (base) and unknown (novel) categories using a network trained with fine-grained data only on base categories. This approach reduces the need for extensive training data and improves adaptability across various detection tasks. To achieve OVOD, models often leverage Vision-Language Models (VLMs) like CLIP (Radford et al. 2021), which align visual and textual feature spaces by learning relationships from large image-text datasets. In this paper, we uti-

*These authors contributed equally.
†Corresponding Author

lize the Faster R-CNN (Ren et al. 2015) architecture for its balance of model size, computational efficiency, ease of deployment, and accuracy across multiple categories.

Building on this foundation, previous OVOD approaches have focused on extracting knowledge of base and novel categories from VLM. For example, the DK-DETR (Li et al. 2023), OADP (Wang et al. 2023b), and OC-OVD (Bangalath et al. 2022) employ both semantic and relational knowledge distillation to explore the relationships between different categories. BARON (Wu et al. 2023a) discovers potential novel objects by exploring relationships between different objects in space. However, it does not learn feature knowledge that is strongly related to specific categories. The LBP (Li et al. 2024a) classifies proposals according to foreground, background, novel category, and base category, establishing a corresponding alignment method for each classification. The DST-Det (Xu et al. 2023) utilizes all class embeddings extracted by CLIP text encoder and measure the similarity between RPN proposals and class embeddings to find out novel proposals. GOAT (Wang et al. 2023a) does something like DST-Det by introducing an open-corpus, which is consisted of class names and their synonyms.

In order to solve the problem of intra-category differences and learn global knowledge that is strongly related to the categories, we introduce the Category-Specific Knowledge Distillation (CSKD) module, which contains an object-level object distillation branch to learn the individual semantic knowledge and relational knowledge, and a cluster-level object distillation branch to cluster features of the same category and enclose corresponding clusters of student and teacher model. Furthermore, we construct a categories-explicit visual semantic space to guide detectors to recognize different categories more clearly than simply use text features, and no need for more class information. To alleviate the overfitting phenomenon of base categories, we propose a Category Generalization Region Proposal Network (CG-RPN), which utilizes the feature space obtained from CSKD to optimize the generation of objectness scores in the RPN process, thereby helping the detector discover a wider range of objects. We present the overview of our framework and the performance enhancement on challenging novel categories compared to the baseline in Fig. 1. Our main contributions are summarized as follows:

- We propose a Category-Specific Knowledge Distillation (CSKD) module for extracting and generating image features that are strongly related to specific categories, which effectively enhances the efficiency of knowledge distillation.
- We introduce the Category-Generalized Regional Proposal Network (CG-RPN) to discover features related to novel categories, which improve the perception ability of novel objects and alleviate the overfit on base category.
- We develop the Category-Aware Knowledge Extraction (CAKE) framework, a plug-and-play method that can be integrated with various OVOD algorithms to improve their performance. CAKE achieves state-of-the-art results on both OV-LVIS and OV-COCO datasets.

Related Work

Open Vocabulary Object Detection

The OVOD (Cai et al. 2022; Kamath et al. 2021; Minderer et al. 2022; Yao et al. 2022; Lin et al. 2024; Qu et al. 2024; Gong et al. 2024) aims to train a model that can detect objects of any categories, even if some categories are unseen to the detector during training. OVR-CNN (Zareian et al. 2021) is the seminal work that proposes this task and achieves great performance by combining box annotations and image-text pairs. Subsequently, prevailing approaches to tackle OVOD mainly rely on VLMs as an effective alternative to directly accessing image-text pairs. These methods can be roughly divided into two types: (1) Fine-tuning VLMs (Kim, Angelova, and Kuo 2023; Kuo et al. 2022; Li et al. 2022; Lin et al. 2022),(2) distillation from VLMs (Bangalath et al. 2022; Ma et al. 2022; Wu et al. 2023b; Gu et al. 2021; Chen et al. 2023). Methods of the first type add large-scale learnable model weights to retrain VLMs and use them for feature extraction or object detection, which is resource-intensive. Moreover, VLM is primarily trained on image-level data and has difficulty in accurately performing instance-level tasks. The methods based on knowledge distillation leverage knowledge distilled from CLIP (Radford et al. 2021) to obtain perception on novel categories. Efforts in this aspect are mainly reflected in the improvement of distillation methods or promoting efficient learning of data of different granularity. However, these methods do not pay attention to the impact of intra-category differences on distillation quality and are all bothered by overfitting to the base category. Our approach addresses both issues by introducing category-specific distillation and category-general proposal generation.

Knowledge Distillation

Knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015) leverages a powerful teacher model to train a student model with fewer parameters, enabling the student to achieve comparable performance. Several works have applied KD to object detection frameworks. For instance, (Chen et al. 2017) implements distillation in Faster R-CNN using feature-based and response-based loss. DeFeat (Guo et al. 2021) simultaneously distills foreground and background regions with different factors. G-DetKD (Yao et al. 2021) provides a general distillation framework for object detectors. FKD (He and Ozay 2022) focuses on distilling attention maps, while FGD (Yang et al. 2022) combines focal and global information for a more comprehensive method.

