Weakly Supervised Open-Vocabulary Object Detection

Jianghang Lin[1], Yunhang Shen[2], Bingquan Wang[1], Shaohui Lin[3], Ke Li[2], Liujuan Cao[1]*

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China.
[2]Tencent Youtu Lab, China.
[3]School of Computer Science and Technology, East China Normal University, China.
{hunterjlin007, wangbingquan}@stu.xmu.edu.cn,
{shenyunhang01, shaohuilin007, tristanli.sh}@gmail.com, caoliujuan@xmu.edu.cn,

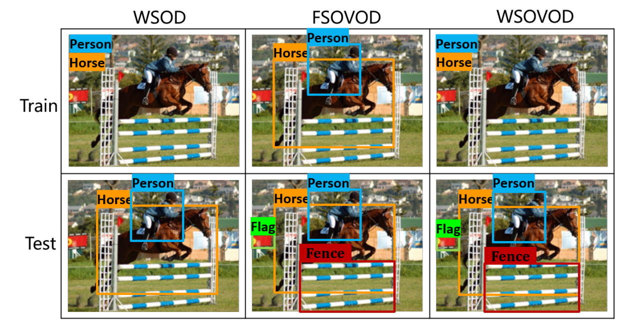
Abstract

Despite weakly supervised object detection (WSOD) being a promising step toward evading strong instance-level annotations, its capability is confined to closed-set categories within a single training dataset. In this paper, we propose a novel weakly supervised open-vocabulary object detection framework, namely WSOVOD, to extend traditional WSOD to detect novel concepts and utilize diverse datasets with only image-level annotations. To achieve this, we explore three vital strategies, including dataset-level feature adaptation, image-level salient object localization, and region-level vision-language alignment. First, we perform data-aware feature extraction to produce an input-conditional coefficient, which is leveraged into dataset attribute prototypes to identify dataset bias and help achieve cross-dataset generalization. Second, a customized location-oriented weakly supervised region proposal network is proposed to utilize high-level semantic layouts from the category-agnostic segment anything model to distinguish object boundaries. Lastly, we introduce a proposal-concept synchronized multiple-instance network, *i.e.*, object mining and refinement with visual-semantic alignment, to discover objects matched to the text embeddings of concepts. Extensive experiments on Pascal VOC and MS COCO demonstrate that the proposed WSOVOD achieves new state-of-the-art compared with previous WSOD methods in both close-set object localization and detection tasks. Meanwhile, WSOVOD enables cross-dataset and open-vocabulary learning to achieve on-par or even better performance than well-established fully-supervised open-vocabulary object detection (FSOVOD).

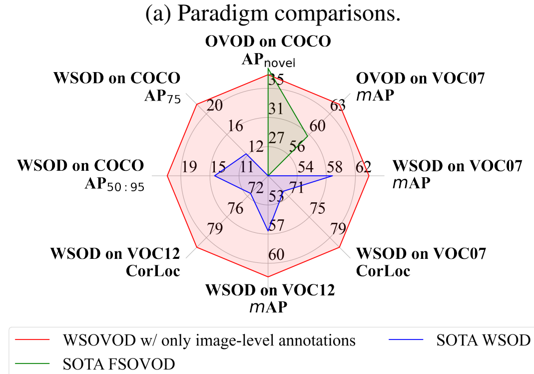
Introduction

In the past decades, the artificial intelligence community has witnessed great progress in object detection. In particular, large amount of human-annotated datasets significantly promotes the prosperity and progress of fully-supervised object detection (FSOD), such as Faster RCNN (Ren et al. 2015), YOLO (Redmon et al. 2016; Redmon and Farhadi 2017), Detr (Carion et al. 2020) and their variants (Zhu et al. 2021; Zheng et al. 2021; Sun et al. 2021). Nevertheless, the laborious and lavish collection of instance-level annotations has severely barricaded the applicability of FSOD in practical application with large-scale categories. By excavating image-level category supervision that indicates the presence or absence of an object, weakly supervised object detection (WSOD) has attracted much attention recently since image-level annotations are widely available in easily-collected classification-like datasets.

Unfortunately, a *de facto* limitation of the existing WSOD methods (Tang et al. 2017, 2018; Kim et al. 2020) stems from their detectors only concentrating on few categories



(a) Paradigm comparisons.



(b) Qualitative comparisons.

Figure 1: Vanilla WSOD is confined to detecting known categories from the training set, *e.g.*, *Person*, and *Horse*. The proposed WSOVOD is generalized to unseen categories, *e.g.*, *Flag*, and *Fence*. WSOVOD outperforms the previous state-of-the-art WSOD methods and achieves on-par or even better performance than FSOVOD.

*Corresponding Author

concepts within individual datasets, such as 20-category Pascal VOC (Everingham et al. 2010) and 80-category MS COCO (Lin et al. 2014). Little effort has been made to explore the limit of WSOD learning at scale toward detecting novel objects. Thus, it may not fully exploit the latent capacity of WSOD whose original intention is to leverage the tremendous amount of tagged images to train object detectors. To solve the above limitation, as illustrated in Fig. 1 (a), we extend WSOD settings to detect and localize open-vocabulary concepts using joint large-scale weakly-annotated datasets that are publicly available. Accordingly, a weakly supervised open-vocabulary object detection, referred to as WSOVOD is put forth.

