Open-Vocabulary Object Detection with Meta Prompt Representation and Instance Contrastive Optimization

Zhao Wang[1]
zwang21@cse.cuhk.edu.hk
Aoxue Li[2]
lax@pku.edu.cn
Fengwei Zhou[2]
fzhou@connect.ust.hk
Zhenguo Li[2]
li.zhenguo@huawei.com
Qi Dou[1]
qidou@cuhk.edu.hk

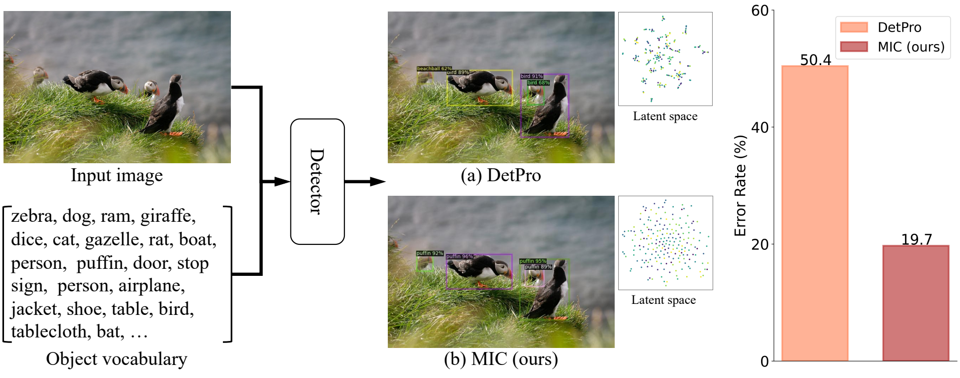
[1] Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
[2] Huawei, China

Abstract

Classical object detectors are incapable of detecting novel class objects that are not encountered before. Regarding this issue, Open-Vocabulary Object Detection (OVOD) is proposed, which aims to detect the objects in the candidate class list. However, current OVOD models are suffering from overfitting on the base classes, heavily relying on the large-scale extra data, and complex training process. To overcome these issues, we propose a novel framework with Meta prompt and Instance Contrastive learning (MIC) schemes. Firstly, we simulate a novel-class-emerging scenario to help the prompt learner that learns class and background prompts generalize to novel classes. Secondly, we design an instance-level contrastive strategy to promote intra-class compactness and inter-class separation, which benefits generalization of the detector to novel class objects. Without using knowledge distillation, ensemble model or extra training data during detector training, our proposed MIC outperforms previous SOTA methods trained with these complex techniques on LVIS. Most importantly, MIC shows great generalization ability on novel classes, *e.g.*, with +4.3% and +1.9% AP improvement compared with previous SOTA on COCO and Objects365, respectively.

1 Introduction

Deep learning models have been successful in closed-set large-scale object detection, in which the carefully designed detectors [2, 15, 27, 43] can accurately localize and classify the objects learned from the training set. However, these classical detectors always fail to generalize to unseen novel class objects during inference. Thanks to recent advancements in vision-language models [18, 21, 26], ViLD [11] extends the traditional closed-set object



(a) (b)

Figure 1: (a) In OVOD, the detector aims to detect any objects within an object vocabulary in an input image. Previous method, *e.g.*, DetPro, can easily misclassify some highly similar classes (puffin v.s. bird). Our method improves the model generalization ability, which can be more discriminative to these similar categories. Note that every point indicates a category in the latent space; (b) The error rate of predicting novel objects as base ones.

detection to an open-set scenario, named Open-Vocabulary Object Detection (OVOD). Under this OVOD setting, the detector is trained with only base classes, then tested on both base and novel classes. In ViLD, an object vocabulary (see Figure. 1a) is given to obtain text embeddings (a.k.a. class embeddings) through pretrained text encoder. The object class is predicted by finding the best matched one from the embedded object vocabulary, and the bounding box is obtained from the class-agnostic regression head.