Compared to these methods, distillation has also been widely applied to Open-Vocabulary Object Detection (OVOD) tasks. For example, BARON (Wu et al. 2023a) utilizes proposals to generate bags of regions that may contain novel categories and performs distillation on both simultaneously. DK-DETR (Li et al. 2023) and OC-OVD (Bangalath et al. 2022) consider the relationships between categories by constructing a feature relationship matrix during distillation. OADP distills both objects and their contextual information (Wang et al. 2023b). In this paper, we propose a CAKE framework to improve the distillation progress fo-

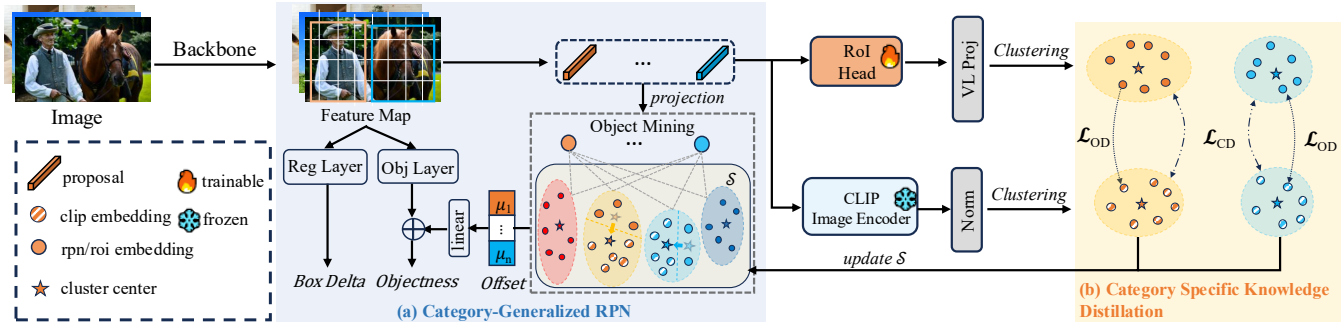


Figure 2: An overall architecture of proposed open-vocabulary detection framework. Object feature embeddings (denoted as circles) are transformed through knowledge distillation into category-specific features (denoted as stars), which are utilized for constructing feature maps. The feature map is applied within a Generalized Region Proposal Network (G-RPN) to mitigate overfitting issues associated with base classes and is further utilized in a feature refinement module for the enhanced delineation of instance-specific features. Shapes of different colors represent distinct categories.

cusing on the category-strong-related information from both proposals and category names.

Methods

Open-vocabulary Object Detection(OVOD) aims to detect objects of categories that do not appear during training. This approach divides categories of objects into two types, which are base category $\mathcal{C}_B$ and novel category $\mathcal{C}_N$, respectively. In addition, to simulate real-world scenarios, an open semantics set $\mathcal{C}_O$ is also included. Noted that $\mathcal{C}_N \neq \varnothing$ and $\mathcal{C}_B \cap \mathcal{C}_N = \varnothing$. In the training dataset $\mathcal{D}_T$, the detector can only access fine-grained data belonging to $\mathcal{C}_B$ and weakly-supervised data belongs to $\mathcal{C}_O$. In the inference dataset $\mathcal{D}_I$, the detector needs to recognize and locate objects belonging to $\mathcal{C}_B \cup \mathcal{C}_N$. By the convention of OVOD, we replace class-specific classifier to text embeddings by CLIP text encoder with predefined template (Wu et al. 2023a).

Preliminaries

The illustration of CAKE is shown in Fig. 2. An image $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$ first passes through backbone to extract feature map $\mathbf{F}$, then the category generalized region proposal network (CG-RPN) will generate proposal boxes $\mathcal{B} \subset \mathbb{R}^4$ on $\mathbf{F}$. $\mathcal{B}$ will pass through the regression layer and the objectness layer to generate the corresponding bounding box deltas and objectness score. On the other hand, $\mathcal{B}$ will be projected into the embedding space and then calculate the correlation with the foreground features from the high-quality feature set $\mathcal{S}$. The calculation result will be adjusted to the objectness score to alleviate the detector's bias towards the base categories. Simultaneously, $\mathcal{B}$ will be input into the category-specific knowledge distillation branch (CSKD) to learn the semantic knowledge and global category knowledge of VLM (*e.g.* CLIP). The CLIP embeddings are then filtered and updated to $\mathcal{S}$ to implement object mining. The overall training objective function is formulated as:

$$\mathcal{L}_{CAKE} = \mathcal{L}_{reg} + \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{CG-RPN} + \lambda_2 \mathcal{L}_{CSKD} \quad (1)$$

where $\mathcal{L}_{CG-RPN}$ is proposal generator loss, which includes box regression loss and classification loss. $\mathcal{L}_{CSKD}$ includes object-level distillation and cluster-level distillation. $\mathcal{L}_{reg}$ and $\mathcal{L}_{cls}$ denote the regression and classification losses generated by related heads, respectively.

