Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

Zongyang Ma[1,2], Guan Luo[1,2], Jin Gao[1,2,†], Liang Li[3,†], Yuxin Chen[1,2], Shaoru Wang[1,2]
Congxuan Zhang[4], and Weiming Hu[1,2,5]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Brain Science Center, Beijing Institute of Basic Medical Sciences [4]Nanchang Hangkong University
[5]CAS Center for Excellence in Brain Science and Intelligence Technology
mazongyang2020@ia.ac.cn, {gluo,jin.gao}@nlpr.ia.ac.cn, liang.li.brain@aliyun.com

Abstract

Open-vocabulary object detection aims to detect novel object categories beyond the training set. The advanced open-vocabulary two-stage detectors employ instance-level visual-to-visual knowledge distillation to align the visual space of the detector with the semantic space of the Pre-trained Visual-Language Model (PVLM). However, in the more efficient one-stage detector, the absence of class-agnostic object proposals hinders the knowledge distillation on unseen objects, leading to severe performance degradation. In this paper, we propose a hierarchical visual-language knowledge distillation method, i.e., HierKD, for open-vocabulary one-stage detection. Specifically, a global-level knowledge distillation is explored to transfer the knowledge of unseen categories from the PVLM to the detector. Moreover, we combine the proposed global-level knowledge distillation and the common instance-level knowledge distillation to learn the knowledge of seen and unseen categories simultaneously. Extensive experiments on MS-COCO show that our method significantly surpasses the previous best one-stage detector with 11.9% and 6.7% AP_{50} gains under the zero-shot detection and generalized zero-shot detection settings, and reduces the AP_{50} performance gap from 14% to 7.3% compared to the best two-stage detector. Code will be released at this url [1].

1. Introduction

The emerging trends in advanced detectors [2, 18, 19, 26–29, 33, 37] have improved the speed and accuracy of traditional object detection tasks significantly, whereas the categories they can recognize are limited. Once the traditional detectors are expected to detect more object categories in the real-world scenarios, the usual solution falls on labeling more categories of objects in training sets. However, the cost may be unaffordable and the long-tail distribution will be exacerbated by increasing the unseen categories linearly according to Zipf's law [31]. To overcome these limitations, zero-shot [1] and open-vocabulary [36] object detection tasks are proposed to recognize objects from unseen categories (novel categories) while the detector is only trained with annotations from seen categories (base categories). The main difference between these two tasks is that the open-vocabulary detector might have seen a novel object during training though its instance-level annotation

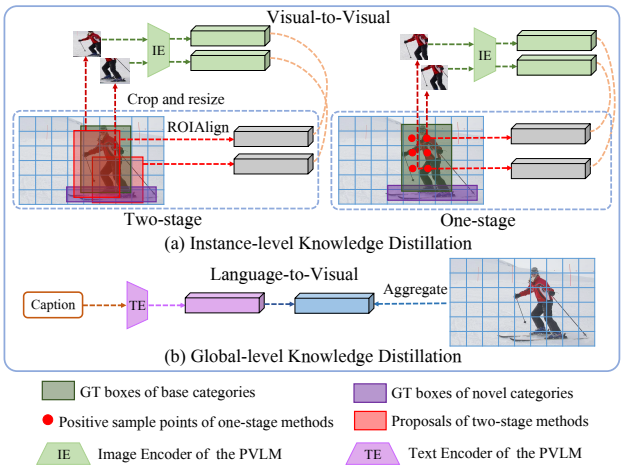

Figure 1. **Comparisons between instance-level and global-level knowledge distillation:** (a) illustrates the pipelines of two-stage methods and one-stage methods with instance-level knowledge distillation. (b) illustrates our proposed global-level knowledge distillation, which directly distills the caption representation from PVLM to the global image representation from detector.

[1] https://github.com/mengqiDyangge/HierKD
† Corresponding authors.

is not available. Therefore, the open-vocabulary detector [8, 35, 36] has developed more rapidly recently, and their performance also lead the former by a large margin.

There have been some works attempting to redesign the traditional detectors to accomplish the above two detection tasks. These works can also be divided into two-stage [1, 8, 17, 36, 39, 40] methods and one-stage [25, 35, 38, 42] methods as in traditional detection. It is known that the traditional state-of-the-art one-stage detectors have comparable performance and more concise pipeline compared to traditional two-stage ones. However, in the open-vocabulary object detection, the current best two-stage method ViLD [8] significantly surpasses the similar one-stage method [35]. As such, it is encouraging to analyze the reason behind this phenomenon and find ways to narrow this performance gap, and then construct a high-performance open-vocabulary one-stage detector.

We show pipelines of recent two-stage and one-stage open-vocabulary detection methods in Figure 1 (a). It can be seen that both of them perform Instance-level visual-to-visual Knowledge Distillation (IKD) on possible instances of interest in the images. The key difference lies in the selection of instances, i.e., object proposals for two-stage methods and positive sample points for one-stage methods. Compared to the object proposals, there are severe inherent limitations in the positive sample points. We argue that these limitations cause the performance gap between two-stage and one-stage methods.

