Open-Vocabulary Object Detection With an Open Corpus

Jiong Wang[1] Huiming Zhang[2] Haiwen Hong[2] Xuan Jin[2]
Yuan He[2] Hui Xue[2] Zhou Zhao[1] *
[1] Zhejiang University, China [2] Alibaba Group
[1]{liubinggunzu, zhaozhou}@zju.edu.cn
[2]{zhm220845, honghaiwen.hhw, jinxuan.jx, heyuan.hy, hui.xueh}@alibaba-inc.com

Abstract

Existing open vocabulary object detection (OVD) works expand the object detector toward open categories by replacing the classifier with the category text embeddings and optimizing the region-text alignment on data of the base categories. However, both the class-agnostic proposal generator and the classifier are biased to the seen classes as demonstrated by the gaps of objectness and accuracy assessment between base and novel classes. In this paper, an open corpus, composed of a set of external object concepts and clustered to several centroids, is introduced to improve the generalization ability in the detector. We propose the generalized objectness assessment (GOAT) in the proposal generator based on the visual-text alignment, where the similarities of visual feature to the cluster centroids are summarized as the objectness. This simple heuristic evaluates objectness with concepts in open corpus and is thus generalized to open categories. We further propose category expanding (CE) with open corpus in two training tasks, which enables the detector to perceive more categories in the feature space and get more reasonable optimization direction. For the classification task, we introduce an open corpus classifier by reconstructing original classifier with similar words in text space. For the image-caption alignment task, the open corpus centroids are incorporated to enlarge the negative samples in the contrastive loss. Extensive experiments demonstrate the effectiveness of GOAT and CE, which greatly improve the performance on novel classes and get new state-of-the-art on the OVD benchmarks.

1. Introduction

As a fundamental task in computer vision community, object detection is previously bounded in a close-set vocabulary to localize and categorize objects in images. Extending to new categories requires exhaustive human annotated data and specific retraining strategies. Inspired by success of the large-scale visual-language (VL) models [28, 18] applied to open vocabulary recognition, recent open vocabulary object detection (OVD) works transfer the multi-modal capabilities of pre-trained VL models to object detection. The semantic categories are represented as text embeddings and the alignment of region-text feature space enables to detect novel categories described by text inputs.

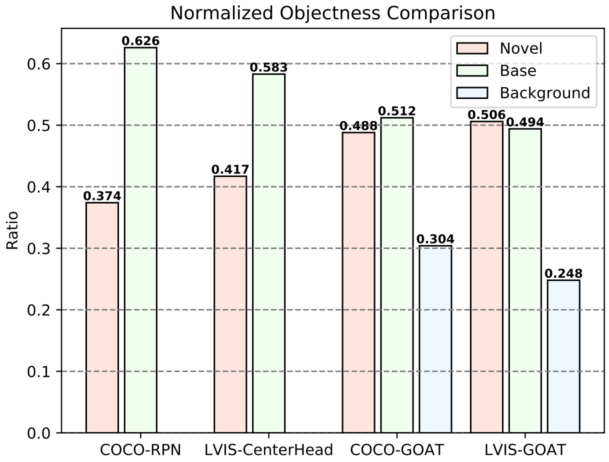


Figure 1: Illustration of the objectness comparison between the base and novel objects assessed by the proposal generators and the proposed GOAT. Conventional proposal generators prefer base categories and the proposed GOAT is generalized for open categories. The objectness scores after sigmoid function are summarized for all base and novel categories and normalized for comparison.

OVD expands traditional object detection to open categories and frees the requirement of costly human annotations, which has recently attracted growing interest. In the seminal work, OVR-CNN [37] defines the notion of OVD and adapts a trained VL model on a Faster-RCNN architecture to align visual features with the category embeddings

[*] corresponding author

for region classification. RegionCLIP [41] and GLIP [19] adapt the CLIP model to region recognition by fine-tuning with grounding data or pseudo-labeled detection data. Zhou *et al*. [42] and Lin *et al*. [21] combine object detection with external recognition or caption datasets to improve the generalization ability. The class-agnostic proposal generators (*e.g*., RPN [29] or Center head [43]) in these works are expected to be generalized to open categories.

However, we found these proposal generators are usually trained on the seen classes and biased toward them. As shown in Figure 1, the proposal generator trained on COCO [22] or LVIS [14] dataset usually gives higher objectness estimations on the seen classes. It indeed undermines the generalization ability on novel categories. In this paper, we propose a generalized objectness assessment (GOAT) strategy by referencing visual features to the open object corpus with a set of object concepts collected from the recognition [7] and caption datasets [5, 30]. The anchor features are projected to the same feature space with text embeddings and the similarities of anchor feature to the centroids of concept clusters are summarized as the generalized objectness. This simple heuristic evaluates objectness with concepts in an open corpus and is theoretically generalized for open categories. As demonstrated in Figure 1, GOAT is generalized for both novel and base objects and suppresses the background for two kinds of proposal generators and datasets.