Category-Generalized RPN

Most existing open vocabulary detectors rely on class-agnostic proposal generators (e.g., RPN (Ren et al. 2015) or Center head (Zhou, Koltun, and Krähenbühl 2021)) to discover foreground objects and it is expected to generalize to open categories, but the parameters are optimized to highlight regions of the base category and suppress other regions, which inevitably leads to overfit to the seen categories (Wang et al. 2023a). We propose to leverage the CLIP's well-structured visual space more fully to guide proposal generation, which is expected to adjust the proposal objectness scores generated by RPN, thus alleviating its bias against unknown categories. We call this process object mining because the regions that contain potential novel category objects with improved objectness scores will have a higher probability of being discovered and recognized by the detector. Similar to RPN from Faster R-CNN (Ren et al. 2015), CG-RPN has a 3×3 convolutional layer followed by two sibling 1×1 convolutional layers, generating box-delta $\Delta = (l, r, u, d)$ and foreground objectness score o, respectively. The difference is that we conduct object mining through a feature space Ω composed of high-quality features and calculate the foreground offset of the proposal to measure the correlation score μ that the proposal belongs to the foreground. Then we combine o and μ to obtain the final objectness score p for each proposal. The process of object mining is shown in Fig. 3.

Object Mining. Specifically, we have proposal embeddings $\{f_p\} \in \mathbb{R}^{|\mathcal{B}| \times d}$, and the feature set $\mathcal{S} = \{K_1, K_2, ..K_n\}$, where K_i denotes the feature cluster which contains several features. Then we use linear layer $\phi(\cdot)$ to fuse the features in the cluster to get the cluster center, ex-

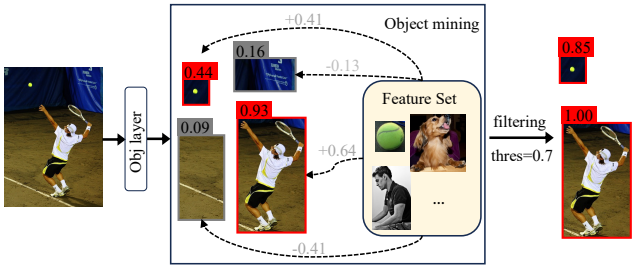


Figure 3: Illustration of object mining, we present proposals and corresponding scores. Positive proposals are denoted as red boxes, negative proposals are denoted as gray boxes. The origin objectness score o and final score $p \in (0, 1)$ are represented by black numbers, correlation coefficient μ is represented by gray numbers. The RPN threshold is set to 0.7.

pressed as:

$$\begin{aligned} \mathcal{W}_{\mathcal{S}} &= \{\phi(K_1), \phi(K_2), ..., \phi(K_n)\} \\ &= \{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_n} | \mathbf{w} \in \mathbb{R}^d\} \end{aligned} \tag{2}$$

Inspired by HRank (Lin et al. 2020; Liu et al. 2023), we use singular value decomposition (SVD) to decouple $\mathcal{W}_{\mathcal{S}}$ to extract main components then get the feature space Ω, which can be formulated as follows:

$$\Omega = V_r V_r^T, \quad where\ V\ satisfies\ \ \mathcal{W}_{\mathcal{S}} = U \Sigma V^T \tag{3}$$

where r is the rank of the feature matrix $\mathcal{W}_{\mathcal{S}}$, which equals the number of non-zero singular values of $\mathcal{W}_{\mathcal{S}}$. V_r corresponds to the first r columns of V and can be regarded as an orthogonal basis of the subspace of $\mathcal{W}_{\mathcal{S}}$, which can then be used to calculate the projection of the vector f_p on $\mathcal{W}_{\mathcal{S}}$. After we have the Ω, we calculate the correlation coefficient μ between f_p itself and the component projected onto Ω, and get the adjusted foreground probability p as:

$$\begin{aligned} \mu &= \sigma(cos(f_p, \Omega f_p)/\tau), \\ p &= max(min(o + \rho(\mu), 1), 0), \\ \mathcal{L}_{CG-RPN} &= \mathcal{L}_{BCE}(p, p^*) + \mathbb{1}_{p^*=1}\{\mathcal{L}_{IoU}(\Delta, \Delta^*)\} \end{aligned} \tag{4}$$

where τ denotes a temperature, $\sigma(\cdot)$ denotes the sigmoid function, $cos(\cdot)$ denotes cosine similarity and $\rho(\cdot)$ denotes the linear layer to revise. Its initialization is set to the identity transformation. p^* and Δ^* are the classification and regression assignments of GT boxes, respectively. For the detailed proof of Eq. 3, please see the supplemental material.