Specifically, as illustrated in Figure 1 (a), the positive sample points (red points) only cover the area of the objects from base categories (green boxes), so the one-stage methods can only learn the semantic knowledge about the base categories from the PVLM during the distillation. On the contrary, the class-agnostic proposals (red boxes) in two-stage methods usually cover the regions of the objects from novel categories (purple boxes), which enables the two-stage methods to implicitly learn the semantic knowledge of novel categories from the PVLM (See sec 4.3 for a clearer analysis). This advantage can effectively expand the semantic category space and further improve performance. What's more, the number of positive sample points is much less than the object proposals in most images, and each positive sample point only covers a smaller area on the feature maps than the proposals. This sparse sampling of the feature map areas during distillation also makes the semantic supervision from PVLM shrink a lot in one-stage methods.

To compensate for these inherent limitations, a straightforward approach is to make use of more sample points of the feature maps for knowledge distillation. Thus, in this work, we propose a weakly supervised global-level language-to-visual knowledge distillation method (GKD) to achieve this approach. As shown in Figure 1 (b), GKD exploits the visual captions that potentially contain semantic knowledge of novel categories, and performs language-to-visual knowledge distillation between caption representation and global-level image representation. In this way, GKD implicitly aligns all sample points in the image with the caption semantics, so that the sample points belonging to the novel categories can also learn their related semantic knowledge from the PVLM.

Finally, our proposed GKD is combined with the commonly used IKD to perform open-vocabulary one-stage detection in an end-to-end fashion, leading to a hierarchical knowledge distillation mechanism-based detector, namely HierKD. We summarize our contributions as follows:

- A weakly supervised global-level language-to-visual knowledge distillation method is explored to learn novel category knowledge beyond training labels for one-stage detection.
- An end-to-end hierarchical visual-language knowledge distillation mechanism is proposed to achieve a high-performing open-vocabulary one-stage detector.
- The proposed HierKD detector significantly surpasses the previous best open-vocabulary one-stage detector with 11.9% and 6.7% AP_{50} gains under the zero-shot detection and generalized zero-shot detection settings respectively on MS-COCO dataset.

2. Related Work

Zero-shot Learning: As the capability of image recognition with supervised learning has reached a high-level status, researchers begin to explore how well the classification models can recognize objects of novel categories beyond training sets, which is usually referred as zero-shot learning (ZSL). The earliest works start from modeling the attributes of objects by encoding the label space with binary attribute vectors for recognizing objects [5, 13, 22], while the later works focus more on the semantic representation of the visual space [7, 21, 34]. Recently, PVLM, e.g., CLIP [23], learns to model visual concepts based on the natural language like human and acquires powerful zero-shot recognition ability. Different from these image-level zero-shot recognition works, we aim at exploring the open-vocabulary instance-level detectors. Nevertheless, PVLM is also closely related to our work for we hope to transfer its zero-shot recognition ability to the open-vocabulary detection by knowledge distillation.

Zero-shot and Open-vocabulary Detection: Zero-shot and open-vocabulary detection both focus on designing a detector that can recognize and localize objects of novel categories beyond the training sets. Some works explore the two-stage detectors [1, 17, 35, 36, 39, 40] and achieve the state-of-the-art performance. Zareian *et al.* [36] designed a projection layer for aligning visual space with

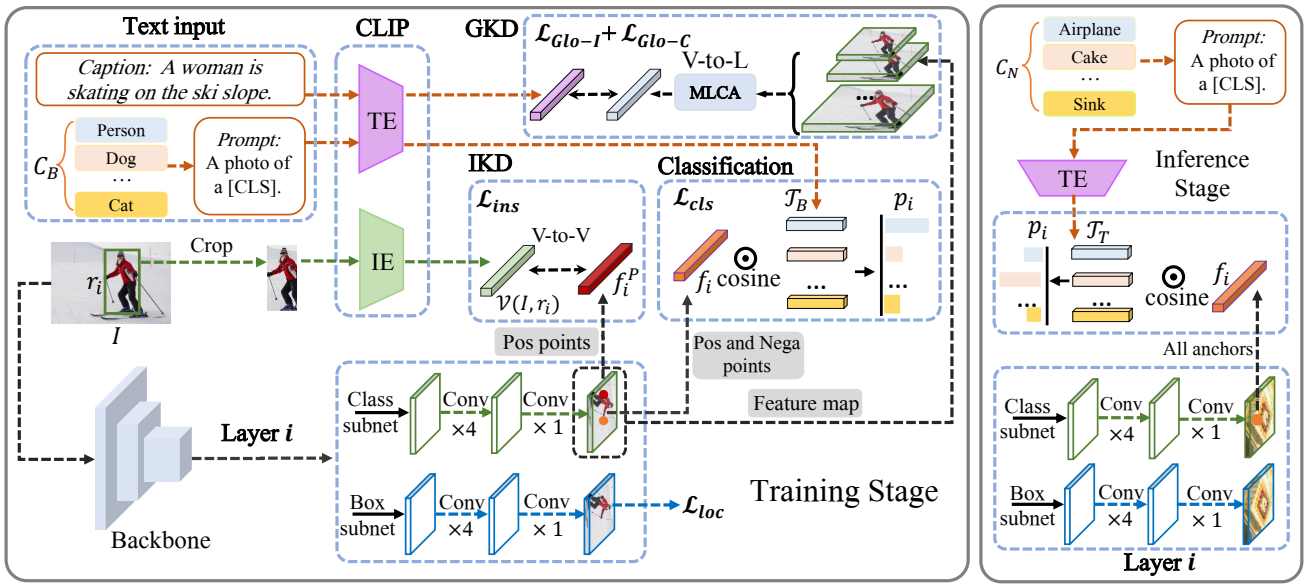

Figure 2. **Overview of our open-vocabulary one-stage detector with hierarchical visual-language knowledge distillation:** In the training stage, the classification branch is initialized with the CLIP textual embedding of base categories. For IKD, the aim is to minimize the distance between the features of sparse positive sample points on feature maps and the CLIP visual embedding of the cropped regions. The GKD aggregates all the multi-layer feature maps to directly align with the captions by cross attention. During inference, the knowledge distillation modules are removed and the CLIP textual embedding is initialized with the novel categories. It is noteworthy that the distillation has less impact on the regression branch thanks to the inherent characteristics of disentanglement in one-stage detectors.