To this end, three main challenges, as we start in this paper, obstacle to the implementation of WSOVOD. First, non-identical data distributions may bring dataset bias (Kim, Lee, and Choo 2021; Torralba and Efros 2011; Jiang et al. 2022) to affect the feature learning, hindering the vision-language alignment introduced as followed. For example, ILSVRC (Russakovsky et al. 2015) is an object-centric dataset with a balanced category distribution, while LVIS (Gupta, Dollár, and Girshick 2019) has many complex scenes with Zipfan distribution. Second, the reliance of existing WSOD methods upon traditional object proposal generators prevents models from learning proposal extraction at different semantic levels, since they only use low-level features computed on super-pixel (Felzenszwalb and Huttenlocher 2004) or counters (Dollár and Zitnick 2013). Third, weak supervision hardly aligns vision-language representation. In the existing open-vocabulary studies (Ma et al. 2022a; Gu et al. 2022; Zang et al. 2022), the visual-semantic alignments are realized in a fully-supervised manner where classification embeddings and box knowledge are necessary. Though recent methods (Zhou et al. 2022b; Kamath et al. 2021; Zareian et al. 2021) resort to weak information, *e.g.*, captions, they deeply rely on strong box-level annotations.

To solve the above three problems and overcome the limitations of common WSOD approaches, our WSOVOD framework (in Fig. 2) innovates in three aspects: 1) We extract data-aware features to generate for each image input-conditional coefficients and combine dataset attribute prototypes to identify dataset bias in proposal features of different distributions. Explicitly, an additional branch learns to squeeze the global image feature into a channel-wise global vector as coefficients to weight dataset attribute prototypes for re-calibrating final proposal features. 2) A location-oriented region proposal network is proposed to leverage high-level semantic layouts from the image segmenter to distinguish object boundaries. Recent interactive segmentation work SAM (Kirillov et al. 2023) exhibits strong image segmentation capabilities, but it lacks semantic recognition ability. Here, we transfer the knowledge from SAM to a customized region proposal network upon high-quality proposals. 3) We introduce a proposal-concept synchronized multiple-instance network that implements object mining to discover objects under image-level classification embeddings, as well as instance refinement to align vision-language representation. Specifically, we obtain text embeddings of the target vocabulary from the pre-trained text encoder, which are considered as category prototypes for multiple-instance learning. Also, we transform the multi-branch refinement heads in the common WSOD framework into open-vocabulary learning to further align object and concept representation. In addition, we leverage SAM to refine the box coordinates of the supervision between multi-branch refinement heads.

Extensive experiments demonstrate that the proposed WSOVOD achieves on-par or even better performance compared to fully-supervised open-vocabulary detection methods, which paves a new way to explore the large number of visual concepts from image-level supervisions. For example, our method significantly outperforms OVR-CNN (Zareian et al. 2021), ViLD (Gu et al. 2022) and Detic (Zhou et al. 2022b) that require box-level annotations of base categories, by 13.9%, 9.1% and 8.9% AP, respectively, for novel categories in MS COCO. Moreover, WSOVOD achieves new state-of-the-art performance compared to the previous WSOD methods under the close-set and single-dataset settings while being able to detect novel categories.

Related Work

Weakly Supervised Object Detection

Combining multiple-instance learning (MIL) (Dietterich, Lathrop, and Lozano-Pérez 1997) with convolutional neural networks (CNNs) has made great progress in WSOD. WSDDN (Bilen and Vedaldi 2016) is the prior work to introduce MIL into CNN and model WSOD as a proposal classification. However, WSDDN suffers from local optimization problems since the detector tends to detect high-activated regions. OICR (Tang et al. 2017) further attaches multi-branch refinement to WSDDN, which gradually propagates the scores of the salient regions to the complete objects. These methods are highly dependent on traditional proposal generation methods (Uijlings et al. 2013; Pont-Tuset et al. 2016) and do not regress the final proposal boxes. Furthermore, UWSOD (Shen et al. 2020a) learns multi-scale features and the region proposal network in an end-to-end unified framework. Nevertheless, the region proposal network is prone to be saturated due to the noisy pseudo-ground-truth boxes in the early training period, which has inferior performance than the cutting-edge WSOD methods. Different from these methods, we exploit knowledge transfer from the category-agnostic segmenter to pursue high-quality and high-recall object proposals.

Open-Vocabulary Object Detection

Open-vocabulary object detection (OVOD) (Zareian et al. 2021; Gu et al. 2022; Minderer et al. 2022; Zhou et al. 2022b) is an attractive research topic in recent years, whose goal is to detect unseen or novel classes that occupy a particular semantically-coherent region within an image. OVOD differs from zero-shot object detection (Bansal et al. 2018) in that it can access large-scale novel objects with weakly-supervised labels, *e.g.*, tags, and captions. However, they share the same paradigm of learning a cross-modal vision-language representation space to model image regions and

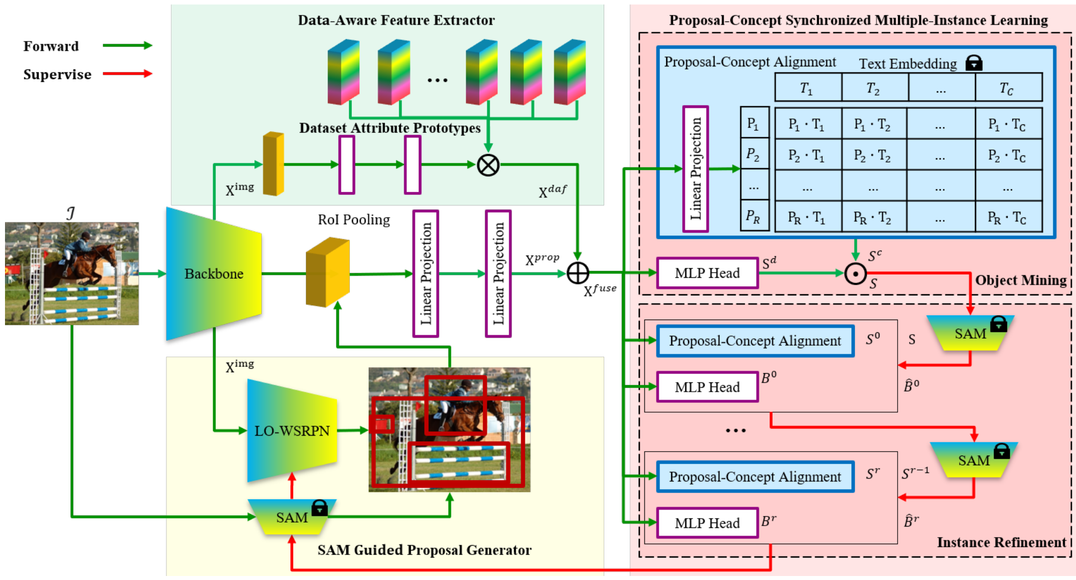


Figure 2: Illustration of the proposed WSOVOD framework. The proposal generator combines candidate regions from LO-WSRPN and SAM that may potentially contain objects for subsequent object mining. The data-aware feature extractor outputs unbiased dataset attribute features by identifying dataset bias from dataset attribute prototypes. The proposal-concept synchronized multiple-instance learning discovers potential objects that match the target vocabularies in image-level labels.