Category Specific Knowledge Distillation

In order to better utilize the visual language knowledge in VLM, many methods have explored different knowledge distillation methods. However, these methods mainly focus on the semantic information and the relationship of **objects**. We notice that the differences between objects of the same category will interfere with the detector's learning of unified global **category** knowledge, which is exactly what open vocabulary object detection focuses on for different categories. The method we proposed introduces extra cluster-level distillation based on the original object-level distillation. The cluster-level distillation is essentially a kind of knowledge with a strong correlation between categories, which is beneficial to help the detector establish an explicit category understanding and make the boundaries between different categories clearer.

Algorithm 1: Update Process of Feature Set

Input:
The CLIP embeddings $\widetilde{\mathcal{E}}$ in current batch;
The feature set $\mathcal{S}_t = \{K_1, K_2, ..., K_n\}$ in current time t, which include n clusters, each cluster contains a feature list with no more than l features with quality score;
The high-quality threshold T_{QS};
Output:
The updated feature set $\mathcal{S}_{t+1} = \{K_1, K_2, ..., K_{n'}\}$

Calculate the QS of $\widetilde{\mathcal{E}}$ and $\mathcal{S}_t$ according to Eq. 6.
for f_c in $\widetilde{\mathcal{E}}$ and $QS_e > T_{QS}$ **do**
  Get the corresponding cluster i, if there is no corresponding cluster, then create a new cluster for f_c
  if $|K_i| == l$ **then**
    $index, value = min(K_i, key = QS)$
    `// get the index of feature in` K_i `with minimum` QS
    if $QS_e > value$ **then**
      $K_i[index] \leftarrow f_c$
    end if
  else
    $K_i \leftarrow K_i \cup f_c$
  end if
end for
$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$
return $\mathcal{S}_{t+1}$

Distillation Branch. To this end, we designed the category-specific knowledge distillation framework (CSKD). CSKD includes two parts: object-level distillation(OD) and cluster-level distillation(CD).

Specifically, for region proposals $\{r\}$ output by CG-RPN, the roi head $\mathcal{R}$ generates pooled features, which are then projected to the semantic feature space shared by the text embedding $\mathcal{T}$ using the linear layer $\psi(\cdot)$. The results are denoted as the regional embeddings $\mathcal{E} = \psi(\mathcal{R}(r)) \in \mathbb{R}^{N \times d}$. Meanwhile, the CLIP image encoder $\mathcal{I}$ will calculate the CLIP embedding offline and normalize it to $\widetilde{\mathcal{E}} = \mathcal{I}(r) \in \mathbb{R}^{N \times d}$. In addition, we calculate the similarity matrix between feature embeddings, denoted as $S_{\mathcal{R}} = \mathcal{E}\mathcal{E}^T \in \mathbb{R}^{N \times N}$ and $S_{\mathcal{I}} = \widetilde{\mathcal{E}}\widetilde{\mathcal{E}}^T \in \mathbb{R}^{N \times N}$, respectively. Furthermore, we also explore text information to guide feature alignment. Specifically, we collect the similarity between the region embedding $\mathcal{E}$ and the fixed language embeddings $\mathcal{T}$ generated by CLIP to obtain the classification result of the corresponding proposals. The region embeddings belonging to the same category are clustered as described in Eq. 2, expressed as $\{\mathbf{w}\} \in \mathbb{R}^{K \times d}$ and $\{\widetilde{\mathbf{w}}\} \in \mathbb{R}^{K \times d}$. The cluster center is essentially the overall representation of the corresponding category, due to the richness of visual features, it has better generalization ability than directly using category

text embeddings. We then enclose the cluster center based on object-level distillation, and the training losses of object-level distillation(OD) and cluster-level distillation(CD) are respectively denoted as:

$$\begin{aligned}\mathcal{L}_{OD} &= \frac{1}{N}\sum_{i} \parallel \mathcal{E}_i - \widetilde{\mathcal{E}}_i \parallel + \frac{1}{N} \parallel S_{\mathcal{R}} - S_{\mathcal{I}} \parallel \\ \mathcal{L}_{CD} &= \frac{1}{K}\sum_{i} \parallel \mathbf{w_i} - \widetilde{\mathbf{w}}_\mathbf{i} \parallel \\ &\quad + \frac{1}{K}\sum_{\substack{i,j \\ i\neq j}} (\mathbf{1} - \parallel \mathbf{w_i} - \widetilde{\mathbf{w}}_\mathbf{j} \parallel)\end{aligned} \tag{5}$$

where the $\parallel \cdot \parallel$ denotes any normalized distance metric operation, usually using L1 distance or cosine similarity to measure the individual difference and use KL divergence or L2 distance to calculate the inter-embedding relations. N is the number of embeddings and K is the number of clusters. The distillation loss is $\mathcal{L}_{CSKD} = \mathcal{L}_{OD} + \lambda_3 \mathcal{L}_{CD}$.