textual semantic space based on PixelBERT [12]. Benefit from the development of knowledge distillation [11, 20, 30] and pre-trained visual-language model [23], Gu *et al*. [8] proposed to distill region-level visual features from CLIP. Another direction focuses on designing more efficient one-stage detectors by modifying loss functions [25], introducing transductive learning [24], and synthesizing features for unseen objects [42]. Xie *et al*. [35] also distilled knowledge from CLIP with a baseline one-stage detector YOLO-v5 [2]. Although it significantly surpasses the previous one-stage methods, there is still a large performance gap compared to the advanced two-stage methods. We have analyzed the reason behind the poorly performing one-stage methods during instance-level knowledge distillation, and concentrate on compensating for their inherent limitations.

3. Approach

Figure 2 illustrates the overall framework of our proposed open-vocabulary one-stage detector HierKD. It consists of a teacher pre-trained visual-language model and a student detector during the training phase. Here we employ a pre-trained visual-language model named CLIP [2] for its superior performance. The student model aims to learn the teacher model's zero-shot recognition ability by our proposed hierarchical visual-language knowledge distillation mechanism. In particular, the positive sample points learn from Image Encoder (IE) of the teacher model by instance-level visual-to-visual knowledge distillation, and the multi-scale feature maps from the detector directly transfer knowledge from Text Encoder (TE) of teacher by global-level language-to-visual knowledge distillation.

Notations: The categories in the training set, *i.e*., base categories is denoted as C_B, and the novel categories in the testing set is denoted as C_N. In addition, TE and IE of CLIP are denoted as $\mathcal{T}$ and $\mathcal{V}$, respectively. The textual embedding $\mathcal{T}_B$ used in training is initialized offline by feeding each category in C_B with a prompt, *i.e*. "a photo of a [CLS].", into the text encoder $\mathcal{T}$. During inference, the only modification is to replace C_B with C_N or the union $C_B \cup C_N$ under different settings.

3.1. Choosing and Modifying a Base Detector

The first challenge is how to adapt an off-the-shelf one-stage base detector to the open-vocabulary object detection task with necessary structural modifications.

Choosing a Base One-stage Detector: We first leverage ATSS [37] as the base one-stage detector for two reasons: (1) The adaptive training sample selection mechanism makes it a top performer in the traditional object detection task; (2) There is only one anchor at each location on the feature maps, which is important because modifying the classification layer (see below) will dramatically increase memory consumption as the number of anchors increases.

Modifying the Base Detector: We then make two modifi-

[2]CLIP ViT-B/32 is selected for fair comparisons with other methods.

cations to the original ATSS, as illustrated in Figure 2: (1) The original convolution-based classification layer is modified to the classification form of CLIP with the names or descriptions of the dataset's categories embedded by TE. A background embedding $\mathcal{T}_{bg}$ is also required since this modification would lose the original detector's ability to distinguish the background samples. The $\mathcal{T}_{bg}$ [3] is initialized by feeding "a photo of background." into $\mathcal{T}$, which allows to learn the background in the training stage. The sigmoid function is also replaced with the softmax function, and the final classification loss is based on the softmax focal loss.

$$p_i = \text{SoftMax}([\tau_c \cdot (\mathcal{T}_B f_i^T), \tau_c \cdot (\mathcal{T}_{bg} f_i^T)]) \,,$$
$$\mathcal{L}_{cls} = \frac{1}{N_{pos}} \sum_{i=0}^{N} \mathcal{L}_{\text{focalloss}}(p_i, y_i) \,, \tag{1}$$

where f_i, p_i and y_i denote the anchor feature, classification result and label of the anchor respectively. τ_c is a learnable temperature coefficient during training, and N_{pos} is the number of positive sample points while N denotes the total number of positive and negativet samle points; (2) The centerness branch in ATSS is replaced with an IOU branch [14] for mitigating the misalignment between classification task and regression task to a certain extent.

3.2. Instance-level Knowledge Distillation

We then introduce the instance-level knowledge distillation, which aims at transferring knowledge from the image encoder $\mathcal{V}$. Following the common practice, only the features of positive samples are fetched for distillation. Since the positive sample points in ATSS may have relatively small IOU values with respect to the ground-truth boxes, we set a fixed IOU threshold to further filter out the positive samples with small IOUs and acquire the features for the remaining positive sample points $\{f_1, f_2, ..., f_{N_{pos}}\}$. Unlike ZSD-YOLO [35], we use the predicted boxes of regression branch instead of the ground-truth boxes to crop regions from image I for the sake of data augmentation. These cropped regions are then resized to 224 $\times$ 224 to adapt to the input image size of $\mathcal{V}$. We use a resizing method that can keep more image information, *i.e.*, "Long side + padding", which resizes the long side to 224 and pads the short side with 0. Next, the features to be mimicked $\{\mathcal{V}(I, r_i), r_i \in \mathcal{R}\}$ can be obtained by feeding these resized regions $\mathcal{R}$ into the image encoder $\mathcal{V}$. Finally, the knowledge is transferred from the CLIP image encoder to detectors with distillation as follows:

$$\mathcal{L}_{ins} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \left\| \frac{f_i}{\|f_i\|_2} - \frac{\mathcal{V}(I, r_i)}{\|\mathcal{V}(I, r_i)\|_2} \right\|_1 . \tag{2}$$

[3] We also try to randomly initialize it [35] and set a fixed zero vector with a bias [36], but we finally get similar performance.