word descriptors. The main challenge in this field is aligning proposal features with category text embeddings, thus it is crucial to use image-text knowledge efficiently (Radford et al. 2021; Li et al. 2022). OVR-CNN (Zareian et al. 2021) pre-trains the detector on image-text pairs using contrastive learning and fine-tunes it on detection data with a limited vocabulary. OWL (Minderer et al. 2022) further transfers knowledge from vision-language models to transformer-based detectors with contrastive image-text pre-training and detection fine-tuning. Detic (Zhou et al. 2022b) improves OVOD performance of long-tail categories via image-level annotated data. Different from these approaches, our proposed WSOVOD uses MIL-based object mining to discover potential objects and refines them by multi-branch refinement open-vocabulary heads gradually. All of the above methods are highly dependent on bounding-box annotations, while WSOVOD is devoted to efficiently exploring weakly-annotated data.

Methodology

As illustrated in Fig. 2, an image $\mathcal{I}$ first goes through the vision backbone to extract global image features X^{img}. Then, the data-aware feature extractor takes in X^{img} to generate coefficients for combining dataset attribute prototypes as data-aware features X^{daf}. Meanwhile, a proposal generator also takes in X^{img} to hypothesize object locations. Next, RoI pooling crops the pooled features from global feature X^{img}, and two fully-connected layers transform them to get proposal features $X^{\text{prop}} \in \mathcal{R}^{R\times D}$, where R is proposal number in image $\mathcal{I}$ and D is feature vector length. We further fuse proposal features X^{prop} with data-aware features X^{daf} to deal with dataset bias, resulting in X^{fuse}. Finally, a proposal-concept synchronized multiple-instance learning takes in X^{fuse} to discover objects constrained by image-level classification embeddings and align representation between objects and concepts. The overall training objective function is formulated as:

$$\mathcal{L}_{\text{WSOVOD}} = \mathcal{L}_{\text{PG}} + \mathcal{L}_{\text{OM}} + \mathcal{L}_{\text{IR}}, \tag{1}$$

where $\mathcal{L}_{\text{PG}}$ is proposal generator loss. And $\mathcal{L}_{\text{OM}}$ and $\mathcal{L}_{\text{IR}}$ are object mining and instance refinement losses for proposal-concept synchronized multiple-instance learning.

Data-Aware Feature Extractor

To better align vision-language representation, it is necessary to learn as many categories as possible, however, an individual dataset contains limited concepts. This motivates us to train one detector upon multiple datasets jointly to generalize the detection scope of WSOD. The main challenge stems from domain incompatibility over non-identical data distribution. Much of the bias can be accounted for by the divergent goals of the different datasets: For example, LVIS (Gupta, Dollár, and Girshick 2019) has an average of 11.2 instances from 3.4 categories per image with long-tail Zipfan distribution, while most images in ILSVRC (Russakovsky et al. 2015) are object-centric with single category. Such variant dataset biases hurt the representation learning, thus simply combining multiple datasets has poor performance as observed in our experiments. In contrast, such bias could be well recognized even from a single image by humans and classifiers (Torralba and Efros 2011).

To this end, we design a data-aware feature extractor (DAFE) to generate generalized dataset-level features for cross-dataset learning with different scenarios and different categories. The key intuition of DAFE is to capture the unique and identifiable "signature" of each dataset conditioned on full-image information and adjust proposal features accordingly. Specifically, it consists of a global average pooling layer to squeeze the input information from image feature maps X^{img}. Then two fully-connected layers followed by the Tanh activation function learn to generate coefficients based on the image input to combine dataset attribute prototypes for identifying the dataset bias from the squeezed global features and produce data-aware feature X^{daf} with the same dimension of the proposal features. Finally, we aggregate X^{daf} with the proposal features X^{prop} by element-wise summation: $X^{\mathrm{fuse}} = X^{\mathrm{prop}} + X^{\mathrm{daf}}$. Thus, the input-conditional vector X^{daf} aims to re-calibrate the original proposal features to de-bias the different distributions, which are then used for the subsequent open-vocabulary object mining and refinement.

Discussion. The proposed DAFE, to some extent, is related to recent prompt tuning (Jia et al. 2022) that adapts large foundation models to downstream tasks with a small amount of task-specific learnable parameters. Our approach differs from theirs in two folds. First, most prompt learning methods perform data-space adaptation by transforming the input. For example, approaches in (Feng et al. 2022) append a sequence of learnable vectors to the textual input, and method in (Bahng et al. 2022) learns an image perturbation to convert the image to the formats of downstream tasks. Different from the above methods, our DAFE eliminates the different dataset distributions by feature-space adaptation with an input-conditional vector. Second, existing prompt tuning mainly focuses on fully-supervised learning, which is difficult to generalize to wide unseen categories. Input-conditional prompt learning (Zhou et al. 2022a) still relies on an online text encoder to generate input-specific weights for each image. Our adaptation does not require an online text encoder during training and inference.