From recent works on OVOD [11, 42], it has been widely studied that the proposal generator in a trained detector can generalize to novel classes well even the detector is trained only on the base classes. However, the current performance of open-vocabulary object detector degenerates and results in low AP when generalized to novel classes. Intuitively, the low performance is caused by the uncertain matching between proposal features and class embeddings. Under the OVOD setting with a large amount of classes, as shown in Figure 1a, many classes (*e.g.*, puffin v.s. bird) are very similar, which may lead to mismatching between proposal features and class embeddings. Furthermore, as the detector can only be learned on the base classes, the detector fails to recognize unseen novel classes. There can be a set of highly similar candidate classes (including both base and novel classes) when the detector meets a novel class object, which brings high uncertainty to novel object class prediction. This infers too close clustering of data points from similar classes in the latent feature space, and thus the decision boundary crosses high density regions [9, 39]. Meanwhile, the trained detector may misclassify a novel class object as background [13, 19]. Current OVOD methods can not handle these issues effectively. DetPro [6] makes class embeddings learnable, while other works [1, 8, 42] train the model with more classes. Since only the base classes are given during training, the corresponding class embeddings in DetPro [6] result in overfitting on the base classes, which has been pointed out in few-shot vision-language learning [41]. In [1, 8, 42], they supplement lots of extra data from ImageNet21k [5], Conceptual Captions [30], or LAION-400M [28] in the training process, but it is unfair by including novel classes during training as demonstrated in OWL-ViT [25].

To alleviate the overfitting issue more efficiently, we propose a meta prompt learning scheme. Specifically, during each training iteration, in addition to batch-wise training data and its annotations, we randomly sample more class names from base object vocabulary to simulate the novel-class-emerging scenario as in OVOD setting. Then, these enriched text samples are used to learn prompts of text representations. With meta representation sampling scheme, we can obtain more discriminative text representations and thus improve the generalization ability to novel classes. Meanwhile, instead of randomly initializing the background class embedding, we learn the background prompt representation. With the learnable background class embedding, the detector can better distinguish the negative (background) proposals from positive (foreground) ones. The object vocabulary list can be extended with more easily-accessed class names and to further alleviate the overfitting issue (see Table 1). Further, we incorporate an instance-level contrastive learning scheme to promote the intra-class compactness and inter-class separation, which expands low-density regions in the latent feature space by narrowing the cluster of base classes during the detector training. With such instance contrastive learning strategy, the novel classes are potentially separated from the base ones in the latent space and thus benefits novel class generalization. As shown in Figure 1b, MIC decreases the error rate of predicting novel objects as base ones significantly, from 50.4% to 19.7%. These results demonstrate the effectiveness of our method for alleviating the extreme overfitting issue.

Our main contributions are summarized as: 1) We introduce a novel meta prompt learning scheme to simulate a novel-class-emerging scenario, which boosts the generalization ability. Meanwhile, the learnable background prompt is incorporated to help the detector distinguish the positive and negative proposals. 2) We propose an instance-level contrastive learning strategy to promote intra-class compactness and inter-class separation, in which the contrastive pairs are built among foreground and background proposal samples. 3) We conduct extensive experiments on the benchmark dataset LVIS. Without knowledge distillation, ensemble model or extra training data, our method outperforms previous SOTA methods equipped with these complex techniques. Most importantly, our method shows great generalization ability on directly transferring to other datasets, such as COCO and Objects365.

2 Related Work

Open-Vocabulary Object Detection. Classical object detectors [2, 15, 23, 24, 27, 34, 43] heavily rely on the large-scale training data and can not generalize to unseen classes during inference. To alleviate these issues, lots of specific methods have been designed, such as semi/self-supervised [31, 33], few/zero-shot [17, 32], and open-world detection [14, 19]. However, zero-shot detection can only tackle in-domain setting, such as splitting COCO as seen and unseen classes, while open-world detection can only detect the unknown objects without classifying them. Recently, ViLD [11] proposes a new setting called Open-Vocabulary Object Detection (OVOD), which aims to detect any objects within an object vocabulary in an input image. To achieve this, ViLD replaces the traditional classifier weights with the class embeddings generated by CLIP [26] to make it generalize to novel classes. Following ViLD, some works are done to improve the detection performance by learning prompt representations [6] or supplementing extra training data [1, 8, 42]. For the former, the learned prompt representation can be easily overfitting on the base classes, which can be harmful to the generalization ability of the detector. For the latter, it is unfair to train on the novel classes from extra data, which is not strictly OVOD setting [25].