Based on the open corpus, we also propose category expanding (CE) to enlarge the positive and negative category sets in two aspects. For the region classification, we additionally introduce an open corpus classifier (OCC) by reconstructing each category embedding in original classifier with similar words in the text feature space. OCC encourages the feature to fit open categories similar to the seen classes and avoids to over-fit the seen classes. It essentially enlarges the positive sets for all the categories with the surrounding clusters. For the image-caption matching, the open corpus clusters are used to enrich the negative samples, which is proved effective to improve the performance in contrastive loss. Category expanding enables the model to perceive more positive and negative categories in the feature space and get more reasonable optimization directions.

In summary, we introduce an open object corpus to improve the generalization ability of OV detector. The contributions can be summarized as follows:

- Based on the open corpus, we propose a GOAT module in the proposal generator. GOAT assesses visual objectness with open object concepts and is thus generalized to open categories.
- Based on the open corpus, we propose the category expanding to enrich the positive and negative samples in two stags. In the classification stage, an open corpus classifier is reconstructed to avoid over-fitting the seen classes. In the alignment stage, the open clusters are incorporated to enlarge the negative samples in contrastive loss.
- We give detailed illustrations and qualitative results to dissect the proposed method, which achieves new state-of-the-art performance on the OVD benchmarks.

2. Related Works

Open vocabulary object detection. OVD expands conventional object detection to open categories and frees the requirement of costly human annotations, which has been extensively studied recently. OVR-CNN [37] defined the notion of OVD and pre-train a vision-language model with image-caption pairs. The trained VL model is adopted as the backbone of a Faster-RCNN architecture to align visual features with the category embedding for region classification.

Following works resort to the powerful multi-modal representation ability of CLIP model [28]. ViLD [13] proposes a knowledge distillation strategy to distill knowledge from the visual encoder of CLIP and guide the visual-text alignment in a student detector [13]. RegionCLIP [41] adapts the CLIP to align the fine-grained region-text pairs with pseudo region labels, thus closing the image-region gap. To the same end, GLIP [19] aligns fine-grained region-text pairs in a self-training fashion to learn from both detection and grounding data. Detic [42] trains the classifiers of a detector on image classification data and thus expands the vocabulary of detectors to tens of thousands of concepts. VLDet [21] translates the region-word alignment in image-text pairs to a bipartite matching problem between image regions and word candidates, and optimizes the fine-grained region-word alignment to complement the OVD. The pseudo region labels strategies [11, 39] enlarge the set of base classes by automatically generating pseudo annotations of diverse objects, with the help of the pre-trained class-agnostic detector and VL model. Prompt learning strategies [10, 8] design learnable text prompts to preferably facilitate the alignment of regional and textual feature space. The transformer architectures [36, 27] have also been discovered for OVD and present impressive results.

Compared to these works, the proposed method is advanced in two aspects. Firstly, the detector of these works is optimized in the base classes and the proposal generator may be biased to the base classes. The proposed GOAT method assesses the objectness with an open corpus that is generalized to open categories. Even though the pseudo-label strategies enlarge the base classes, the proposal generator is still biased to the seen data and sensitive to the quality of annotations. Secondly, the classifiers of these works are optimized to distinguish seen objects. The proposed category expanding method enables perceiving more samples

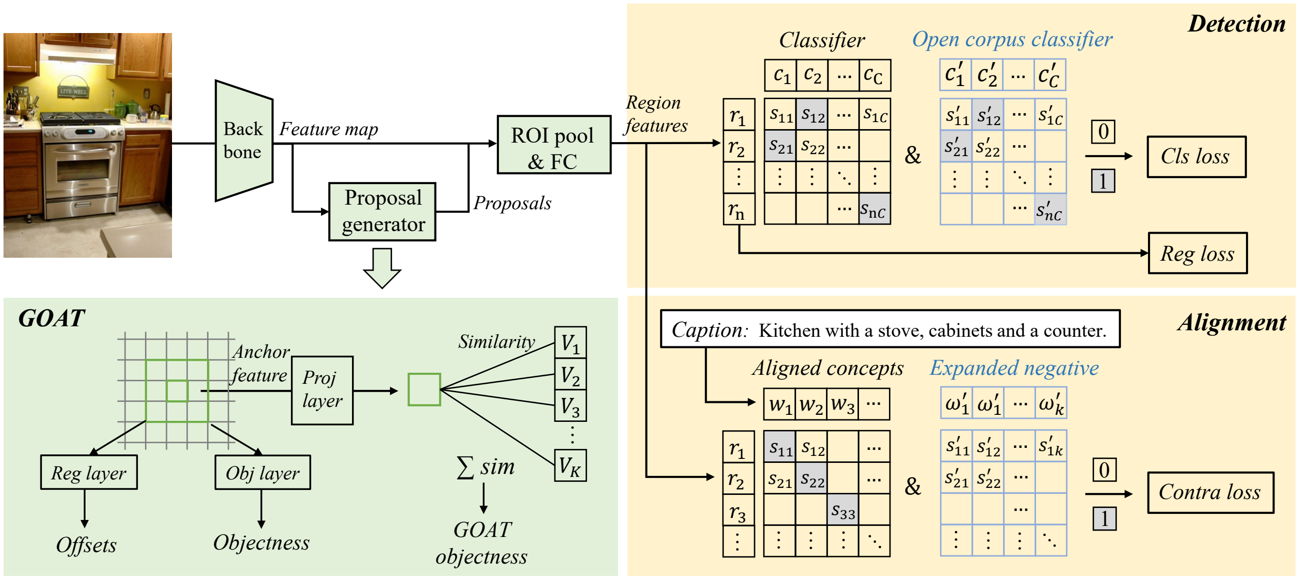


Figure 2: Illustration of the open vocabulary detection framework with the proposed generalized objectness assessment and category expanding based on an open corpus. The GOAT module is plugged in the proposal generator to complement the objectness layer. With the thought of category expanding, we reconstruct an open corpus classifier to complement the original classifier in detection task and expand negative clusters from open corpus in the region-word alignment task.