SAM Guided Proposal Generator

Most existing WSOD methods use traditional proposal methods with low-level features to generate region candidates, which prevents the models from end-to-end learning with high-level semantic information. We design a location-oriented weakly supervised region proposal network (LO-WSRPN) to recognize category-agnostic potential objects, which further transfers knowledge of high-level semantic layouts from SAM (Kirillov et al. 2023). In detail, similar to RPN from Faster RCNN (Ren et al. 2015), LO-WSRPN has a 3×3 convolutional layer with 256 channels followed by three sibling 1×1 convolutional layers for localization and shape quality estimations, respectively. The first two convolutional layers are responsible for localization quality estimation, predicting centerness c and foreground probabilities p, respectively. We use $s = \sqrt{c \cdot p}$ as the localization quality for each region proposal during inference. The last convolutional layer is responsible for shape quality estimation. Different from anchor-based detectors, we directly view locations as training samples instead of anchor boxes. We replace the standard box-delta targets (x, y, h, w) with distances $t = (l, r, t, b)$ from the location to four sides of the ground-truth box as in (Tian et al. 2019). The training objective function of this module is formulated as:

$$\begin{aligned} \mathcal{L}_{\mathrm{PG}} = & \mathcal{L}_{\mathrm{BCE}}(p, p^*) + \\ & \mathbb{1}_{p^*=1}\{L_1(c, c^*) + \mathcal{L}_{\mathrm{IoU}}(t, t^*)\}, \end{aligned} \tag{2}$$

where $\mathcal{L}_{\mathrm{BCE}}$ is the binary cross-entropy loss function, L_1 constrains the distance between the sampling anchor points and the pseudo-ground-truth (PGT) boxes, $\mathcal{L}_{\mathrm{IoU}}$ measures the shape difference between the predicted boxes and the PGT boxes, thereby constraining the shape of the predicted boxes. We use WSOVOD's final predictions as PGT boxes and assign the corresponding targets, *i.e.*, p^*, c^*, and t^*.

However, object proposals from LO-WSRPN are extremely noisy in the early stage of training as observed in (Shen et al. 2020a), which has a negative impact on subsequent object mining and brings in inferior PGT boxes. Inspired by large-scale interactive segmentation models (Kirillov et al. 2023), we leverage SAM to generate additional proposals during training, which helps stabilize subsequent object mining. In detail, we first sprinkle evenly 32×32 grid points as the prompt input of SAM to generate additional proposals. Then, we concatenate the SAM proposals with the learned proposals from LO-WSRPN as input to subsequent object mining. Incorporating knowledge from SAM not only helps enrich the high-quality object proposals but also leverages high-level semantic layouts from the image segmenter to distinguish object boundaries.

Proposal-Concept Synchronized Multiple-Instance Network

The central idea of fully-supervised open-vocabulary object detection (FSOVOD) is to align object features with text embeddings which are pre-trained on large-scale image-text pairs like CLIP (Radford et al. 2021). In detail, FSOVOD methods convert a generic two-stage object detector to an open-vocabulary detector by replacing the learnable classifier head with fixed language embeddings, corresponding to the category names. Thus, object-level annotations are required during training to maximize the embedding similarities of positive region-category pairs and minimize that of negative ones. However, it is challenging to align object-level vision-language representation with only image-level supervision. Fortunately, WSOD is often formulated as multiple-instance learning (MIL) (Dietterich, Lathrop, and Lozano-Pérez 1997) and implicitly learns instance-based classifier from image-level information.

Therefore, our WSOVOD extends the common MIL-based WSOD framework (Bilen and Vedaldi 2016) to mine large-scale category concepts in an open-vocabulary manner. The fused proposal features X^{fuse} are forked into two fully-connected layers parallel, namely classification stream $W^c \in \mathcal{R}^{D \times C}$ and detection stream $W^d \in \mathcal{R}^{D \times C}$, producing two score matrices $S^c, S^d \in \mathcal{R}^{R \times C}$ respectively, where C and R denote the number of categories and proposals during training in image I, respectively. Different to work in (Bilen and Vedaldi 2016), we adapt text embedding $T \in \mathcal{R}^{D \times C}$ of category names as the parameters W^c of classification stream so that it imposes explicit visual-semantic constraint during MIL optimization. At the same time, the detection stream is still learnable, since it focuses on localizing the foreground proposals, which is expected to be category-agnostic. Thus, the two score matrices are computed as: $S^c = \frac{X^{\text{fuse}}}{\|X^{\text{fuse}}\|}\frac{T}{\|T\|}$ and $S^d = X^{\text{fuse}}W^d$. Then, both score matrices are normalized by softmax functions $\sigma(\cdot)$ over category and proposal axes, respectively. The final score S for assigning category c to region r is computed via an element-wise product: $S = \sigma(S^c) \odot \sigma((S^d)^T)^T \in [0, 1]$. To acquire image-level classification scores for training, S is summed for all regions $\varphi_c = \sum_{r=1}^{R} S_{rc}$. Then the object-mining objective function is binary cross-entropy loss:

$$\mathcal{L}_{\text{OM}} = \sum_{c=1}^{C} y_c \log(\varphi_c) + (1 - y_c)\log(1 - \varphi_c), \quad (3)$$

where $y \in \{0, 1\}^C$ is the category one-hot label indicating image-level existence of category c.