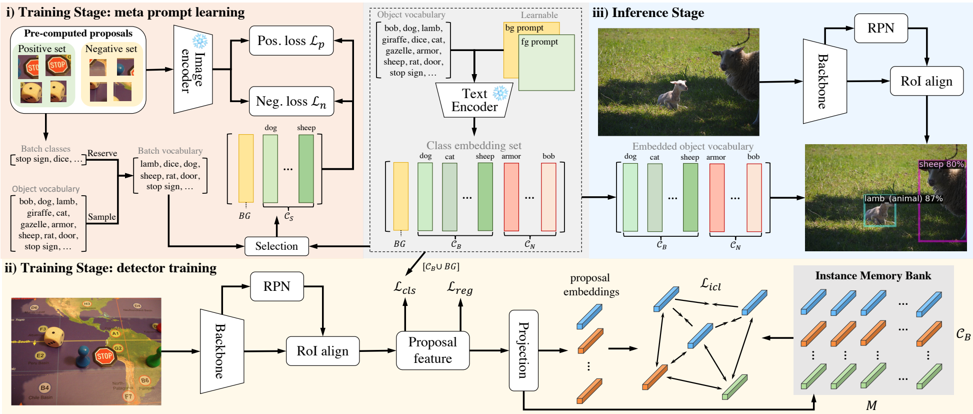


Figure 2: **Overview of our proposed method.** The training stage is divided into two consecutive parts: i) meta prompt learning and ii) detector training. During i) meta prompt learning, to simulate a novel-class-emerging scenario, we sample a batch-wise varying object vocabulary with $\mathcal{C}_S$ from $\mathcal{C}_B$, which improves the generalization ability of learned foreground prompt. Also, we integrate the learnable background prompt to help the model distinguish foreground and background proposals. Further, in ii) detector training, we introduce an instance-level contrastive learning scheme to promote intra-class compactness and inter-class separation. During iii) inference stage, we use the learned foreground prompt representation to generate class embeddings for novel classes.

Prompt Learning. Prompt learning is a light-weight framework to adapt large vision-language pretrained models[18, 21, 26] to downstream task. CoOp [40] proposes to learn the prompt representations but not just use a human-designed prompt template. CoCoOp [41] further handles the issue in CoOp that the model overfits on the base classes during training and fails on the novel classes during inference. Naively transferring prompt learning to OVOD can also cause overfitting, but the instance-level adaption strategy in CoCoOp is impossible for OVOD with the limited GPU memory.

Contrastive Learning. Contrastive pretraining [3, 4, 10, 16] has been validated effective for learning discriminative feature representations. Such pretraining strategy is based on contrastive image pairs built from two different strong augmentations of an image. Further, to increase the diversity of contrasive image pairs, supervised contrastive learning [20] is proposed. It builds contrastive pairs across different images, which better exploit the inherent information from images. In this work, we take the superiority and extend it to instance-level contrastive learning, which promote intra-class compactness and inter-class separation.

3 Method

3.1 Preliminary

In OVOD, the model can only access base classes $\mathcal{C}_B$ during training and is expected to detect the objects within a wider class vocabulary $\mathcal{C}=\mathcal{C}_B \cup \mathcal{C}_N$, where $\mathcal{C}_N$ represents the set of novel classes. We adopt CLIP [26] as our vision-language model during meta prompt learning, in which the text encoder $E_{\mathcal{T}}(\cdot)$ produces the class embeddings and the image encoder $E_{\mathcal{I}}(\cdot)$

outputs the image embeddings. ViLD [11] takes the class embeddings obtained from prompt engineering, which is not effective enough for OVOD since the text encoder $E_{\mathcal{T}}(\cdot)$ is highly sensitive to prompt templates. Regarding this issue, DetPro [6] learns a class-shared foreground prompt representation following CoOp [40] instead of using human-defined prompt templates. The prompt representation $\boldsymbol{V}_c$ for class $c \in \mathcal{C}_B$ is $\boldsymbol{V}_c = \left\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{L_p}, \boldsymbol{w}_c\right\}$, where $\boldsymbol{v}_i$ denotes the i-th learnable context vector, L_p is the length of learnable context, and $\boldsymbol{w}_c$ is the fixed word embedding for class c. Here, context vectors can be analogue to the human-designed prompt, such as "a photo of". The word embedding $\boldsymbol{w}_c$ is obtained from the corresponding class name, and the learnable context vectors $\boldsymbol{V}_{fg} = \{\boldsymbol{v}_i\}_{i=1}^{L_p}$ are randomly initialized. The class embedding $\boldsymbol{t}_c$ of class c is obtained by $\boldsymbol{t}_c = E_{\mathcal{T}}(\boldsymbol{V}_c)$. Given a positive (foreground) proposal, it is fed into image encoder $E_{\mathcal{I}}(\cdot)$ to obtain its embedding $\boldsymbol{f}_p$, and the probability of $\boldsymbol{f}_p$ to be classified as class c is $p_c^p = \frac{\exp(\text{sim}(\boldsymbol{f}_p, \boldsymbol{t}_c)/\tau)}{\sum_{i \in \mathcal{C}_B} \exp(\text{sim}(\boldsymbol{f}_p, \boldsymbol{t}_i)/\tau)}$, where τ is a temperature parameter, and $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity. The cross-entropy loss is used to optimize the context vectors $\boldsymbol{V}_{fg}$ while the base CLIP is fixed:

$$\mathcal{L}_p = -\log p_c^p. \tag{1}$$

Negative (background) proposals can not be recognized as any foreground objects. Given a negative proposal, its embedding $\boldsymbol{f}_n$ should be dissimilar to any foreground class embeddings. In practice, with a large $|\mathcal{C}_B|$, simply optimizing prediction probability p_c^n of a negative proposal to $\frac{1}{|\mathcal{C}_B|}$ can achieve this goal. Thus, the negative loss is $\mathcal{L}_\text{n} = -\frac{1}{|\mathcal{C}_B|}\sum_{c=1}^{|\mathcal{C}_B|} \log p_c^n$, where p_c^n is computed in the same way as p_c^p by replacing $\boldsymbol{f}_p$ with $\boldsymbol{f}_n$.

3.2 Meta Prompt Learning

With learnable foreground prompt $\boldsymbol{V}_{fg}$, the model performs better compared with prompt engineering. However, one urging issue is that the learned context can be easily overfitting on base classes $\mathcal{C}_B$ and lack of generalization ability to novel classes $\mathcal{C}_N$. In this section, we introduce a Meta Prompt Learning (MPL) scheme to promote generalization, as shown in Figure 2.

Meta Representation Sampling. Recall that there are $|\mathcal{C}_N|$ novel classes during inference in OVOD. Regarding this, we aim to simulate such a novel-class-emerging scenario to promote the prompt representation generalizing to novel classes well, in which a batch-wise varying vocabulary $\mathcal{C}_S$ is sampled from the base class vocabulary $\mathcal{C}_B$. Specifically, given a batch of proposal samples during prompt learning, with the base class embeddings $\boldsymbol{T}_B = \{\boldsymbol{t}_i\}_{i \in \mathcal{C}_B}$, we sample a subset $\boldsymbol{T}_S$=$\{\boldsymbol{t}_i\}_{i \in \mathcal{C}_S}$ of $\boldsymbol{T}_B$ ($\boldsymbol{T}_S \subset \boldsymbol{T}_B$). In this subset $\boldsymbol{T}_S$, the class embeddings of classes existed in the current batch proposal samples are reserved, and other class embeddings are randomly sampled from the remaining base classes. Then the probability of $\boldsymbol{f}_p$ to be classified as the corresponding class c is changed to

$$p_c^p = \frac{\exp\left(\text{sim}\left(\boldsymbol{f}_p, \boldsymbol{t}_c\right)/\tau\right)}{\sum_{i \in \mathcal{C}_S} \exp\left(\text{sim}\left(\boldsymbol{f}_p, \boldsymbol{t}_i\right)/\tau\right)}. \tag{2}$$

With batch-wise varying class embeddings, the learned $\boldsymbol{V}_{fg}$ is more generalizable and robust to the unseen novel classes, which helps the generalization of detector to novel classes.

Background Prompt Representation. A typical challenge for a classical detector is how to distinguish the negative (background) proposals from positive (foreground) ones, which becomes more serious in OVOD. With only base classes during training, detector will easily misclassify novel class objects as background. Previous works [6, 42] randomly initialize the background class embedding, while the randomness of the background class embedding can mislead the judgement of detector. To this end, we introduce the learnable background prompt representation with L_n learnable context vectors: $\boldsymbol{V}_{bg} = \{\boldsymbol{v}_1^{bg}, \boldsymbol{v}_2^{bg}, \ldots, \boldsymbol{v}_{L_n}^{bg}\}$. Also, we can obtain the background class embedding $\boldsymbol{t}_{bg} = E_{\mathcal{T}}(\boldsymbol{V}_{bg})$. Although the background proposals can not be recognized as the foreground objects, sometimes objects are partially located in the background proposals. So directly including the background proposals into prompt learning can be harmful. Instead, we consider the background prompt as a negative item to foreground classes. Then, Eq. (2) becomes

$$p_c^p = \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{f}_p, \boldsymbol{t}_c\right)/\tau\right)}{\sum\limits_{i \in \mathcal{C}_S} \exp\left(\mathrm{sim}\left(\boldsymbol{f}_p, \boldsymbol{t}_i\right)/\tau\right) + \exp\left(\mathrm{sim}\left(\boldsymbol{f}_p, \boldsymbol{t}_{bg}\right)/\tau\right)}. \tag{3}$$