in the feature space and get reasonable optimization directions. Even though Detic and VLDet enlarge the concept sizes with external datasets, the proposed method with open corpus is complementary to them and greatly improves the performance on novel classes.

Deep models with external knowledge. There have been previous works that improves data-driven deep models with external knowledge bases in natural language processing [16, 17, 3] and computer vision communities [6, 33]. In image captioning works [12, 34, 4, 20], the knowledge bases are incorporated to expand vocabulary and describe novel objects. The knowledge bases [1, 32, 23] have also been adopted in the visual reasoning tasks [31, 25, 38, 40, 26, 35] with a set of supporting facts. The knowledge distillation strategy in ViLD [13] adapts the knowledge of CLIP visual encoder of to the student detector. Similarly, the pseudo-labeling works [11, 39] adopt the pretrained VL models to enlarge the seen classes, which can also be deemed as knowledge transfer. The external recognition and caption datasets adopted in Detic [42] and VLDet *et al.* [21] are another form of external knowledge.

Different from these works, only the open corpus concepts are adopted in this paper, without paired captions [21], images [42] or (pseudo) annotations [11, 39]. With this easily available and simple form of knowledge, the generalization ability of the detector can be improved with the proposed GOAT and category expanding strategies.

3. The Proposed Method

In this section, we first introduce the conventional and open-vocabulary object detection setting (Section 3.1). Then we detail the generalized objectness assessment (GOAT) module (Section 3.2) and category expanding (Section 3.3) based on the open corpus. The category expanding consists of the open corpus classifier in detection task and the negative cluster expanding in region-word alignment task.

3.1. Preliminary

Conventional object detectors are trained on the object detection dataset $\mathcal{D}^{\text{det}} = \{(\mathcal{I}, \{(b, l)_k\})_i\}$ with region annotation of box b and category label l. The training and testing steps follow the same localization and classification objectives on the whole category vocabulary $\mathcal{V}$. In open vocabulary object detection (OVD), $\mathcal{V}$ is divided into base vocabulary $\mathcal{V}^{base}$ and novel vocabulary $\mathcal{V}^{novel}$. The OV detectors are usually trained with annotations of base categories and test on the whole vocabulary. The conventional practice replaces the region classifier with frozen text embedding extracted from pre-trained VL models and optimizes the region-text alignment for categorizing open categories.

Considering a two-stage detector with a proposal generator (*e.g.*, RPN or Centerhead) to generate proposal candidates and a ROI-head for region classification and box regression. An image $\mathcal{I}$ is passed to the backbone, the feature map $F \in \mathbb{R}^{h\times w\times d}$ is got with height h, width w and dimension d. The multi-scale anchors are generated at each spatial position in the proposal generator with corresponding objectness O_{obj} predicted in the objectness layer and offsets predicted in the regression layer.

The proposals with higher objectness are then passed to the ROI-head with ROI alignment and transformation to get region features $\mathcal{R} = \{r_1, r_2, ..., r_n\}$, which are used for region classification and box regression as shown in Figure 2. The binary cross-entropy loss is adopted as follow:

$$\begin{aligned} Loss_{cls} &= \sum_{i}^{n} -(\log \boldsymbol{\sigma}(s_{i,c}) + \sum_{k\neq c} \log \boldsymbol{\sigma}(1 - s_{i,k})), \\ s_{i,k} &= \mathcal{C}_k r_i^T, \end{aligned} \tag{1}$$

where $\mathcal{C} \in \mathbb{R}^{C\times d}$ is the classifier, $\boldsymbol{\sigma}$ is the sigmoid function and c is the target labels.

In this paper, we introduce an open corpus with large amount of external object concepts and apply it in the proposal generator with GOAT and in the ROI-Head with category expanding.

3.2. Generalized Objectness Assessment With Open Corpus

In the proposal generator, the objectness is evaluated by the objectness layer with convolution operation. Even though it is class-agnostic, the parameters are optimized to highlight the regions of base category and suppress others. It inevitably over-fits the seen classes as demonstrated in Figure 1. We thus propose to assess the objectness by referencing to an open corpus, which is constructed by collecting the object concepts in the object recognition [7] and caption datasets [5, 30].