Recently, WSOD works (Tang et al. 2017, 2018) also explicitly assign pseudo labels from the above mining module to learn more discriminative classifiers, which are also called instance refinement modules. Thus, we also develop multiple open-vocabulary classification heads which uses a shared vision-language representation space to refine discovered object. In addition, to reduce miss-localization, for each refinement branch we regress the bounding boxes which need high-quality proposals to provide PGT boxes. Therefore, the PGT boxes mined by the object mining module are used as box prompt input to SAM to refine boxes to supervise the first refinement branch, and the former refinement branch supervises the latter one. Thus, the objective function of this multi-branch refinement is the sum of classification and regression losses over all branches:

$$\mathcal{L}_{\text{IR}} = \sum_{k=1}^{K} \mathcal{L}_{\text{cls}}^k + \mathcal{L}_{\text{reg}}^k, \quad (4)$$

We concatenate the text embedding T with a background zero-vector as the classifier parameters $W^r \in \mathcal{R}^{D\times(C+1)}$ of refinement branch k. The classification loss is defined as:

$$\mathcal{L}_{\text{cls}}^k = \sum_{r=1}^{R}\sum_{c=1}^{C+1} w_c^k \hat{y}_{r,c}^k \log S_{r,c}^k, \quad (5)$$

where w_c^k is the weight factor to smooth the learning process following (Tang et al. 2017), $S^k \in \mathcal{R}^{R\times(C+1)}$ is computed by $\frac{X^{\text{fuse}}}{\|X^{\text{fuse}}\|}\frac{W^r}{\|W^r\|}$ and $\hat{y}_{r,c}^k$ is the PGT labels of proposal r for category c in branch k. And $\mathcal{L}_{\text{reg}}^k$ is the smooth L1 loss (Ren et al. 2015) in branch k.

Experiments

Datasets. We evaluate the proposed WSOVOD framework on Pascal VOC 2007, 2012 (Everingham et al. 2010) and MS COCO (Lin et al. 2014), which are widely used for WSOD. In addition, we also use ILSVRC (Russakovsky et al. 2015) and LVIS (Gupta, Dollár, and Girshick 2019) for open-vocabulary learning, both of which are widely used for FSOVOD.

Evaluation Metrics. Following the common setting of FSOVOD, we also split COCO into 17 novel classes and 48 base classes, and use AP_N and AP_B to evaluate the results of 17 novel classes and 48 base classes, respectively. We also use AP to evaluate the results of $17 + 48$ total classes. To compare the model performance in the WSOD setting, we use two evaluation metrics: CorLoc and mAP. Correct localization (CorLoc) is a commonly-used measurement that quantifies the localization performance by the percentage of images that contain at least one object instance with at least 50% IoU to the ground-truth boxes. Mean average precision (mAP) follows standard Pascal VOC protocol to report the mAP at 50% IoU of the detected boxes with the ground-truth ones. Furthermore, we report standard COCO metrics for WSOD, including AP at different IoU thresholds.

Implementation Details. We use VGG16 (Simonyan and Zisserman 2015), RN18/50-WS-MRRP (Shen et al. 2020b), initialized with the weights pre-trained on ILSVRC as vision backbones. We use synchronized SGD training on Nvidia 3090 with a batch size of 4, a mini-batch involving 1 images per GPU. We use learning rates of $1e^{-3}$ and $1e^{-2}$ for VGG16 and RN18/50-WS-MRRP backbone, respectfully, a momentum of 0.9, a dropout rate of 0.5, a learning rate decay weight of 0.1. We fix the backbone weights for WSOD but set a $1e^{-5}$ learning rate to backbones for OVOD.

Open-Vocabulary and Cross-Dataset Detection

Since we are the first exploration for WSOVOD, we compare the proposed WSOVOD framework with fully-supervised open-vocabulary object detection (FSOVOD). Noted that FSOVOD divides the MS COCO categories into 48 base and 17 novel classes (Bansal et al. 2018), and uses object-level annotations of 48 base classes during training. In addition, in order to expand vocabulary learning, some works (Zareian et al. 2021; Zhou et al. 2022b; Zareian et al. 2021) use weak annotation information including novel classes, such as tags, captions, and etc. The first and second parts of Tab. 1 shows the state-of-the-art FSOVOD results without and with image-level annotation, respectively. The $6th$ row in the second part removes COCO captions in Detic (Zhou et al. 2022b), which results in a dramatic performance drop in novel classes with only 1.3% AP_N. This shows that fully-supervised methods are hard to generalize well to detect novel classes if they lack the supervision information of a large vocabulary. Therefore, it is necessary to study WSOVOD on large-vocabulary datasets with only category annotations. As shown in the third part of Tab. 1, WSOVOD exhibits strong generalization ability despite large differences between train and test distributions. In particular, WSOVOD significantly improves the AP_N performance of novel classes compared to FSOVOD with only object-level supervision. On COCO novel classes, WSOVOD even surpasses FSOVOD methods, which require both image-level and object-level supervision.

We further conduct experiments to train our WSOVOD with multiple datasets jointly in the bottom part of Tab. 1. We observe that: (1) Cross-dataset learning achieves superior or at least comparable results to the corresponding single dataset. For instance, combing VOC07 and VOC12 sig-