Feature Set. We use the CLIP embeddings to construct a high-quality feature set $\mathcal{S}$, using it to implement object mining and guide the CG-RPN. The current CLIP embeddings are filtered and updated to $\mathcal{S}$ to modify the corresponding clusters. During the training process, $\mathcal{S}$ is continuously optimized to learn category-specific global features. Apparently, high-quality features are crucial in feature generation as the basis for guiding RPN. We argue that a high-quality feature should have these characteristics: credible and representative. The credibility can be expressed by objectness itself, the representativeness can be measured by the density of proposals. Thus we introduce a quality score QS, which is derived from the objectness and density of proposals. Considering that the early stage of training of the model is very noisy, we also introduce a time factor to ensure that newer features have a higher probability of being updated to $\mathcal{S}$.

Specifically, since the CLIP embedding and RoI features are derived from the same region, the objectness can directly utilize the results by CG-RPN. In addition, the density of proposals can be expressed by the proposals that are suppressed in the NMS stage. Based on the above description, assume the RoI features f_i corresponding to CLIP embedding $\widetilde{\mathcal{E}}_i$ that are suppressed in the NMS stage are represented as $\{\widetilde{f}_j\}$ (which could be empty set), the quality score QS of $\widetilde{\mathcal{E}}_i$ is defined as:

$$QS_i = (o_i + \sum_j \frac{e^{IoU_{i,j}}}{\sum_a e^{IoU_{i,a}}} o_j) \cdot \omega(t) \tag{6}$$

where the $IoU_{i,j}$ represents the intersection over the union between f_i and $\widetilde{f}_j$, $\omega(t)$ denotes a time-varying parameter that suppresses previously calculated scores as training progresses. The QS represents the quality credibility of the feature. The data in a batch that has a QS larger than the threshold T_{QS} is possible to update to $\mathcal{S}$ to replace the low-quality features in the corresponding cluster. The updating process is shown in Algorithm 1.

Experiments

Experimental Settings

Datasets. We mainly conduct experiments on LVIS v1 and COCO datasets, which are the datasets most often used in evaluating OVOD methods. For the open-vocabulary setting, we follow the approach of OV-RCNN and ViLD to split COCO and LVIS v1 into OV-COCO and OV-LVIS, respectively, and the annotations of novel categories are removed from the training set for both OV-COCO and OV-LVIS. OV-COCO is divided into 48 base categories and 17 novel categories and the remaining 15 categories are unused. For OV-LVIS, the rare categories are defined as novel categories(including 337 categories), and common and frequent categories are defined as base categories(including 886 categories).

Evaluation Metrics. We evaluate our method on the validation set for both OV-COCO and OV-LVIS. For OV-COCO, the main metric that evaluates the open-vocabulary detection performance is AP_{50}^{novel}, which denotes box mAP at IoU threshold 0.5 for novel categories. We also calculate AP_{50}^{base} for base categories and AP_{50} for all categories. For OV-LVIS, the main metric is AP_r, which denotes box mAP for rare categories. We also calculate AP_c for common categories, AP_f for frequent categories, and AP for all categories.

Implementation Details. We construct CAKE using Faster R-CNN and Mask R-CNN with ResNet-50 backbone for OV-COCO and OV-LVIS, respectively, and adopt Object-centric-OVD(OC-OVD) as our baseline method. For OV-COCO, we train the supervised model on 48 base categories for 1x schedule(90000 iterations) and report box AP_{50}. For OV-LVIS, we conduct our detector with federated loss (Zhou, Koltun, and Krähenbühl 2021) and sigmoid cross-entropy, following OC-OVD's approach, and report box AP_{50} and mask AP. The object-level distillation is done in the same way as OC-OVD and also adds the weight-transfer module for efficiency. For the feature set, each cluster contains no more than 10 features, their quality score is real-time and will be re-calculated every time they are accessed. The temperature τ is set to 5. Regarding CLIP, we choose the CLIP model based on ViT-B/32 (Radford et al. 2021). For the generation of category prompts, we employ hand-crafted prompts following ViLD (Gu et al. 2021). In OV-COCO experiments, the λ_1 to 1, λ_2 to 0.15, λ_3 to 0.5. In OV-LVIS experiments, the λ_1 to 1, λ_2 to 0.2, λ_3 to 0.25. These hyper-parameters are chosen through the randomized search on the corresponding datasets. In our experiments, we use SGD with a weight decay of $1e^{-4}$ and a momentum of 0.9, training on Nvidia 3090 with a batch size of 8, a mini-batch involving 2 images per GPU. We use a learning rate of 0.01 which drops by 10 at the 8th and 11th epochs.