Having an open corpus $\mathcal{T} \in \mathbb{R}^{N\times d}$ with N object concepts, we extract their text embeddings with the CLIP model [28] and align the visual-text feature space with a projection layer on the anchor features. It is intuitive that visual objects are close to these concept embeddings while the background is on the opposite. Exhaustively referencing all the concepts is time-consuming and we propose to cluster the corpus into k_{obj} cluster centroids $V_{obj} \in \mathbb{R}^{k_{obj}\times d}$ as a substitution. As illustrated in the GOAT part of Figure 2, the anchor features with different spatial scales are generated for objectness assessment in the objectness layer and offset estimation in the offset layer. In the projection layer, each anchor feature f is aligned with the centroids and their similarities are then averaged as the GOAT objectness as follow:

$$O_{goat} = \frac{1}{k_{obj}} \sum_{i=1}^{k_{obj}} s_i, \quad s = V_{obj} f^T, \tag{2}$$

For brevity, we uniformly adopt s to indicate the similarity, with different indications in different equations. The GOAT score is used to complement the original objectness score in proposal generator, and the final objectness is formulated as $O'_{obj} = O_{goat} + O_{obj}$.

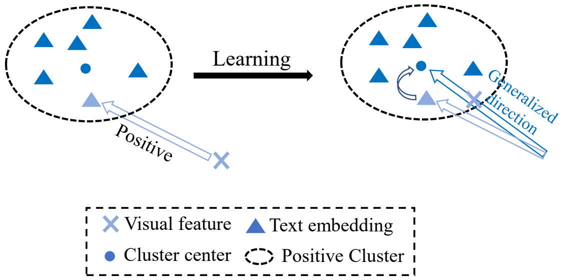


Figure 3: Illustration of the open corpus classifier. The open corpus classifier reconstructs the original category embedding with a set of similar text embeddings. Fitting such an open classifier gets more general optimizing directions to the open classes beyond seen classes.

3.3. Category Expanding With Open Corpus

Open corpus classifier (OCC). The classification layer in the detector encourages a region feature to be close to the positive category embedding and far from other known categories. A distinct performance gap between seen and unseen classes can be observed in OV classifier and the reason is known that the number of seen categories is limited. We thus propose an open corpus classifier (OCC) C' which reconstructs the original classifier with similar objects in the open corpus. Specifically, we retrieve a category embedding within the corpus and choose the top-k most similar concepts with similarities larger than a threshold (0.9). The similarities are transformed to reconstruction weights with a softmax function and the OCC classifier is got by weighted summation of similar concept embeddings:

$$C'_i = softmax(S_i \alpha)\mathcal{T}, \quad S = mask(sort(\mathcal{T} C_i^T)) \tag{3}$$

where α is the temperature, $S_i \in \mathbb{R}^{1\times N}$ indicate the concepts weights after sorting and masking operations, where only top-k similar concepts are kept. The open classifier gets closer to the original classifier when α gets larger and we empirically set it to 30.

A schematic diagram of OCC is illustrated in Figure 3, where the original category embedding is replaced by the center of cluster surrounding it. Optimizing with OCC gets generalized directions that close to all the open categories in the cluster rather than a seen category. Essentially, OCC can be considered as category expanding where the positive and negative sets are expanded beyond the seen classes.

As illustrated in Figure 2, the OCC is adopted to complement the original classifier. The total classification loss is the normalized summation of two classifiers as follow:

$$\begin{aligned} Loss_{cls}^{all} &= \frac{1}{1+\lambda_{cls}}(Loss_{cls} + \lambda_{cls} * Loss_{cls}^{occ}), \quad (4) \\ Loss_{cls}^{occ} &= \sum_{i=1}^{n} -(\log \boldsymbol{\sigma}(s'_{i,c}) + \sum_{k \neq c} \log \boldsymbol{\sigma}(1 - s'_{i,k})). \end{aligned}$$

The additional classifier is discarded and only the original classifier is retained when testing.

Negative cluster expanding (NCE). Following VLDet [21] and Detic [42], we also incorporate a caption dataset and optimize fine-rained region-word matching with bipartite matching to improve the generalization ability. After the linear assignment step [21], a set of region feature $\mathcal{R} = \{r_1, r_2, ..., r_n\}$ are aligned with corresponding concept embeddings $\mathcal{W} = \{w_1, w_2, ..., w_n\}$, and a set of negative words in same batch $\mathcal{W}' = \{w'_1, w'_2, ..., w'_m\}$. The region-word similarity matrix of positive and negative sets are calculated within the contrastive loss as follow:

$$Loss_{contra} = \sum_{i=1}^{n} -(\log \boldsymbol{\sigma}(s_{i,i}) + \sum_{j \in \mathcal{W}'} \log \boldsymbol{\sigma}(1 - s_{i,j})). \quad (5)$$

Contrastive loss [28, 18] is known to benefit from larger batch size with more negative samples. We thus propose to mine negative clusters in the open corpus to facilitate the contrastive loss, where each cluster centroid summarizes words with similar patterns. Specifically, we cluster the open corpus into k_{cap} clusters $V_{cap} \in \mathbb{R}^{k_{cap} \times d'}$ and expand the negative set $\mathcal{W}'$ with V_{cap}. The expanded negative loss is formulated as follow:

$$Loss_{contra}^{nce} = \sum_{i=1}^{n} -(\sum_{j \in V_{cap}} m_{i,j} \log \boldsymbol{\sigma}(1 - s_{i,j})). \quad (6)$$

where m is the mask and $m_{i,j} = 0$ to indicate i and j belong to the same cluster and we neglect this negative loss. The final contrastive loss with expanded negative clusters is formulated as follow:

$$Loss_{contra}^{all} = Loss_{contra} + \lambda_{cap} * Loss_{contra}^{nce}. \quad (7)$$

In the training stage, we alternatively optimize the detection with OCC (Equation 4) and region-word alignment tasks (Equation 7) with NCE following VLDet [21].