We use $\mathcal{L}_p$ in Eq. (1) to optimize p_c^p in Eq. (3). For the negative proposals, the prediction probability is computed by

$$p_c^n = \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{f}_n, \boldsymbol{t}_c\right)/\tau\right)}{\sum_{i \in \mathcal{C}_S} \exp\left(\mathrm{sim}\left(\boldsymbol{f}_n, \boldsymbol{t}_i\right)/\tau\right)}, \tag{4}$$

and we use the following loss to optimize p_c^n:

$$\mathcal{L}_{\mathrm{n}} = -\frac{1}{|\mathcal{C}_S|} \sum_{c=1}^{|\mathcal{C}_S|} \log p_c^n. \tag{5}$$

3.3 Instance Contrastive Learning

With $\boldsymbol{V}_{fg}$ and $\boldsymbol{V}_{bg}$ learned in MPL, we then train our detector as shown in Figure 2. As the novel classes are not seen during training, the detector can easily classify a novel class object as a base class one. Such phenomenon reflects the decision boundary crosses the high density regions by mistake [36]. In this section, we propose Instance Contrastive Learning (ICL), aiming to expand low-density regions in the latent space by narrowing the cluster of base classes, which potentially pulls the cluster of novel classes from base ones.

Class-balanced Memory Bank. Given a large object class vocabulary, previous contrastive learning methods [4, 16] use a huge batch size (*e.g.*, 8192) to learn with more samples, which is unaffordable for detection. Even we perform instance-level contrastive learning, the number of proposals contained in a single batch is only around 400. So regarding this issue, we build an instance memory bank $\mathcal{Q}$ to collect diverse proposal samples. Moreover, to avoid frequent class dominating, we make this memory bank class balanced, in which the memory bank $\mathcal{Q}_c$ for each base class $c \in \mathcal{C}_B$ and background class contains M proposal samples. The memory bank is updated every iteration as the following: i) We filter out the foreground proposals with high Intersection of Union (IoU) $> U_p$ and identify background proposals with low IoU $< U_n$. Here, we set large U_p and small U_n to make the proposal samples representative. ii) We sample m ($m < M$) proposal samples that are the most dissimilar with the existing ones in $\mathcal{Q}$ to enrich the diversity of the samples in $\mathcal{Q}$. Along with the learning process, the memory bank $\mathcal{Q}$ is maintained in a first in and first out manner.

Optimization. Directly applying compact regularization to the high dimension proposal embeddings may result in an over-constraint to the network that hinders its convergence. Therefore, we introduce a projection network $\mathcal{Z}_\phi$ to map the proposal embeddings $\boldsymbol{f}$ into another low dimensional space as z. Inspired by supervised contrastive learning [20], we propose an instance-level contrastive loss to learn compact embeddings of proposal samples. The contrastive loss is defined as

$$\mathcal{L}_{icl} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\left|\mathcal{Q}_{\boldsymbol{c}(i)}\right|}\sum_{j}^{\left|\mathcal{Q}_{\boldsymbol{c}(i)}\right|}\log\frac{\exp\left(z_i\cdot z_j/\gamma\right)}{\sum_k^{\left|\mathcal{A}_{\boldsymbol{c}(i)}\right|}\exp\left(z_i\cdot z_k/\gamma\right)}, \tag{6}$$

where N is the number of proposal samples, $\boldsymbol{c}(i)$ is the class label of i-th proposal sample, $\mathcal{Q}_{\boldsymbol{c}(i)}$ denotes the memory bank of class $\boldsymbol{c}(i)$, γ is a temperature hyperparameter, and $\mathcal{A}_{\boldsymbol{c}(i)} = \mathcal{Q}\backslash\mathcal{Q}_{\boldsymbol{c}(i)}$. By expanding the low density regions of base classes cluster in the hidden space, the detector can learn more robust and generalizable feature representations.