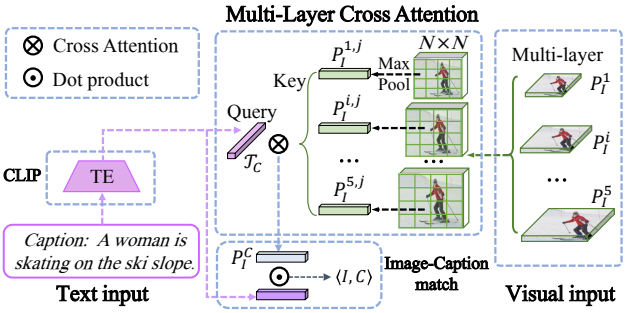

Figure 3. **Global-level knowledge distillation:** This GKD module takes the caption as textual input and the feature maps from multi layers as visual input, and learns to match the image-caption pairs by mimicking the contrastive learning in CLIP through the multi-layer cross-attention.

We have also tried with L_2 norm for mimicking, and there are no obvious differences among different measures after adjusting the appropriate loss weights.

3.3. Global-level Knowledge Distillation

To overcome the limitation of only learning from the base categories, a weakly supervised GKD module is explored by exploiting the image captions to learn the semantic knowledge of novel categories beyond training labels. GKD mimics the contrastive learning in CLIP to match the image-caption pairs and aims at transferring CLIP's large-scale semantic knowledge to the one-stage detector.

Figure 3 illustrates the overall process of GKD. Specifically, an arbitrary image denoted by I and its paired caption denoted by C are matched by Multi-Layer Cross Attention (MLCA). For the visual input, feature maps from different FPN layers are evenly divided into $N \times N$ patches, and the Max Pooling operation is performed inside all patches of different feature maps to obtain the patch-level representations. The set of pooled patch features is denote by $\{P_I^{i,j} | i = 1, 2, 3, 4, 5, j = 1, ..., N \times N\}$, where i indicates the FPN layer and j is the patch location on the feature maps of each layer. Next, for the textual input, the whole caption C is encoded directly by text encoder $\mathcal{T}$ to represent the textual feature $\mathcal{T}_C$. As the CLIP model is great at extracting the overall high-level textual feature while not at word-level representation in some simple visual-grounding experiments, we choose to take the feature of the entire caption instead of each word.

After obtaining the textual feature and the set of multi-layer patch features, the cross attention takes these multi-modal inputs to aggregate the patch features. Specifically, the caption is regarded as query, all patches are regarded as keys, and the response between the query and each key can be calculated via cosine similarity. Hence, the aggregation of all patch features is obtained with the normalized simi-

larities as follows:

$$e_{i,j} = \frac{\mathcal{T}_C \cdot P_I^{i,j}}{\|\mathcal{T}_C\| \|P_I^{i,j}\|} ,$$
$$P_I^C = \sum_{i=j=1}^{5,k} \frac{exp(e_{i,j})}{\sum_{i'=j'=1}^{5,k} exp(e_{i',j'})} P_I^{i,j} , \tag{3}$$

where P_I^C represents the caption-aware visual feature aggregation, and $e_{i,j}$ is the response between the caption and the j_{th} patch of i_{th} layer. Finally, the matching score $\langle I, C \rangle$ between the image-caption pair (I, C) is:

$$\langle I, C \rangle = \frac{P_I^C \cdot \mathcal{T}_C}{\|P_I^C\| \|\mathcal{T}_C\|} . \tag{4}$$

Since the aim of our global-level knowledge distillation is to transfer CLIP's large-scale semantic knowledge to the detector, it is naturally to mimic the contrastive learning in CLIP and also the recent self-supervised learning works [4, 9]. The paired images and captions are regarded as positive pairs in a batch while the others are negative pairs. We introduce a symmetrical contrastive loss function to push the positive pairs and pull the negatives in semantic space:

$$\mathcal{L}_{Glo-I} = -log \frac{exp(\tau_m \cdot \langle I, C \rangle)}{\sum_{C_i=1}^{b} exp(\tau_m \cdot \langle I, C_i \rangle)} ,$$
$$\mathcal{L}_{Glo-C} = -log \frac{exp(\tau_m \cdot \langle I, C \rangle)}{\sum_{I_i=1}^{b} exp(\tau_m \cdot \langle I_i, C \rangle)} , \tag{5}$$

where τ_m is a trainable temperature coefficient, and b denotes the batch size.