Method		Bakcbone	Train Supervision		COCO			VOC07
			Image-level	Object-level	AP_N	AP_B	AP	mAP
ZS-LAB	(Bansal et al. 2018)	Incept.-Res. v2	–	COCO 48 cls.	0.3	*24.9*	–	–
DELO	(Zhu, Wang, and Saligrama 2020)	DarkNet19	–	COCO 48 cls.	3.4	–	*13.0*	–
PL	(Rahman, Khan, and Barnes 2020)	RN50-FPN	–	COCO 48 cls.	4.1	*35.9*	*7.4*	–
SAN	(Rahman, Khan, and Porikli 2020)	RN50	–	COCO 48 cls.	2.6	*13.9*	*4.3*	–
BLC	(Zheng et al. 2020)	RN50	–	COCO 48 cls.	4.5	*42.1*	*8.2*	–
SSB	(Khandelwal et al. 2023)	RN101	–	COCO 48 cls.	10.2	*48.9*	*16.9*	–
RRFS	(Huang et al. 2022)	RN101	–	COCO 48 cls.	**13.4**	*42.3*	*20.4*	–
OVR-CNN	(Zareian et al. 2021)	RN50-C4	COCO Cap.	COCO 48 cls.	22.8	*46.0*	*39.9*	52.9
ViLD	(Gu et al. 2022)	RN50-FPN	CLIP400M	COCO 48 cls.	27.6	*59.5*	*51.3*	–
ZSD-YOLO	(Xie and Zheng 2022)	CSP-DarkNet53	CLIP400M	COCO 48 cls.	13.6	*31.7*	*19.0*	–
HierKD	(Ma et al. 2022b)	RN50-FPN	Conceptual Cap.	COCO 48 cls.	20.3	*51.3*	*43.2*	–
Detic	(Zhou et al. 2022b)	RN50-C4	COCO Cap.	COCO 48 cls.	27.8	*47.1*	*45.0*	–
Detic	(Zhou et al. 2022b)	RN50-C4	–	COCO 48 cls.	1.3	–	*39.3*	–
RKDWTF	(Bangalath et al. 2022)	RN50-C4	COCO Cap.	COCO 48 cls.	36.6	*54.0*	*49.4*	–
SGDN	(Shi, Hayat, and Cai 2023)	RN50	Flickr30K, VG	COCO 48 cls.	**37.5**	*61.0*	*54.9*	–
PBBL	(Gao et al. 2022)	RN50	COCO Cap., VG, SBU Cap.	COCO 48 cls.	30.8	*46.1*	*42.1*	59.2
WSOVOD		RN50-WS-MRRP	VOC07	–	15.4	*7.8*	*9.8*	63.4
WSOVOD		RN50-WS-MRRP	VOC12	–	17.0	*9.3*	*11.3*	**64.8**
WSOVOD		RN50-WS-MRRP	ILSVRC	–	9.1	*6.4*	*7.0*	26.7
WSOVOD		RN50-WS-MRRP	LVIS	–	16.7	*11.0*	*13.2*	31.0
WSOVOD		RN50-WS-MRRP	COCO	–	**35.0**	**27.9**	**29.8**	60.9
WSOVOD		RN50-WS-MRRP	VOC07, VOC12	–	19.2	*12.4*	*15.1*	**65.0**
WSOVOD		RN50-WS-MRRP	COCO, VOC07, VOC12	–	35.4	*27.3*	*29.8*	**65.0**
WSOVOD		RN50-WS-MRRP	COCO, ILSVRC	–	35.6	*27.7*	*30.0*	61.2
WSOVOD		RN50-WS-MRRP	COCO, LVIS	–	**36.7**	**28.4**	**30.3**	62.3

Table 1: Comparison with the state-of-the-art OVOD methods on MS COCO and Pascal VOC.

nificantly improves the COCO AP_N with gains of 3.8% and 2.2% compared to separately using VOC07 and VOC12 datasets, respectively. (2) Adding more image-level concepts to COCO further improves the COCO AP_N. For instance, adding ILSVRC to COCO performs better than adding VOC07 and VOC12 to COCO. (3) Adding denser image-level annotations significantly improves results. For example, LVIS and COCO share the same training set, and directly combining LVIS and COCO improves 1.7% AP_N, although the image-level labels in LVIS are incomplete.

Rescuing Federated and Long-Tail Data

We further conduct experiments on the difficult federated and long-tail distribution LVIS dataset, as shown in Tab. 3. When only using LVIS for training, the performance of WSOVOD reaches saturation around 1 epoch. This is because LVIS is a federated dataset with sparse annotations where image-level labels are not exhaustively annotated with all classes. The missing classes are treated as background and generate incorrect supervision signals. To this end, we introduce a batch-class-aware sampling, termed BCAS. In BCAS, the data sampler first picks a category and then selects multiple images containing that category to form a mini-batch. When equipped with BCAS for LVIS, WSOVOD reaches saturation at about 3 epochs and improves 3.8% $\text{AP}_{0.5}$ on COCO. We further add COCO to LVIS training without BCAS and observe substantial performance improvements on VOC07 with gains of 30.7% mAP and 34.8% CorLoc, respectively. Compared to single COCO, combing LVIS with COCO also significantly improves the VOC07 mAP from 60.5% to 61.7% and CorLoc from 78.2% to 79.3%, respectively. This reveals that incomplete image-level annotated data is helpful for WSOVOD.

Weakly Supervised Object Detection

We compare our proposed method with the state-of-the-art WSOD methods. Tab. 2 shows the performance comparisons on the VOC 2007, VOC 2012, and MS COCO, where $\mathcal{I}$, $\mathcal{O}$ and $\mathcal{B}$ denote image-level supervision, object-level supervision with class labels, and object-level supervision without class labels, respectively. With the VGG16 backbone, the proposed WSOVOD suppresses the performances of all previous WSOD methods for mAP and CorLoc on VOC and $AP_{0.5:0.95}$ on MS COCO. The proposed WSOVOD with RN18-WS-MRRP backbone reaches the new state-of-the-art of 80.6% and 81.0% CorLoc on VOC 2007 and 2012, respectively, and 29.7% $\text{AP}_{0.5}$ on MS COCO. With RN50-WS-MRRP backbone, WSOVOD further sets new state-of-the-art of 63.4% and 62.1% mAP on VOC 2007 and VOC 2012, respectively, and 20.5% $AP_{0.5:0.95}$ and 21.4% $\text{AP}_{0.75}$ on MS COCO. Furthermore, with object-level supervision without class labels, our proposed WSOVOD even shows comparable performance compared to FSOD in all datasets.