4. Experiments

4.1. Datasets

COCO-2017. The COCO-2017 dataset is manually divided into 48 base classes and 17 novel classes in open-vocabulary COCO setting (OV-COCO) [2, 37]. Following previous works, 107,761 training images with base class annotations are used for training. Correspondingly 4,836 test images with both base and novel classes are used for evaluation. The box mAPs for base and novel are reported.

LVIS. The LVIS dataset [14] contains 1203 categories which are further split into frequent, common and rare categories. Following previous works, we combine the frequent and common categories as base classes and keep all rare classes as novel, resulting in 866 base and 337 rare classes. The novel class annotations are removed in training and all categories are evaluated in testing images. The mask mAPs for base and novel are reported.

COCO Caption and Conceptual Captions. The COCO Caption [22] and Conceptual Caption (CC3M) [30] are used as the caption datasets for alignment task in the training stage. Following [21], we pair COCO Caption with OV-COCO and CC3M with OV-LVIS training data. The object concepts from COCO Caption and CC3M dataset are separately filtered with the concepts frequency are larger than 100, resulting in 4,764 and 6,790 concepts.

For the open object corpus, we construct it by summarizing the object concepts in ImageNet21k [7], COCO caption and CC3M datasets. After removing duplicate words, we get the open object corpus with 28,535 object concepts.

4.2. Implementation Details

Following [21, 42], the parameters of the detector is initialized by a fully-supervised detector trained on the detection data of base category. The pre-trained CLIP text encoder (RN50) is adopted to embed the caption and object words. For the OV-COCO setting, we adopt the ImageNet pre-trained model or the RegionCLIP model (RN50) as backbone of a Faster-RCNN architecture. The learning rate is set to 0.02 for the detector and 0.002 for the backbone. The model is optimized for 90,000 iterations using SGD optimizer with 1000 iterations warm up, and the learning rate is scaled down by a factor of 10 at 60,000 and 80,000 iterations. For OV-LVIS setting, we follow [21, 42] to adopt Center-Net2 [43] with ResNet50 [15] backbone as the detector. The learning rate is set to 2e-4 and the model is optimized for 90,000 iterations using Adam optimizer with 1000 iterations warm up. The large-scale jittering and repeat factor sampling [14] are adopted as data augmentation. The alternative training scheme of VLDet is adopted and other training details follow previous works [21, 42, 41].

The cluster numbers in GOAT and NCE are separately set to 128 and 1,024. The λ_{cls} and λ_{cap} in OCC and NCE are set to 0.1 and 0.2 for OV-COCO, 0.2 and 0.3 for OV-LVIS. The ablations are presented in Figure 4.

4.3. Open-Vocabulary Object Detection

OV-COCO. The comparisons with the state-of-the-art works on COCO dataset datasets are presented in Table 1. RegionCLIP [41] adapts the CLIP model for region recognition task and ViLD [13] distills the knowledge from CLIP

Method	Architecture	Extra supervision	Novel AP	Base AP	All AP
Base-only [42]	RN50-IN		1.3	52.8	39.3
OVR-CNN [37]	RN50-IN	Caption	22.8	46.0	39.9
Detic [42]	RN50-IN	Image	27.8	47.1	42.0
RegionCLIP [41]	RN50-CLIP	Caption	26.8	54.8	47.5
ViLD [13]	RN50-IN	CLIP	27.6	**59.5**	**51.3**
PBOVD [11]	RN50-IN	Caption	30.8	46.1	42.1
VLDet [21]	RN50-IN	Caption	32.0	50.6	45.8
VLDet*	RN50-IN	Caption	30.1	51.3	45.7
Ours	RN50-IN	Caption	31.7	51.3	46.1
Ours	RN50-CLIP	Caption	**36.4**	53.0	48.6

Table 1: Compared with existing OVD works on COCO dataset. The novel AP is a primary indicator to reflect the performance. The best results are highlighted in bold. * indicates the re-implementation results.

Method	Backbone	mAP$^{mask}_{novel}$	mAP$^{mask}_{com}$	mAP$^{mask}_{freq}$	mAP$^{mask}_{all}$
Base-only [42]	RN50	16.3	31.0	35.4	30.0
Detic [42]	RN50	19.5	-	-	30.9
RegionCLIP [41]	RN50	17.1	27.4	34.0	28.2
ViLD [13]	RN50	16.6	24.6	30.3	25.5
DetPro [8]	RN50	19.8	25.6	28.9	25.9
VLDet [21]	RN50	21.7	29.8	34.3	30.1
Ours	RN50	**23.3**	29.7	34.3	30.4
Base-only [42]	Swin-B	21.9	40.5	43.3	38.4
Detic [42]	Swin-B	23.9	40.2	42.8	38.4
VLDet [21]	Swin-B	26.3	39.4	41.9	38.1
Ours	Swin-B	**27.4**	40.0	42.2	**38.5**

Table 2: Compared with existing open vocabulary object detection works on LVIS dataset with the ResNet50 and Swin-B backbone.