Finally, the hierarchical knowledge distillation of our one-stage detector can be formulated by combining the instance-level knowledge distillation and global-level knowledge distillation:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{loc}\mathcal{L}_{loc} + \lambda_{ins}\mathcal{L}_{ins} + \lambda_{Glo}(\mathcal{L}_{Glo-I} + \mathcal{L}_{Glo-C}) . \tag{6}$$

3.4. Sampling the Negative Samples

Advanced one-stage detectors often combine focal Loss [19] or its variants [6, 15, 16] with all negative samples to solve the imbalance problem between positive and negative samples. However, this setting is troubling in open-vocabulary detection, for the detectors will identify more foreground regions as background when generalizing to novel categories in experiments. On the other hand, sampling the negative samples to 1:1 with positive samples as in two-stage methods will boost the performance on novel categories, whereas it seriously affects the base categories. To make a trade-off between the above options, we adopt a sampling strategy by sampling 10% negative samples to boost the recall performance on novel categories while maintaining the performance on base categories.

3.5. Direct Inference Alternative with CLIP

As the zero-shot recognition ability of the proposed method is transferred from CLIP, we can thus measure the mimicking ability of our method by comparing the performance gap between our model and this CLIP direct inference. We design a simple CLIP direct inference way in algorithm 1. Essentially, it compares the difference in the classification results of the same sample points between the detector and the CLIP.

Algorithm 1: CLIP Direct Inference

Input: CLIP image encoder $\mathcal{V}$ and text encoder $\mathcal{T}$, novel categories C_N, trained model $\mathcal{M}$, test images $\mathcal{D}_T$

Output: Detection boxes B

$\mathcal{T}_N \leftarrow \mathcal{T}(Prompt(C_N))$ and normalize;
for $I \in D_T$ **do**
  $A_I \leftarrow \mathcal{M}_{anchor}(I)$;
  $A_I^{fore} \leftarrow \arg\max_k(\mathcal{M}_{cls}(A_I) \times \mathcal{M}_{IOU}(A_I))$;
  $\mathcal{V}_I^{fore} \leftarrow \mathcal{V}(I, \mathcal{M}_{loc}(A_I^{fore}))$ and normalize;
  $S_I^{fore} \leftarrow Softmax(\tau \cdot \mathcal{T}_N \mathcal{V}_I^{fore})$;
  $B \leftarrow B \cup NMS(S_I^{fore}, box_I^{fore})$;
end

4. Experiments and Results

4.1. Dataset and Evaluation Protocol

We validate our method on the MS-COCO 2017 benchmark under both zero-shot detection (ZSD) and generalized zero-shot detection (GZSD) settings. In the previous ZSD literature, two different types of base/novel split settings are available: the 48/17 and the 65/15 base/novel splits by Bansal et al. [1] and Rahman et al. [25], respectively. We evaluate both split settings in this paper. Our data preprocessing is the same as Rahman et al. [36]. Following the most previous ZSD methods, we evaluate our method using mAP and Recall@100 at IOU=0.5, and mainly focus on the performance of novel categories.

4.2. Implementation Details

Our implementation and hyper-parameter settings are based on MMdetection [3]. A standard ResNet-50 [10] is adopted as the backbone, and all hyper-parameters remain the default settings unless otherwise specified. We set the thresholds of NMS and classification score to 0.4 and 0.0 respectively. The temperature coefficients τ_c and τ_m are initialized to 100 and 10 respectively. We also add a gradient clip at 10.0 during the training stage. For the knowledge distillation, the teacher model CLIP is frozen, and the feature maps of different FPN layers are divided into 3×3 patches. We train the model on 4 Tesla V100 GPUs and use a batch

IOU	Base/Novel	AR@100	AR@300	AR@1000
0.5	48/17	61.9	76.9	87.5
0.75	48/17	37.4	48.1	57.4

Table 1. Generalization ability of RPN

Norm	Weight	Region	Area	AR_{50}	AP_{50}
L_1	1	*pred*	1×	62.4	14.6
L_2	1	*pred*	1×	**65.1**	12.8
L_2	10	*pred*	1×	63.6	14.6
L_1	1	*GT*	1×	62.8	14.5
L_1	1	*pred*	1.5×	64.5	**15.3**

Table 2. Comparisons between different sub-module options in IKD. *pred* and *GT* mean cropping regions from prediction boxes and ground-truth boxes, respectively. 1× and 1.5× represent cropping the original box and its 1.5× center expansion respectively.

size of 16 in IKD and 32 in GKD and HierKD. The learning schedule follows the traditional object detection settings.

4.3. Test on generalization ability of RPN

To more clearly illustrate the generalization ability of the RPN, we train the RPN on the base categories and directly transfer it to test on the novel categories. As shown in Table 1, the category-agnostic proposals in RPN of two-stage methods usually cover the regions of the novel objects, and AR is still up to 37.4 when generating 100 proposals and IOU=0.75, which contributes to feature learning on novel categories during knowledge distillation.

4.4. Ablation Study

We conduct ablation studies on the MS-COCO ZSD benchmark to verify the effectiveness of design choices. All the results are reported on the novel categories under the 48/17 base/novel split setting unless otherwise specified.

Instance-level Knowledge Distillation: We compare the impact of different sub-module options in the instance-level knowledge distillation in Table 2. Compared to our distillation using L_1 norm, replacing it with L_2 loss norm will cause a 1.8% AP_{50} drop, and this gap can be reduced through increasing the loss weight. It is essentially because the distance between the features measured by the L2 norm requires a larger weight to be consistent with the result of the L1 norm. The cropped region factor used in knowledge distillation is not sensitive to using prediction boxes or ground-truth boxes. However, it can improve performance by cropping the 1.5× expanded box area to provide more contextual information.