3.4 Training

The training process of the proposed method consists of two parts: i) meta prompt learning and ii) detector training. For meta prompt learning, the overall loss is $\mathcal{L}_{mpl} = \mathcal{L}_p + \mathcal{L}_n$, where $\mathcal{L}_p$ and $\mathcal{L}_n$ are defined in Eq. (1) and Eq. (5), respectively. For detector training, the overall loss is $\mathcal{L}_{det} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{reg} + \alpha\mathcal{L}_{icl}$, where $\mathcal{L}_{rpn}$ is RPN loss, $\mathcal{L}_{reg}$ is regression $L1$ loss, $\mathcal{L}_{cls}$ is classification cross-entropy loss, $\mathcal{L}_{icl}$ is defined in Eq. (6), and $\mathcal{L}_{icl}$ is weighted by α. The overall training algorithm is given in the appendix.

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on the large-scale open-vocabulary benchmark LVIS [12]. LVIS v1 is a large-scale dataset with 1203 categories for object detection and instance segmentation task. The long-tailed distribution of LVIS dataset is very suitable for OVOD setting. We take frequent and common classes as the base classes $\mathcal{C}_B$ (866 classes), and rare classes as the novel classes $\mathcal{C}_N$ (337 classes). This dataset contains 100k and 20k images for training and validation. The models are trained only on base classes and evaluated on both base and novel classes. We also conduct transfer experiments to validate the generalization ability by directly evaluating LVIS-trained model on Pascal VOC [7], COCO [22], and Objects365 [29]. The implementation details and training time comparison are described in the appendix.

Evaluation Metrics. Following the previous works [6, 11, 42], we use Average Precision (AP) to evaluate the performance of our model. For LVIS, we take AP_r as the main metric as it is the performance of model generalizing to the novel classes $\mathcal{C}_N$, and also report AP_c, AP_f, and AP. For transfer experiments, we report AP, AP_{50}, AP_{75}, AP_s, AP_m and AP_l.

4.2 Main Results

Experiment on LVIS. We select the most recent SOTA OVOD methods for comparison, including ViLD [11], RegionCLIP [38], DetPro [6], OV-DETR [37], PromptDet [8], Detic

Method	KD?	Ens?	Extra data?	Detection AP_r	Detection AP_c	Detection AP_f	Detection AP	Instance segmentation AP_r	Instance segmentation AP_c	Instance segmentation AP_f	Instance segmentation AP
ViLD [11]	yes	yes	no	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
RegionCLIP [38]	no	no	CC3M	17.1	27.4	34.0	28.2	-	-	-	-
DetPro [6]	yes	yes	no	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OV-DETR [37]	yes	no	no	-	-	-	-	17.4	25.0	32.5	26.6
PromptDet [8]	no	no	LAION-400M	-	-	-	-	19.0	18.5	25.8	21.4
Detic [42]	no	no	CC3M	-	-	-	-	19.8	-	-	31.0
Rasheed *et al.* [1]	yes	no	ImageNet21k	-	-	-	-	19.3	23.6	27.9	24.1
MIC (ours)	no	no	no	**22.1**	33.9	40.0	33.8	**20.3**	30.6	35.2	30.6
MIC* (ours)	no	no	100 class names	**22.9**	34.0	39.9	34.4	**20.8**	30.5	35.4	30.7

Table 1: **Comparison of our method with previous SOTA methods on LVIS benchmark.** Note: KD (knowledge distillation); Ens (ensemble model). * indicates we train the prompts with 100 extra class names during MPL.

Method	Pascal VOC AP_{50}	Pascal VOC AP_{75}	COCO AP	COCO AP_{50}	COCO AP_{75}	COCO AP_s	COCO AP_m	COCO AP_l	Objects365 AP	Objects365 AP_{50}	Objects365 AP_{75}	Objects365 AP_s	Objects365 AP_m	Objects365 AP_l
Supervised	78.5	49.0	46.5	67.6	50.9	27.1	67.6	77.7	25.6	38.6	28.0	16.0	28.1	36.7
ViLD [11]	73.9	57.9	34.1	52.3	36.5	21.6	38.9	46.1	11.5	17.8	12.3	4.2	11.1	17.8
DetPro [6]	**74.6**	57.9	34.9	53.8	37.4	22.5	39.6	46.3	12.1	18.8	12.9	4.5	11.5	18.6
MIC (ours)	73.0	**58.3**	**39.2**	**56.8**	**42.2**	**27.2**	**43.1**	**51.1**	**14.0**	**20.1**	**15.2**	**6.6**	**16.6**	**24.6**

Table 2: **Comparison of our method with previous SOTA methods on transfer experiments.** We directly evaluate LVIS-trained model on Pascal VOC test set, COCO validation set and Objects365 validation set, together with a supervised baseline.