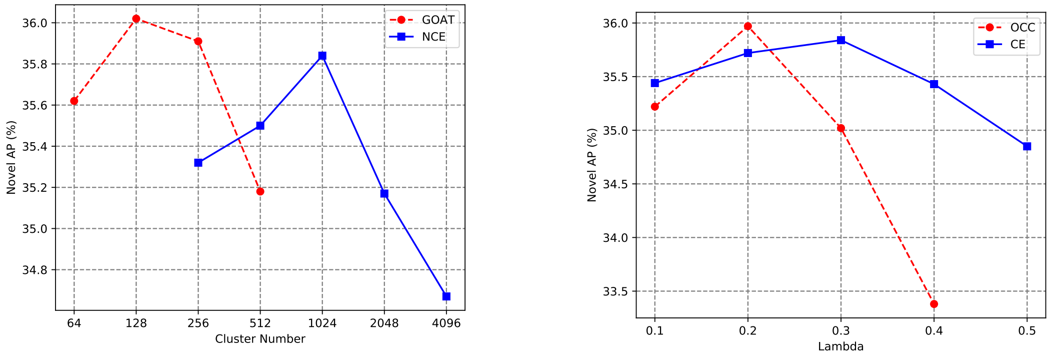
model to the student backbone. The PBOVD [11] discovers pseudo labeling strategy on the caption datasets to enlarge the seen classes in the training stage. The Detic [42] and VLDet [21] adopt external image recognition or caption datasets to facilitate open region recognition. All these methods adopt the ResNet50 architecture as backbone with different extra supervisions.

The proposed method is based on re-implementation of VLDet and the ImageNet pre-trained model (RN50-IN) or the RegionCLIP pre-trained model (RN50-CLIP) is adopted as backbone. Together with the proposed generalized objectness assessment (GOAT) and category expanding (CE), our model greatly improves the novel performance and outperforms all existing works. Note that the re-implementation result of VLDet is lower than the provided results in VLDet paper. Even though the ViLD gets better performance on base categories, Novel AP is a primary indicator to reflect the open vocabulary recognition performance. Compared with the most similar works (Detic and VLDet), the proposed method comprehensively outperforms these works on the three metrics and surpasses the SOTA method VLDet by 4.4% Novel AP.

OV-LVIS. The comparisons with the state-of-the-art works on LVIS dataset are presented in Table 2 with both ResNet50 and Swin-B backbone [24]. The Base-only model which is trained only on the seen classes gets considerable performance thanks to the dense annotations and larger categories. Based on the SOTA method VLDet, the proposed method adopts the GOAT and CE with the help of an open corpus. It can be observed that our model surpasses all existing methods on the novel mask AP by a large margin. Similar observation can be seen in the results based on Swin-B backbone, which reflects the effectiveness and extensive applicability of the proposed method.

4.4. Ablation Study

In this subsection, we give exhaustive ablation studies to demonstrate the effectiveness of generalized objectness assessment (GOAT) and category expanding (CE) consisting of open corpus classifier (OCC) and negative cluster expanding (NCE). The baseline method is based on our reproduction of the alternative training framework proposed



(a) Ablation on cluster number

(b) Ablation on λ

Figure 4: (a) The ablation of the number of open clusters in GOAT and the negative clusters in NCE. (b) The ablation of parameter λ in OCC and NCE. Both ablations are conducted on the OV-COCO protocol and only the novel APs are presented.

GOAT cluster		OV-COCO		
Train	**Test**	Novel AP	Base AP	All AP
None	None	35.0	52.7	48.1
Base	All	35.3	53.1	48.5
Open C	Open C	36.0	**53.1**	**48.6**
Open C	All	**36.1**	53.1	48.6

Table 3: Ablation of GOAT with different variants. Open C indicates the proposed open corpus clusters in GOAT.

by VLDet [21]. The differences are in two aspects: For OV-COCO, we adopt CLIP (RN50) visual encoder rather than the ImageNet pre-trained model as backbone. This practice gets higher performance than what reported in VLDet because of the alignment of CLIP visual and text encoder. For OV-LVIS, we follow the implementation details and the performance is lower than what reported in VLDet. Even so, we surpass VLDet on both COCO and LVIS datasets with the proposed GOAT and CE (Section 4.3)).

Generalized objectness assessment. To demonstrate the effectiveness of the proposed GOAT with open corpus clusters, we compare it with baseline (No cluster) and some variants. The first variant is based on the constrained clusters, where the open clusters are replaced by the base category embeddings. The base embeddings describe the objects in the training data and are replaced by all the category embeddings when testing. We also ablation the usage of all category and open clusters in GOAT when testing. The comparison is given in Table 3, where some observations can be summarized. GOAT surpasses the baseline regardless of the constrained or open clusters are adopted, while the open clusters perform better than the constrained counterpart. When replacing the open clusters with all the category embeddings in the testing stage, there are small improvements. In practice, we consistently adopt the open clusters in our experiment except as otherwise noticed.