Method		Sup.	Bakcbone	VOC 2007		VOC 2012		MS COCO Avg. Precision, IoU:		
				mAP	CorLoc	mAP	CorLoc	0.5:0.95	0.5	0.75
WSDDN	(Bilen and Vedaldi 2016)	$\mathcal{I}$	VGG16	34.8	53.5	–	–	9.5	19.2	8.2
OICR	(Tang et al. 2017)	$\mathcal{I}$	VGG16	41.2	60.6	37.9	62.1	–	–	–
PCL	(Tang et al. 2018)	$\mathcal{I}$	VGG16	43.5	–	–	–	8.5	19.4	–
WSOD^2	(Zeng et al. 2019)	$\mathcal{I}$	VGG16	53.6	69.5	47.2	71.9	10.8	22.7	–
C-MIDN	(Gao et al. 2019)	$\mathcal{I}$	VGG16	52.6	68.7	50.2	71.2	9.6	21.4	–
MIST	(Ren et al. 2020)	$\mathcal{I}$	VGG16	54.9	68.8	52.1	70.9	12.4	25.8	10.5
UWSOD	(Shen et al. 2020a)	$\mathcal{I}$	RN18-WS-MRRP	45.0	63.8	46.2	65.7	3.1	10.1	1.4
SPE	(Liao et al. 2022)	$\mathcal{I}$	CaiT	51.0	70.4	–	–	7.2	18.2	4.8
Seo et al.	(Seo et al. 2022)	$\mathcal{I}$	RN101	58.7	69.8	56.2	71.2	14.4	29.0	12.4
WSOVOD		$\mathcal{I}$	VGG16	59.1	77.2	59.8	79.7	18.8	27.1	19.7
WSOVOD		$\mathcal{I}$	RN18-WS-MRRP	63.0	**80.6**	61.9	**81.0**	20.1	**29.7**	21.2
WSOVOD		$\mathcal{I}$	RN50-WS-MRRP	**63.4**	80.1	**62.1**	80.7	**20.5**	29.1	**21.4**
Fast RCNN	(Girshick 2015)	$\mathcal{O}$	VGG16	66.9	–	65.7	–	18.9	38.6	–
Faster RCNN	(Ren et al. 2015)	$\mathcal{O}$	VGG16	69.9	–	67.0	–	21.2	41.5	–
WSOVOD		$\mathcal{B}$	VGG16	67.2	88.2	65.4	84.5	16.4	31.1	15.3
WSOVOD		$\mathcal{B}$	RN18-WS-MRRP	68.8	90.9	66.3	89.2	19.8	37.6	18.5
WSOVOD		$\mathcal{B}$	RN50-WS-MRRP	**71.8**	**91.0**	**69.7**	**90.0**	**21.6**	**40.6**	**20.8**

Table 2: Comparison with the state-of-the-art WSOD methods on PASCAL VOC 2007, 2012 and MS COCO.

Train	Test: VOC07		Test: MS COCO Avg. Precision, IoU:		
	mAP	CorLoc	0.5:0.95	0.5	0.75
LVIS	31.0	44.5	4.8	12.9	5.9
LVIS*	31.7	47.7	6.6	16.7	7.8
COCO	60.5	78.2	20.1	29.7	21.2
LVIS, COCO	**61.7**	**79.3**	**21.0**	**30.1**	**22.2**

Table 3: Comparison with the results of WSOVOD trained on LVIS with RN18. ("*" refers to sampling by BCAS.)

Train Data	without DAFE mAP	without DAFE CorLoc	with DAFE mAP	with DAFE CorLoc
VOC07	62.6	78.7	63.0 (↑ **0.4**)	80.6 (↑ **1.9**)
VOC07, VOC12	63.5	79.2	64.1 (↑ **0.6**)	82.2 (↑ **3.0**)
VOC07, COCO	61.4	78.2	63.0 (↑ **1.6**)	80.5 (↑ **2.3**)

Table 4: Ablation study of DAFE with RN18-WS-MRRP backbone on VOC 2007.

Ablation Study

We conducted two sets of ablation studies to investigate the effectiveness of the proposed modules. We firstly ablate DAFE in Tab. 4 to verify the effectiveness of DAFE for training multiple datasets. We test all models on VOC07 *test*. When training on VOC12 and VOC07, DAFE improves mAP by 0.6% and CorLoc by 3.0%. Thus, DAFE significantly improves the detection and localization performance, indicating that DAFE is simple and effective. When training on COCO and VOC07, DAFE improves mAP by 1.6% and CorLoc by 2.3%. It demonstrates that DAFE also deals well with the large distribution gap. DAFE also performs well on a single dataset. Thus, introducing global image-level context to local proposal-level features is helpful to WSOD. This reveals that DAFE not only successfully gathers dataset-level bias but also image-level context. Secondly, we ablate the proposal generator in Tab. 5. It shows that, as observed in (Shen et al. 2020a), only using predictions from the model itself as supervision results in noisy training. When using proposals from MCG, the performance is significantly improved, but compared with SAM proposals based on high-level semantic information, it is still much worse. When adding SAM proposals to LO-WSRPN proposals with the refinement mechanism, our method improves mAP and CorLoc by 16.3% and 15.9%, respectively.

Proposal	mAP	CorLoc
LO-WSRPN	46.7	65.1
MCG (Pont-Tuset et al. 2016)	54.2 (↑ **7.50**)	71.9 (↑ **6.80**)
SAM (Kirillov et al. 2023)	61.7 (↑ **15.0**)	77.5 (↑ **12.4**)
LO-WSRPN + SAM	62.5 (↑ **15.8**)	79.9 (↑ **14.8**)
LO-WSRPN + SAM + refine	63.0 (↑ **16.3**)	81.0 (↑ **15.9**)

Table 5: Ablation study of proposal generator with RN18-WS-MRRP backbone on VOC 2007.

Conclusion

In this paper, we propose a weakly supervised open-vocabulary object detection framework, namely WSOVOD, which extends WSOD to detect and localize open-vocabulary concepts and utilizes diverse and large-scale datasets with only image-level annotation.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U21B2037, No.U22B2051, No.62176222, No.62176223, No. 2176226, No.62072386, No.62072387, No.62072389, No.62002305, NO. 62102151 and No.62272401), the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), and CCF-Tencent Rhino-Bird Open Research Fund.

References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring Visual Prompts for Adapting Large-Scale Models. *arXiv*.

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. In *NeurIPS*.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-Shot Object Detection. In *ECCV*.

Bilen, H.; and Vedaldi, A. 2016. Weakly Supervised Deep Detection Networks. In *CVPR*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-To-End Object Detection with Transformers. In *ECCV*.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *AI*.