[42], and Rasheed *et al.* [1]. From Table 1, it can be seen that without knowledge distillation and ensemble model, our method outperforms DetPro by 1.3% bbox AP_r and 0.5% mask AP_r. Moreover, without extra training data, our method outperforms PromptDet, Detic, and Rasheed *et al.* [1] by 1.3%, 0.5%, and 1.0% mask AP_r, respectively. We also train the prompts with extra 100 class names from ImageNet21k and find that bbox AP_r can be further improved to 22.9%. Such results demonstrate the effectiveness and robustness of our method when generalizing to novel classes.

Transfer Experiment. We further validate the generalization ability of our method by directly evaluating LVIS-trained model on Pascal VOC, COCO, and Objects365. We use the learned prompt representation and class names of corresponding dataset to generate the class embeddings. From Table 2, we can observe that our method improves the performance by a large margin, especially on more difficult COCO (+4.3% AP) and Objects365 (+1.9% AP).

4.3 Ablation Studies

In this subsection, we do comprehensive ablation experiments on LVIS. To study MPL and ICL schemes, we randomly sample a subset with 5k images from LVIS validation set for hyper-parameter selection.

Overall Analysis. We study the effect of different components in our proposed framework. As shown in Table 3, MPL helps learnable foreground prompt generalize, with +0.9% AP_r improvement. With the learnable background prompt, the detection performance on novel classes improves by 0.6%. Further, our proposed ICL boosts the performance AP_r to 22.1%, surpassing fixed prompt (+4.5% AP_r) and naive learnable prompt (+2.4% AP_r).

Number of Sampled Classes in MPL. We study the model performance when trained with different number of sampled classes in Figure 3. We range the number of classes from 200

Prompt FG	Prompt BG	Strategy MPL	Strategy ICL	Detection AP_r	Detection AP_c	Detection AP_f	Detection AP
fixed	✗	✗	✗	17.6	34.4	40.2	33.8
learnable	✗	✗	✗	19.7	34.0	39.8	33.8
learnable	✗	✓	✗	20.6	33.5	39.8	33.7
learnable	learnable	✓	✗	21.2	34.0	39.9	34.1
learnable	learnable	✓	✓	**22.1**	33.9	40.0	34.2

Table 3: **Effect of different components** of our approach. Note: FG (foreground prompt); BG (background prompt).

$[L_p, L_n]$	[4, 6]	[8, 10]	[16, 18]
AP_r	25.2	**26.4**	25.8
AP	39.3	40.1	39.7

(a) Context lengths

Position	Front	Middle	End
AP_r	23.8	25.4	**26.4**
AP	39.0	39.8	40.1

(b) Different positions of class token

Table 4: **Learnable context study.**

Figure 3: **Sampling strategy in MPL.** We study the effect of sampled classes.

U_n	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
U_p	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.9
AP_r	**26.4**	25.1	26.0	26.0	24.7	24.3	25.4	25.2	24.8
AP	40.1	40.0	39.8	39.9	40.1	40.2	39.7	39.9	39.4

(a) IoU threshold U_p and U_n

m	8	16	32	8	16	32	8	16	32
M	64	64	64	128	128	128	256	256	256
AP_r	24.3	24.5	24.6	26.0	25.5	24.9	25.7	**26.4**	23.9
AP	39.8	40.2	39.8	39.9	39.7	40.2	39.6	40.1	39.4

(b) Batch sampling size m and memory size M

Table 5: **Sampling strategy in ICL.** (a) Effect of foreground and background instance IoU threshold U_p and U_n. (b) Effect of batch sampling size m and memory size M.

to 850 in intervals of 50, along with a special case that only the batch classes are reserved (about 150 classes). We can learn that AP_r shows a trend of first rising and then falling as the sample size increases.

Context Lengths. We study the effect of using different foreground and background context lengths. Due to lack of class word embedding, we set L_n always longer than L_p by 2. As shown in Table 4a, too short context is not effective while too long context may cause over-fitting. We set the foreground and background context as 8 and 10, respectively.

Position of Class Token. In Table 4b, we study the effect of different positions of class token in the foreground prompt, including front, middle, and end. The best position of class token usually depends on the dataset [40]. We find that inserting it in the end performs best.

Sampling Thresholds in ICL. Here, we study the effect of selection thresholds under the different combinations of U_p and U_n, as shown in Table 4a. We can learn the following: 1) the model performs better with smaller background selection threshold which infers few overlap between proposal and gt bounding box; 2) the model performs best with foreground selection threshold 0.7, which demonstrates that too high threshold will filter out too many foreground proposals and lead to sub-optimal solution.