Comparison with State-Of-The-Arts

OV-COCO. Tab. 1 compares our CAKE with other methods on the OV-COCO dataset. Similar to OADP (Wang et al. 2023b), we conducted extensive experiments on four OVD benchmarks: Vanilla OVD (V-OVD), Caption-based OVD (C-OVD), Generalized OVD (G-OVD), and Weakly Super-

Method	Benchmark	Detector	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
OADP (Wang et al. 2023b)	V-OVD	Faster R-CNN	30.0	53.3	47.2
DK-DETR (Li et al. 2023)	V-OVD	DeformableDETR	32.3	**61.1**	**53.6**
BARON (Wu et al. 2023a)	V-OVD	Faster R-CNN	33.1	54.8	49.1
LBP (Li et al. 2024b)	V-OVD	Faster R-CNN	37.8	58.7	53.2
CAKE (Ours)	V-OVD	Faster R-CNN	**38.2**	58.0	52.8
OV-DETR (Zang et al. 2022)	G-OVD	DeformableDETR	29.4	**61.0**	52.7
OADP (Wang et al. 2023b)	G-OVD	Faster R-CNN	35.6	55.8	50.5
VL-PLM (Zhao et al. 2022)	G-OVD	Faster R-CNN	32.3	54.0	48.3
CAKE (Ours)	G-OVD	Faster R-CNN	**39.1**	58.1	**53.1**
RegionCLIP (Zhong et al. 2022)	C-OVD	CLIP	26.8	54.8	47.5
CoDet (Ma et al. 2024)	C-OVD	CLIP	30.6	52.3	46.6
BARON (Wu et al. 2023a)	C-OVD	Faster R-CNN	35.8	58.2	52.3
CAKE (Ours)	C-OVD	Faster R-CNN	**41.3**	**60.2**	**55.3**
Detic (Zhou et al. 2022)	WS-OVD	Faster R-CNN	28.4	53.8	47.2
GOAT (Wang et al. 2023a)	WS-OVD	Faster R-CNN	36.4	53.0	48.6
OC-OVD (Bangalath et al. 2022)	WS-OVD	Faster R-CNN	36.6	54.0	49.4
CAKE (Ours)	WS-OVD	Faster R-CNN	**41.8**	**60.6**	**55.7**

Table 1: Comparision with other state-of-the-art methods on the OV-COCO dataset. The best results for each benchmark are shown in **bold** numbers, the best results over all benchmarks are shown in underlined numbers. Detector "CLIP" means the model uses CLIP visual encoder as backbone and CLIP textual encoder as classifier.

Method	Detector	Detection				Segmentation			
		AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
DetPro (Du et al. 2022)	Faster R-CNN	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OC-OVD (Bangalath et al. 2022)	Faster R-CNN	21.1	25.0	29.1	25.9	-	-	-	-
OADP (Wang et al. 2023a)	Faster R-CNN	21.9	28.4	32.0	28.7	21.7	26.3	29.0	26.6
CORA (Wu et al. 2023b)	Deformable DETR	22.2	-	-	-	-	-	-	-
DK-DETR (Li et al. 2023)	Deformable DETR	22.2	32.0	**40.2**	33.5	20.5	28.9	**35.4**	**30.0**
BARON (Wu et al. 2023a)	Faster R-CNN	23.2	29.3	32.5	29.5	22.6	27.6	29.8	27.6
CoDet (Ma et al. 2024)	CLIP	23.4	30.0	34.6	30.7	-	-	-	-
LBP (Li et al. 2024b)	Faster R-CNN	24.1	29.5	32.8	29.9	23.7	27.7	30.1	28.0
CAKE (Ours)	Faster R-CNN	**25.0**	**34.8**	38.4	**34.9**	**23.9**	**29.1**	33.6	28.7

Table 2: Comparision with other state-of-the-art methods on the OV-LVIS dataset. The best results are shown in **bold** numbers.

vised OVD (WS-OVD). In the V-OVD setting, CAKE construct feature set without weakly-supervised data and human prior on novel categories, the experiment shows the great generalization performance of our proposed method. CAKE does not reach the best performance on base categories. We argue the reason is our method alleviates the overfitting on base categories, this problem is solved when using more training data, just as in WS-OVD. The CAKE on WS-OVD setting reaches the state-of-the-art performance on both novel categories and all categories among all Faster R-CNN based methods, and it can also surpass some methods with more powerful backbones(ViT (Maaz et al. 2021), Swin-T (Liu et al. 2021), *etc.*) or DeformableDETR (Zhu et al. 2020), a detector known for its superiority over Faster R-CNN.

OV-LVIS. Tab. 2 shows the CAKE performance on OV-LVIS dataset. Compared to OV-COCO, OV-LVIS has a larger number of instances and categories but fewer implicit novel instances. The CAKE boosts OC-OVD's performance by 18.4% on AP_r. We also introduce the segmentation version which doesn't exist in the baseline method and outperforms the latest methods, demonstrating the model's robust ability to recognize a wide range of categories.

Plug-and-Play evaluation. To demonstrate the transfer ability of CAKE, we implement CSKD and CG-RPN on several methods and report the box AP_{50} on the OV-COCO dataset. The experiment settings follow the corresponding baseline. We also show the comparison of another plug-and-play method RALF (Kim et al. 2024), the latest and powerful trick that can improve OVOD model performance. Related results are shown in Tab. 3.