Global-level Knowledge Distillation: As shown in Table 3, the different choices of sub-modules have great impacts on the performance. First, We observe that the AP_{50} achieved by using Average Pooling is only about half of Max Pooling. This is caused due to the loss of distinguishability of the patch features obtained through Average Pooling. Moreover, compared to dividing the feature maps into a small number of patches, such as 3×3 or 4×4, dividing it into more patches, such as 8×8, brings a significant AP_{50} drop. It can be attributed to the reason that more training iterations are required to converge for more patches. Additionally, there is no obvious difference between 8 *bs/gpu* and 4 *bs/gpu*. We infer that both of them we can afford are too small for contrastive learning to make a difference. Finally, replacing contrastive learning with only pushing the positive pairs brings a 2.8% AP_{50} drop. This shows that contrastive learning can better transfer the zero-shot recognition capability of the PVLM.

Patch	*Pool*	*Loss*	*bs/gpu*	AR_{50}	AP_{50}
4	*Ave*	*CL*	*8*	59.2	12
4	*Max*	*CL*	*8*	64.2	20.1
3	*Max*	*CL*	*8*	61.1	**20.7**
8	*Max*	*CL*	*8*	60.8	13.7
3	*Max*	*PL*	*8*	60.9	17.9
3	*Max*	*CL*	*4*	**65.6**	20.5

Table 3. Comparisons between different sub-module options in GKD. *Ave* and *Max* represent using Average Pooling and Max Pooling to obtain patch features respectively. *CL* denotes training with the contrastive learning loss, while *PL* only considers the cosine similarities between positive pairs. *bs/gpu* is the batch size on each GPU during training.

IKD	GKD	IOU_b	AR_{50}	AP_{50}	AP_S	AP_M	AP_L
-	-	-	52.4	10.2	8.8	12.5	12.8
√			62.4	14.6	10.1	13.2	19.1
	√		61.1	20.7	10.1	28.5	27.5
√	√		70.1	20.7	11.5	30.2	27.0
√	√	√	**71.3**	**21.6**	**11.6**	**30.7**	**28.1**

Table 4. Verify the effectiveness and compatibility of each module. The first row is the baseline, which is the exploited base detector trained with only classification loss and localization loss.

Distillation Module Analysis: We quantitatively verify the effectiveness of each distillation module and the compatibility of different modules. We additionally report the detection performance on small objects AP_S, medium objects AP_M and large objects AP_L to perform a more detailed analysis. As shown in Table 4, by adding IKD and GKD to the baseline, we can obtain 4.4% and 10.5% AP_{50} gains as well as 10.0% and 8.7% AR_{50} gains respectively. This validates the effectiveness of each distillation module. In addition, compared to applying IKD and GKD separately, the combination of IKD and GKD, *i.e.*, HierKD, further brings 7.7% and 9.0% AR_{50} gains respectively. This shows that the great compatibility of IKD and GKD. The AP_S and AP_M in HierKD are improved by 1.4% and 1.7% compared to GKD, which shows that HierKD has advantages in detecting small and medium objects. Finally, using IOU branch leads to more improvements on the medium and large objects than the small. It may be because objects with low classification scores and high IOUs generally do not appear on small objects.

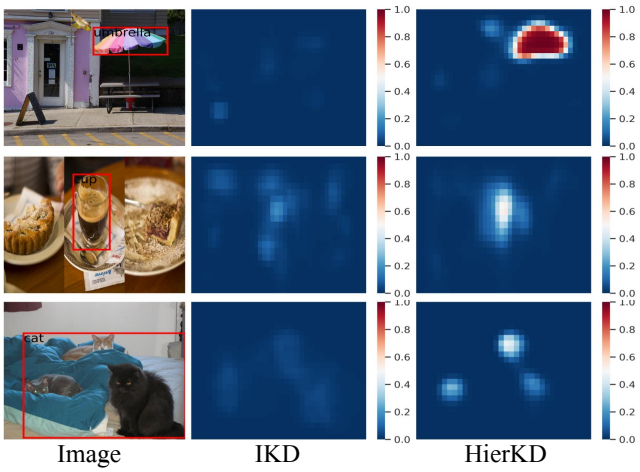

Figure 4. Spatial distribution of classification score. The red boxes in the images are ground-truth of the novel categories. The heatmaps in IKD and HierKD show the classification score of anchors at each location for the categories in the red boxes.

Negative samples	IKD	GKD	Base		Novel	
			AR_{50}	AP_{50}	AR_{50}	AP_{50}
1:1	√		71.0	37.0	**63.0**	**16.8**
10%	√		**75.9**	44.3	62.4	14.6
100%	√		74.5	**44.4**	60.3	9.0
1:1		√	69.2	34.9	60.2	19.3
10%		√	**74.0**	**42.7**	**61.1**	**20.7**
100%		√	72.4	42.6	56.4	18.7

Table 5. Comparisons between different sampling strategies for negative samples. 1:1, 10%, 100% mean sampling the same number of negative samples as the positive samples, sampling 10 % of the negative samples, and using all the negative samples.