Category expanding. To demonstrate the effectiveness of the proposed category expanding (CE) with open corpus classifier (OCC) and negative cluster expanding (NCE), we separately validate them on the OV-COCO and OV-LVIS protocols in Table 4. It can be seen that OCC and NCE consistently improve the baseline method on two datasets. When they are combined together, the category expanding method gets further performance enhancement. Moreover, the category expanding and GOAT are complementary each other to further improve the performance. It can be observed that all the proposed methods consistently improve the baseline performance to reflect their effectiveness.

Cluster number choice. As is known that a cluster centroid is the summation of all the sample embeddings assigned in the cluster. Changes of cluster number influence the average samples assigned in the cluster. The choice of open cluster number in GOAT and NCE are empirically set to be 128 and 1,024. The comparison with other choices is plotted in Figure 4 (a). For the GOATs with cluster number smaller than 512, the performance is similar and 128-cluster performs better. The performance degrades when the cluster number changed to 512, we suppose the reason is some clusters have only one sample and they are weak to evaluate objectness. For NCE in the contrastive loss, properly increasing the negative clusters considerably improves the novel performance and the performance drops in cluster-4096, we suppose the negative gradients overwhelm the positives and a re-weighting strategy is not work. We thus adopt the best-performed cluster number choices by default.

We also give ablations on the choice of λ of OCC and NCE in Figure 4 (b). In OCC, increasing λ weakens the influence of original classifier and strengthen the open corpus classifier. Moderately fitting a noisy classifier increases the generalization ability but it is risky to excessively strengthen the impact. As can be observed, OCC improves the per-

Method	**OV-COCO**			**OV-LVIS**			
	Novel AP	Base AP	All AP	$\text{mAP}^{mask}_{novel}$	mAP^{mask}_{com}	mAP^{mask}_{freq}	mAP^{mask}_{all}
Baseline	35.0	52.7	48.1	21.0	29.1	34.1	29.7
$+GOAT$	36.0	53.1	48.6	21.7	29.8	30.1	
$+OCC_{cls}$	36.0	52.9	48.4	21.3	29.3	34.1	29.9
$+NCE_{caption}$	35.8	53.0	48.5	21.6	29.6	34.1	30.1
$+CE$	35.9	53.1	48.6	22.0	29.7	34.2	30.2
$+GOAT+CE$	**36.4**	53.0	**48.6**	**23.3**	**29.7**	**34.3**	**30.4**

Table 4: Ablation of generalized objectness assessment and category expanding on OV-COCO and OV-LVIS protocols.

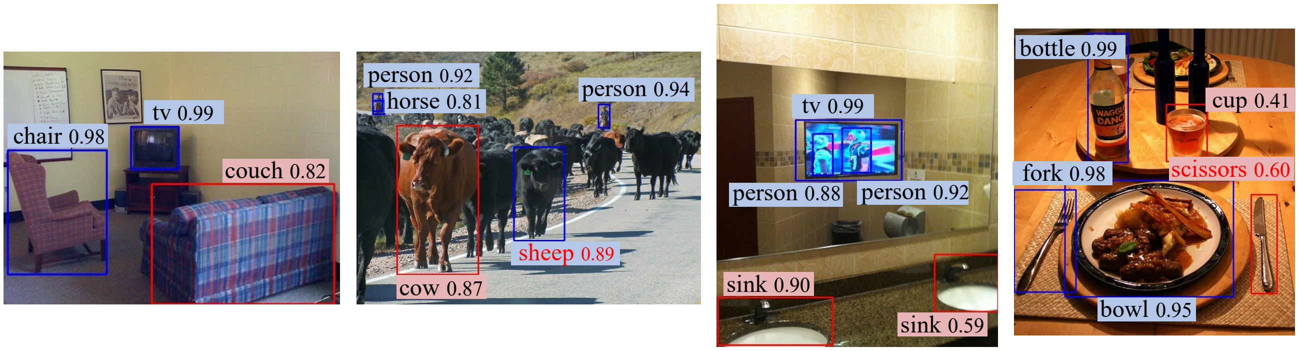


Figure 5: Qualitative results of the detection results on COCO dataset. The top predictions on bounding boxes and categories with confidence scores are plotted. The novel objects are bounded by red boxes and base objects are blue. The wrongly categorized objects are highlighted in red font.

Open classifier	**OV-COCO**		
/ without novel	Novel	Base	All
GOAT	36.0 / 36.1	53.1 / 53.1	48.5 / 48.5
OCC	36.0 / 36.0	52.9 / 52.8	48.4 / 48.3
NCE	35.8 / 35.6	53.0 / 53.0	48.5 / 48.4
GOAT + CE	36.4 / 36.3	53.0 / 53.1	48.6 / 48.6

Table 5: The impact of novel classes in the open corpus.

formance when λ is smaller and decrease the performance sharply when it gets larger than 0.2. Similar observation can be drawn in NCE where the λ control the weight of negative gradients, where best performance is achieved when $\lambda = 0.3$. We thus adopt the best λ choices in OCC and CE by default.