Ablation Study

In this part, we conduct several ablation studies to validate the effectiveness of modules in CAKE. The ablation studies include the effectiveness of CSKD and CG-RPN, hyperpa-

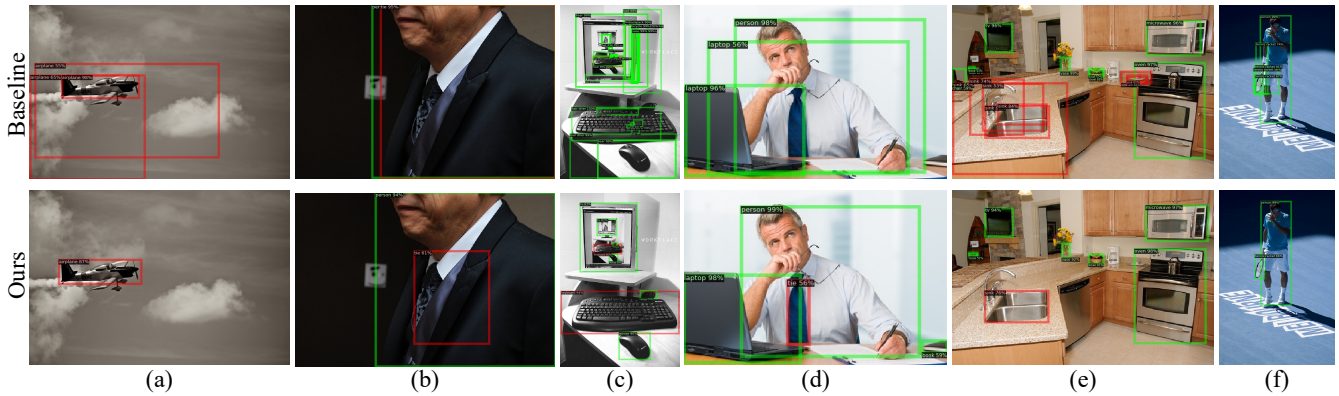


Figure 4: The visualization comparison between CAKE (bottom row) and baseline OC-OVD (top row). The objects belonging to novel categories are highlighted in red colors and base categories objects are green boxes.

Method	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
OC-OVD (Bangalath et al. 2022)	36.6	54.0	49.4
OC-OVD + RALF	41.3	54.3	50.9
OC-OVD + CAKE	**41.8**	60.6	55.7
OADP (Wang et al. 2023a)	30.0	53.3	47.2
OADP + RALF	33.4	54.5	49.0
OADP + CAKE	**36.1**	57.2	51.7
BARON (Wu et al. 2023a)	35.8	58.2	52.3
BARON + RALF	-	-	-
BARON + CAKE	**39.4**	60.0	54.6

Table 3: Results of plug-and-play on OV-COCO. Our proposed CAKE can adapt well to a variety of models and significantly improve their performance.

CSKD		CG-RPN	G-OVD	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
OD	CD					
-	-	-	-	30.4	52.6	46.8
✓	-	-	-	36.6	54.0	49.4
-	✓	-	-	32.5	54.3	48.6
✓	✓	-	-	40.1	56.7	52.4
✓	✓	✓	-	41.3	60.2	55.3
✓	✓	✓	✓	**41.8**	60.6	55.7

Table 4: Effect of individual components in CAKE. The G-OVD option represents whether novel category text embeddings are included or not. The OD includes both object-level distillation and weight-transfer in OC-OVD.

rameters result for $\lambda_1, \lambda_2, \lambda_3, l, \tau, T_{QS}$, the detailed result for CG-RPN. Note that all ablation studies are conducted on the WS-OVD setting and OV-COCO dataset. The result of the first experiment is reported in Tab. 4, the last two experiments are available in Supplementary Material.

In the distillation branch, OD and CD contribute 6.2 and 2.1 AP_{50}^{novel} respectively. When the two modules work together, the model shows great performance on both novel and base categories, reflecting marvelous interaction between two distillation methods. CG-RPN can further improve the model's perception of multiple categories and achieve 41.8 AP_{50}^{novel} when introducing novel category prior knowledge. Noting that novel category prior knowledge does not significantly improve model performance, we believe that CAKE can build a highly generalized knowledge structure based on a large number of image-text pairs, and can even learn general knowledge of foreground features by relying only on base category information.

Visualization

We compare the predictions of CAKE and baseline OC-OVD (Bangalath et al. 2022) on OV-COCO in Fig. 4. Our proposed CAKE can detect more diverse objects of different sizes, which the baseline approach often ignores. For example, CAKE recognizes slender ties in (b) and (d), keyboards on the table, and tv in (c). Moreover, CAKE has a clearer understanding of objects, that would not use duplicate boxes to mark the same object, like the airplane in (a), the person in (d), the sink in (e), and the tennis racket in (f). CAKE also has more accurate predictions. For example, the OC-OVD can detect tie in (b), but the box given by OC-OVD actually frames the whole person. In comparison, our method is more suitable. In general, CAKE achieved more comprehensive and accurate results than OC-OVD in detecting objects of novel categories.