We also visualize some classification score distribution and detection results for qualitative analysis. Figure 4 illustrates the spatial distribution of classification score in IKD and HierKD, respectively. We can see that IKD often fails to identify the objects of novel categories, *e.g.*, the "umbrella" in the first row. In addition, IKD may also have low confidence in recognition of the objects of novel categories, such as the "cup" in the second row and the "cat" in the third row. By introducing GKD, the proposed HierKD can recognize the "umbrella" in the first row, and also significantly increases the confidence in recognition of the "cup" in the second row and the "cat" in the third row. This shows that our HierKD can better transfer the novel category knowledge from CLIP and reduce missed detections while increasing detection confidence. We also show some detection results of novel categories in Figure 5. First, it can be seen that GKD and HierKD can identify more objects of novel categories compared to IKD, such as the "umbrella" in the second row and the "airplane" in the third row. Moreover, GKD and HierKD also have higher classification accuracy, such as correctly classifying the "elephant" in the first row instead of recognizing it as a "cow" like IKD.

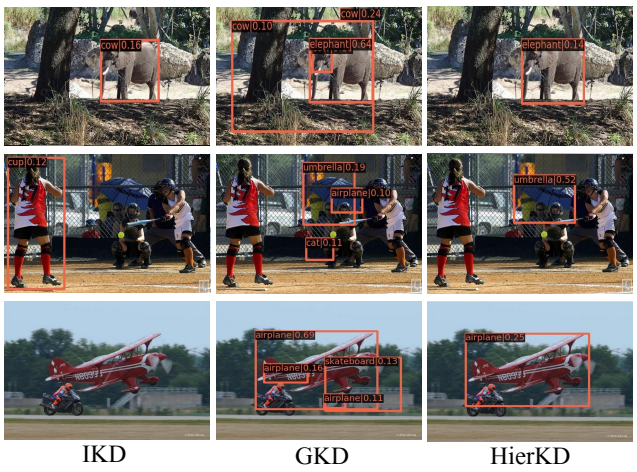

Figure 5. Visualization of some detections on novel categories.

	Model	CLIP	AR_{50}	AP_{50}	AP_S	AP_M	AP_L
IKD	√		62.4	14.6	10.1	13.2	19.1
		√	**66.6**	**24.9**	**18.3**	**28.3**	**32.3**
GKD	√		61.1	20.7	10.1	**28.5**	27.5
		√	**64.5**	**23.3**	**18.2**	27.9	**30.3**
HierKD	√		**70.1**	20.7	11.5	**30.2**	27.0
		√	65.8	**22.8**	**18.8**	27.4	**29.9**

Table 6. Comparison with the direct inference with CLIP. Model and CLIP represent inference with the detector from distillation and direct inference with CLIP respectively.

Compared to GKD, HierKD can suppress more meaningless detection results, such as the multiple partial "airplane" in the third row.

Sampling the Negative Samples: The impact of the sampling strategy for negative samples is shown in Table 5. Taking 100% sampling as the baseline, we can see that 10% sampling does not cause a large AP_{50} drop on the base categories in comparison with 1:1 sampling. When generalizing to novel categories, the AP_{50} obtained by 10% sampling is not much worse than the 1:1 sampling in IKD while achieving the best in GKD. This validates the effectiveness of the 10% sampling strategy.

Compared to Direct Inference with CLIP: The performance gap between the proposed method and direct inference with CLIP is shown in Table 6. The IKD baseline has only about half of the AP_{50} compared to direct inference with CLIP on all sizes of objects, while our proposed GKD achieves similar performance on medium and large objects. The final HierKD has higher AR_{50} than direct inference with CLIP. However, the AP_S in HierKD lags behind the direct inference with CLIP a lot, which shows that our method has insufficient learning ability for small objects.

Different Training Settings: As shown in Table 8, extending the period of training schedule from 1× to 2×, 3×, introducing scale jitter (480-800), and changing backbone to larger ResNet-101 can improve the performance of both

Method			Base/Novel	ZSD	GZSD		
				Novel	Base	Novel	All
TS	ZS	SB [1]	48/17	0.70	29.2	0.31	24.9
		LAB [1]	48/17	0.27	20.8	0.22	18.0
		DESE [1]	48/17	0.54	26.7	0.27	22.1
		BLC [39]	48/17	9.9	42.1	4.50	32.3
		ZSI* [40]	48/17	11.4	46.5	4.83	35.6
	OV	OVR-CNN [36]	48/17	16.7	-	-	34.3
		ViLD* [8]	48/17	-	59.5	27.6	51.3
OS	ZS	PL* [25]	48/17	10.0	35.9	4.12	27.9
		DELO [42]	48/17	7.6	13.8	3.41	13.0
	OV	ZSD-YOLO* [35]	48/17	13.4	31.7	13.6	27.0
		HierKD(ours)	48/17	**25.3**	**51.3**	**20.3**	**43.2**
TS	ZS	BLC [39]	65/15	13.1	36.0	13.1	31.7
		ZSI* [40]	65/15	13.6	38.7	13.6	34.0
OS	ZS	PL* [25]	65/15	12.4	34.1	12.4	30.0
	OV	ZSD-YOLO* [35]	65/15	18.3	31.7	17.9	29.2
		HierKD(ours)	65/15	**27.4**	**48.9**	**20.4**	**43.6**

Table 7. **Comparison with other state-of-the-art methods:** * denotes the state-of-the-art methods in various settings. "TS" and "OS" are abbreviation of two-stage and one-stage detectors, respectively. Note that we classify Cascade R-CNN based detectors as generalized two-stage methods. "ZS" and "OV" indicate that the models belong to zero-shot and open-vocabulary detectors, respectively.

the base and novel categories. This validates that the proposed HierKD is compatible with the general detection performance improvement techniques.