Dollár, P.; and Zitnick, C. L. 2013. Structured Forests for Fast Edge Detection. In *ICCV*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (voc) Challenge. *IJCV*.

Felzenszwalb, P. F.; and Huttenlocher, D. P. 2004. Efficient Graph-Based Image Segmentation. *IJCV*.

Feng, C.; Zhong, Y.; Jie, Z.; Chu, X.; Ren, H.; Wei, X.; Xie, W.; and Ma, L. 2022. PromptDet: Towards Open-Vocabulary Detection Using Uncurated Images. In *ECCV*.

Gao, M.; Xing, C.; Niebles, J. C.; Li, J.; Xu, R.; Liu, W.; and Xiong, C. 2022. Open Vocabulary Object Detection with Pseudo Bounding-Box Labels. In *ECCV*.

Gao, Y.; Liu, B.; Guo, N.; Ye, X.; Wan, F.; You, H.; and Fan, D. 2019. C-Midn: Coupled Multiple Instance Detection Network with Segmentation Guidance for Weakly Supervised Object Detection. In *ICCV*.

Girshick, R. 2015. Fast R-Cnn. In *CVPR*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2022. Open-Vocabulary Detection via Vision and Language Knowledge Distillation. *ICLR*.

Gupta, A.; Dollár, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*.

Huang, P.; Han, J.; Cheng, D.; and Zhang, D. 2022. Robust region feature synthesizer for zero-shot object detection. In *CVPR*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. In *ECCV*.

Jiang, S.; Zhu, Y.; Liu, C.; Song, X.; Li, X.; and Min, W. 2022. Dataset Bias in Few-shot Image Recognition. *TPAMI*.

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 *ICCV*.

Khandelwal, S.; Nambirajan, A.; Siddiquie, B.; Eledath, J.; and Sigal, L. 2023. Frustratingly Simple but Effective Zero-shot Detection and Segmentation: Analysis and a Strong Baseline. *arXiv*.

Kim, D.; Lee, G.; Jeong, J.; and Kwak, N. 2020. Tell Me What They're Holding: Weakly-Supervised Object Detection with Transferable Knowledge from Human-Object Interaction. In *AAAI*.

Kim, E.; Lee, J.; and Choo, J. 2021. BiaSwap: Removing dataset bias with bias-tailored swapping augmentation. In *ICCV*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*.

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 *CVPR*.

Liao, M.; Wan, F.; Yao, Y.; Han, Z.; Zou, J.; Wang, Y.; Feng, B.; Yuan, P.; and Ye, Q. 2022. End-to-End Weakly Supervised Object Detection with Sparse Proposal Evolution. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft Coco: Common Objects in Context. In *ECCV*.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022a. Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation. In *CVPR*.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022b. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *CVPR*.

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 *ECCV*.

Pont-Tuset, J.; Arbelaez, P.; Barron, J. T.; Marques, F.; and Malik, J. 2016. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *TPAMI*.

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 *ICML*.

Rahman, S.; Khan, S.; and Barnes, N. 2020. Improved Visual-Semantic Alignment for Zero-Shot Object Detection. In *AAAI*.

Rahman, S.; Khan, S. H.; and Porikli, F. 2020. Zero-shot object detection: Joint recognition and localization of novel concepts. *IJCV*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: Better, Faster, Stronger. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.

Ren, Z.; Yu, Z.; Yang, X.; Liu, M.-Y.; Lee, Y. J.; Schwing, A. G.; and Kautz, J. 2020. Instance-Aware, Context-Focused, and Memory-Efficient Weakly Supervised Object Detection. In *CVPR*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*.

Seo, J.; Bae, W.; Sutherland, D. J.; Noh, J.; and Kim, D. 2022. Object discovery via contrastive learning for weakly supervised object detection. In *ECCV*.

Shen, Y.; Ji, R.; Chen, Z.; Wu, Y.; and Huang, F. 2020a. UW-SOD: Toward Fully-Supervised-Level Capacity Weakly Supervised Object Detection. *NeurIPS*.

Shen, Y.; Ji, R.; Wang, Y.; Chen, Z.; Zheng, F.; Huang, F.; and Wu, Y. 2020b. Enabling Deep Residual Networks for Weakly Supervised Object Detection. In *ECCV*.

Shi, H.; Hayat, M.; and Cai, J. 2023. Open-Vocabulary Object Detection via Scene Graph Discovery. In *ACMMM*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Sun, Z.; Cao, S.; Yang, Y.; and Kitani, K. M. 2021. Rethinking Transformer-Based Set Prediction for Object Detection. In *ICCV*.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. 2018. Pcl: Proposal Cluster Learning for Weakly Supervised Object Detection. *TPAMI*.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *CVPR*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *ICCV*.

Torralba, A.; and Efros, A. A. 2011. Unbiased Look at Dataset Bias. In *CVPR*.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective Search for Object Recognition. *IJCV*.

Xie, J.; and Zheng, S. 2022. Zero-shot Object Detection Through Vision-Language Embedding Alignment. In *ICDMW*.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-Vocabulary DETR with Conditional Matching. In *ECCV*.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-Vocabulary Object Detection Using Captions. In *CVPR*.

Zeng, Z.; Liu, B.; Fu, J.; Chao, H.; and Zhang, L. 2019. Wsod2: Learning Bottom-up and Top-down Objectness Distillation for Weakly-Supervised Object Detection. In *ICCV*.

Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; and Dong, H. 2021. End-To-End Object Detection with Adaptive Clustering Transformer. In *BMVC*.

Zheng, Y.; Huang, R.; Han, C.; Huang, X.; and Cui, L. 2020. Background learnable cascade for zero-shot object detection. In *ACCV*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022b. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*.

Zhu, P.; Wang, H.; and Saligrama, V. 2020. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable Detr: Deformable Transformers for End-To-End Object Detection. In *ICLR*.