Conclusion

In this paper, we introduce the Category Aware Knowledge Extraction (CAKE) framework for open-vocabulary object detection, comprising two main components: the Category-Specific Knowledge Distillation (CSKD) branch and the Category Generalization Region Proposal Network (CG-RPN). The CSKD branch constructs a high-quality category-specific feature set by extracting and refining category-specific information, addressing intra-category exclusion issues. The CG-RPN utilizes the feature set to explore and detect novel categories more effectively. Extensive experiments on the OV-COCO and OV-LVIS benchmarks show that CAKE reach state-of-the-art performance.

Acknowledgments

This work was supported by National Science and Technology Major Project (No. 2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No.2022J06001).

References

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35: 33781–33794.

Cai, Z.; Kwon, G.; Ravichandran, A.; Bas, E.; Tu, Z.; Bhotika, R.; and Soatto, S. 2022. X-detr: A versatile architecture for instance-wise vision-language tasks. In *European Conference on Computer Vision*, 290–308. Springer.

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30.

Chen, Z.; Ding, J.; Cao, L.; Shen, Y.; Zhang, S.; Jiang, G.; and Ji, R. 2023. Category-aware Allocation Transformer for Weakly Supervised Object Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6643–6652.

Chen, Z.; Wang, C.; Wang, Y.; Jiang, G.; Shen, Y.; Tai, Y.; Wang, C.; Zhang, W.; and Cao, L. 2022. LCTR: On Awakening the local continuity of transformer for weakly supervised object localization. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 410–418.

Chen, Z.; Wang, S.; Cao, L.; Shen, Y.; and Ji, R. 2024. Adaptive Zone Learning for Weakly Supervised Object Localization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14093.

Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; and Xu, C. 2021. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2154–2164.

He, B.; and Ozay, M. 2022. Feature kernel distillation. In *International Conference on Learning Representations*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1780–1790.

Kim, D.; Angelova, A.; and Kuo, W. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11144–11154.

Kim, J.; Cho, E.; Kim, S.; and Kim, H. J. 2024. Retrieval-Augmented Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17427–17436.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2022. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*.

Li, J.; Zhang, J.; Li, J.; Li, G.; Liu, S.; Lin, L.; and Li, G. 2024a. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16678–16687.

Li, J.; Zhang, J.; Li, J.; Li, G.; Liu, S.; Lin, L.; and Li, G. 2024b. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16678–16687.

Li, L.; Miao, J.; Shi, D.; Tan, W.; Ren, Y.; Yang, Y.; and Pu, S. 2023. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6501–6510.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; and Cai, J. 2022. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*.

Lin, J.; Shen, Y.; Wang, B.; Lin, S.; Li, K.; and Cao, L. 2024. Weakly supervised open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3404–3412.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. Hrank: Filter pruning using high-rank feature map. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1529–1538.

Liu, Y.; Wang, J.; Huang, C.; Wang, Y.; and Xu, Y. 2023. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23776–23786.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, C.; Jiang, Y.; Wen, X.; Yuan, Z.; and Qi, X. 2024. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14074–14083.

Maaz, M.; Rasheed, H. B.; Khan, S. H.; Khan, F. S.; Anwer, R. M.; and Yang, M.-H. 2021. Multi-modal transformers excel at class-agnostic object detection. *arXiv*.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14482–14491.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, 728–755. Springer.

Qu, Y.; Dai, S.; Li, X.; Lin, J.; Cao, L.; Zhang, S.; and Ji, R. 2024. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5328–5337.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

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

Wang, J.; Zhang, H.; Hong, H.; Jin, X.; He, Y.; Xue, H.; and Zhao, Z. 2023a. Open-Vocabulary Object Detection With an Open Corpus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6759–6769.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023b. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11186–11196.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 2023a. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15254–15264.

Wu, X.; Zhu, F.; Zhao, R.; and Li, H. 2023b. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7031–7040.

Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; and Luo, P. 2021. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8392–8401.

Xu, S.; Li, X.; Wu, S.; Zhang, W.; Li, Y.; Cheng, G.; Tong, Y.; Chen, K.; and Loy, C. C. 2023. Dst-det: Simple dynamic self-training for open-vocabulary object detection. *arXiv preprint arXiv:2310.01393*.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2022. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4643–4652.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; and Xu, H. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35: 9125–9138.

Yao, L.; Pi, R.; Xu, H.; Zhang, W.; Li, Z.; and Zhang, T. 2021. G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3591–3600.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 106–122. Springer.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14393–14402.

Zhao, S.; Zhang, Z.; Schulter, S.; Zhao, L.; Vijay Kumar, B.; Stathopoulos, A.; Chandraker, M.; and Metaxas, D. N. 2022. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, 159–175. Springer.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16793–16803.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. Springer.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2021. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.