4.5. Comparison with the Start-of-the-Art

We compare our HierKD with the other two-stage methods and one-stage methods on the MS-COCO benchmark in Table 7, all metrics reported in Table 7 are AP_{50}. **Limitation:** we can not make a completely fair comparison like the traditional object detection because the factors of batch size, scale jitter, *etc*., used in some works (such as ViLD [8]) are different from the general settings.

We can observe that under the 48/17 base/novel split setting, HierKD achieves 25.3% AP_{50} on novel categories under the ZSD setting. HierKD significantly outperforms the previous best one-stage method ZSD-YOLO with 11.9% AP_{50} gains, and also exceeds the most recent two-stage method OVD (trained without external Conceptual Caption dataset [32]) by 8.6% AP_{50}. Under the GZSD setting, HierKD outperforms ZSD-YOLO with 6.7% gains on novel categories. HierKD also reduces the AP_{50} performance gap from 14% to 7.3% compared to the best two-stage method ViLD. Under the GZSD setting, the AP_{50} of HierKD on the novel categories is 5% lower than that of the ZSD. This is caused by the detection confidence of the novel categories is lower than that of the base categories, so some detection results of novel categories are suppressed during NMS.

Under another 65/15 base/novel split setting, HierKD surpasses the previous best method ZSD-YOLO with 10.1% and 2.5% AP_{50} gains on novel categories under ZSD and GZSD settings respectively.

Backbone	Schedule	Scale Jitter	Base		Novel	
			AR_{50}	AP_{50}	AR_{50}	AP_{50}
ResNet-50	1×		74.8	44.7	71.3	21.6
ResNet-50	2×		77.5	49.0	69.8	23.1
ResNet-50	3×	√	80.0	51.8	70.0	25.3
ResNet-101	3×	√	**80.8**	**53.5**	**71.4**	**27.3**

Table 8. Verification of the compatibility with general detection performance improvement techniques.

	Base/Novel	AR_{50}	AP_{50}	AP_S	AP_M	AP_L
HierKD	48/17	**71.4**	27.3	11.4	39.5	37.3
Upper Bound	48/17	70.7	**68.0**	**36.3**	**74.5**	**87.4**

Table 9. Comparison with the ideal upper bound. All reported metrics are results on the novel categories.

4.6. Upper Bound Analysis

We can get the ideal upper bound of this type of distillation method by directly using CLIP to classify the instances in the ground-truth boxes and then evaluating the detection results, *i.e*., the classification results of ground-truth boxes. As shown in Table 9, our method achieves a relatively high recall, while the total AP_{50} and AP on objects of various sizes, *i.e*. AP_S, AP_M, AP_M are still far from the upper bound. This shows that there is still much room to improve the mimicking ability of the proposed HierKD. In addition, this also reminds us of using techniques such as prompt learning [41] to improve the zero-shot recognition ability of CLIP itself, thereby further improving the upper bound of model performance.

5. Conclusion

In this work, we have developed a hierarchical visual-language knowledge distillation method, namely HierKD, to obtain a top-performing one-stage open-vocabulary detector. HierKD uses image caption to distill knowledge in a language-to-visual manner. The rich vocabulary in captions enables HierKD to transfer the semantic knowledge of novel categories from CLIP during training. The results indicate that the proposed HierKD can identify novel objects more accurately and confidently, and significantly surpasses the previous methods. In the future, we will continue to explore more efficient and advanced distillation methods to transfer the zero-shot recognition ability of teacher models.

Acknowledgment This work was supported by the National Key R&D Program of China (Grant No. 2018AAA0102803, 2018AAA0102800), the Natural Science Foundation of China (Grant No. U2033210, 62172413, 61972394, 62036011, 62192782, 61721004), the Key Research Program of Frontier Sciences, CAS (Grant No. QYZDJ-SSW-JSC040), the China Postdoctoral Science Foundation (Grant No. 2021M693402). Jin Gao was also supported in part by the Youth Innovation Promotion Association, CAS.

References

[1] 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. 1, 2, 5, 8

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1, 3

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[5] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009. 2

[6] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3510–3519, October 2021. 5

[7] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013. 2

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 2, 3, 8

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 5

[10] 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

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[12] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3

[13] Dinesh Jayaraman and Kristen Grauman. Zero shot recognition with unreliable attributes. *arXiv preprint arXiv:1409.4327*, 2014. 2

[14] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 2020. 4

[15] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2021. 5

[16] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 5

[17] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8690–8697, 2019. 2

[18] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 5

[20] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019. 3

[21] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[22] Mark M Palatucci, Dean A Pomerleau, Geoffrey E Hinton, and Tom Mitchell. Zero-shot learning with semantic output codes. 2009. 2

[23] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3

[24] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6082–6091, 2019. 3

[25] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939, 2020. 2, 3, 5, 8

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[27] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1

[28] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[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*, 28:91–99, 2015. 1

[30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[31] Alexander I Saichev, Yannick Malevergne, and Didier Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer Science & Business Media, 2009. 1

[32] 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. 8

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1

[34] 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. 2

[35] Johnathan Xie and Shuai Zheng. Zsd-yolo: Zero-shot yolo detection using vision-language knowledgedistillation. *arXiv preprint arXiv:2109.12066*, 2021. 2, 3, 4, 8

[36] 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, 4, 5, 8

[37] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[38] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang. Gtnet: Generative transfer network for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12967–12974, 2020. 2

[39] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 8

[40] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2593–2602, 2021. 2, 8

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 8

[42] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020. 3, 8