Memory Bank and Sampling Sizes. Further, we study the size of memory bank M and batch sampling size m. Intuitively, with more samples in the memory bank, the contrastive regularization can be stronger. We consider several different combinations of memory size M and batch sampling size m, and the results are shown in Table 4b. We can learn that with memory size 256 and batch sampling size 16, the detection achieves the best performance.

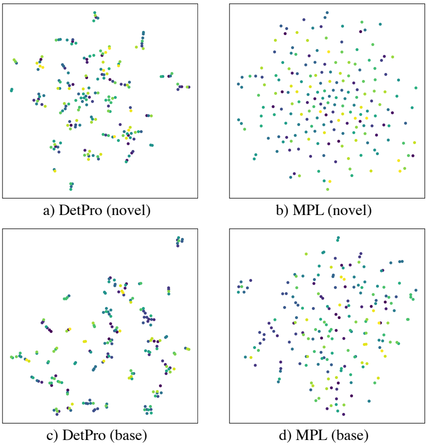


Figure 4: **t-SNE visualization of class embeddings of LVIS.** We randomly sample 200 novel and base classes from LVIS and use t-SNE to visualize the class embeddings.

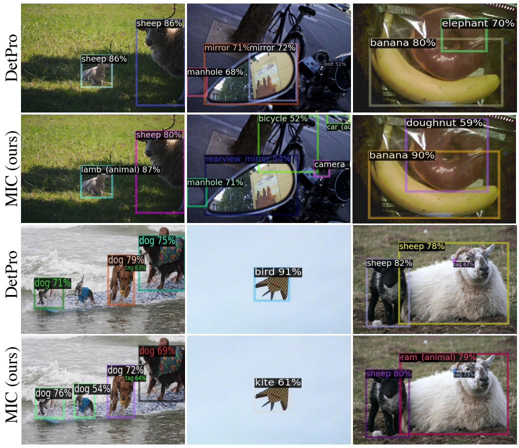


Figure 5: **Qualitative detection visualization results of our proposed method MIC and DetPro.** Our method could better distinguish similar classes, detect smaller objects, and produce less false positives under diverse complex scenarios.

4.4 Visualization Results

Latent Space Embedding. To further validate the effectiveness of our proposed MPL, we randomly sample 200 novel and base classes from LVIS and use t-SNE [35] to visualize the class embeddings from DetPro and our MPL. From Figure 4, it can be seen that MPL can learn more discriminative and generalizable class embeddings, especially for novel classes.

Detection Results. We show the superiority of our method by visualizing the detection results on LVIS validation set. The visualization results are shown in Figure 5. It can be seen that our method can better distinguish similar categories compared with DetPro, such as sheep and lamb. Further, our method can detect smaller objects in the complex environments. Also, with the learnable background prompt, our model produces less false positives.

5 Conclusion

This paper proposes a novel framework MIC for open-vocabulary object detection. MIC consists of two major carefully designed learning schemes, meta prompt and instance contrastive learning. The proposed meta prompt learning strategy simulates a novel-class-emerging scenario, together with the learnable background prompt representation to help the generalization. Further, we propose to use instance-level contrastive learning strategy to help expand the low density regions in the latent feature space. Without complex training techniques and extra training data, extensive experimental results show the strong generalization ability of our proposed method, especially transferring to other datasets, such as COCO and Objects365.

Acknowledgements. We gratefully acknowledge the support of Mindspore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

[1] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[6] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] 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(2):303–338, 2010.

[8] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *Proceedings of the European Conference on Computer Vision*, 2022.

[9] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU.

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[13] 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.

[14] 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.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[17] Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. Robust region feature synthesizer for zero-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7622–7631, 2022.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[19] 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.

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[22] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] 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.

[24] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021.

[25] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *Proceedings of the European Conference on Computer Vision*, 2022.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[27] 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.

[28] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[29] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[31] Yukun Su, Guosheng Lin, Yun Hao, Yiwen Cao, Wenjun Wang, and Qingyao Wu. Self-supervised object localization with joint graph partition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2289–2297, 2022.

[32] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021.

[33] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2291–2301, 2021.

[34] 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.

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[36] Tong Wang, Yousong Zhu, Yingying Chen, Chaoyang Zhao, Bin Yu, Jinqiao Wang, and Ming Tang. C2am loss: Chasing a better decision boundary for long-tail object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6980–6989, 2022.

[37] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 2022.

[38] 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.

[39] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022.

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[42] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *European Conference on Computer Vision*, 2022.

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.