Novel classes in the open corpus. It is intuitive that the pseudo annotations [11, 41] of novel classes facilitate the recognition of corresponding objects when testing. By contrast, the proposed GOAT, OCC and NCE methods are robust to particular classes. In Table 5, we eliminate novel classes in the open corpus and investigate their impacts on GOAT, OCC and NCE. We can observe their impacts are neglectable. In theory, GOAT and NCE are robust to novel classes because each class is a few percent of elements in a cluster. The open classifier is robust to novel classes because they are not close to base classes in feature space. This is our superiority compared to the pseudo-labeling works.

4.5. Qualitative Results

The proposed method based on an open object corpus helps to discover novel objects and we illustrate the detection results of the proposed model on COCO dataset in Figure 5. There are three observations can be summarized. First, the proposed model has considerable ability to localize and categorize the novel objects, such as “couch” and “sink”, with decent precision as Example 1 and 3 illustrated. Second, the confidence scores of novel objects are usually lower than that of base objects. We suppose the visual features are somewhat biased to the seen classes in the training stage. Third, there are also some false categorization results that usually happen when categorizing the novel objects. For example, wrongly categorizing “cow” as “sheep” in Example 2 and categorizing “knife” as “scissors” in Example 4. These evidences reflect that the proposed model is still somewhat biased to the base category, even though it surpasses the state-of-the-art works on the novel classes.

Method	PASCAL VOC	LVIS
OVR-CNN [37]	52.9	5.2
PB-OVD [11]	59.2	8.0
VLDet [21]	61.7	10.0
Ours	**63.6**	**14.0**

Table 6: The comparison of transferring COCO-trained models to the PASCAL VOC and LVIS dataset. The box AP_{50} results are presented.

As aforementioned, the proposed GOAT is generalized to open categories. The normalized score comparison of base and novel objects in two datasets can be seen in Figure 1, where the GOAT objectness scores are generalized for open categories and can suppress the background with lower scores.

4.6. Transfer to Other Datasets

To evaluate the generalization ability, we conduct the transferring experiments where our COCO-trained model is adopted to evaluate on PASCAL VOC [9] test set and LVIS validation set [14] without re-training. The classifier is replaced by the class embeddings of these two datasets for categorization. The adaptation from COCO to PASCAL VOC dataset suffers from the domain gap and the adaptation to LVIS suffers from the augmented semantic space. We compare this setting with former works and give the comparison in Table 6. It can be seen that the best-performed OV-COCO model still performs best in this transfer setting, especially on the difficult LVIS dataset with thousands of categories. It demonstrates the remarkable generalization ability of the proposed method.

5. Conclusion

In this paper, we introduce an open corpus with a set of object concepts to improve the generalization ability in open vocabulary object detection. The open corpus clusters are adopted in the proposal generators to evaluate the visual objectness for generalized objectness assessment (GOAT). Based on the open corpus, We also propose a category expanding strategies that expand the positive and negative samples in two aspects: In the classification stage, we reconstruct the original classifier with similar concepts in the open corpus. In the region-word matching stage, the open corpus clusters are used to enlarge the negative words in the contrastive loss. Extensive experiments demonstrate the effectiveness of incorporating the open corpus with the GOAT and CE strategies, which gets new state-of-the-art results on the open vocabulary benchmarks.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No. 62222211, National Natural Science Foundation of China under Grant No.61836002, National Natural Science Foundation of China under Grant No.62072397.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer, 2007. 3

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 5

[3] Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 967–976, 2014. 3

[4] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 3

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 4

[6] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 48–64. Springer, 2014. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4, 5

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2, 6

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 9

[10] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Prompt-

det: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer, 2022. 2

[11] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, pages 266–282. Springer, 2022. 2, 3, 6, 8, 9

[12] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1969–1978, 2019. 3

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 3, 5, 6

[14] 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. 2, 5, 9

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[17] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric Xing. Deep neural networks with massive learned knowledge. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1670–1679, 2016. 3

[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. 1, 5

[19] 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. 2

[20] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12497–12506, 2019. 3

[21] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2023. 2, 3, 5, 6, 7, 9

[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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5

[23] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 3

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[25] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 3

[26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 3

[27] 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. *arXiv preprint arXiv:2205.06230*, 2022. 2

[28] 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. 1, 2, 4, 5

[29] 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. 2

[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. 2, 4, 5

[31] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4602–4612, 2019. 3

[32] Niket Tandon, Gerard Melo, and Gerhard Weikum. Acquiring comparative commonsense knowledge from the web. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014. 3

[33] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018. 3

[34] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2017. 3

[35] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017. 3

[36] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*. Springer, 2022. 2

[37] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2, 5, 6, 9

[38] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1356–1365, 2021. 3

[39] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 2, 3

[40] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Webly supervised knowledge embedding model for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12445–12454, 2020. 3

[41] 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. 2, 5, 6, 8

[42] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2, 3, 5, 6

[43] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 2, 5