

Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection

Luting Wang^{1,3} Yi Liu^{1,3} Penghui Du^{1,3} Zihan Ding^{1,3} Yue Liao^{1,3*}
 Qiaosong Qi² Biaolong Chen² Si Liu^{1,3}
¹Institute of Artificial Intelligence, Beihang University ²Alibaba Group
³Hangzhou Innovation Institute, Beihang University

Abstract

Open-vocabulary object detection aims to provide object detectors trained on a fixed set of object categories with the generalizability to detect objects described by arbitrary text queries. Previous methods adopt knowledge distillation to extract knowledge from Pretrained Vision-and-Language Models (PVLMs) and transfer it to detectors. However, due to the non-adaptive proposal cropping and single-level feature mimicking processes, they suffer from information destruction during knowledge extraction and inefficient knowledge transfer. To remedy these limitations, we propose an Object-Aware Distillation Pyramid (OADP) framework, including an Object-Aware Knowledge Extraction (OAKE) module and a Distillation Pyramid (DP) mechanism. When extracting object knowledge from PVLMs, the former adaptively transforms object proposals and adopts object-aware mask attention to obtain precise and complete knowledge of objects. The latter introduces global and block distillation for more comprehensive knowledge transfer to compensate for the missing relation information in object distillation. Extensive experiments show that our method achieves significant improvement compared to current methods. Especially on the MS-COCO dataset, our OADP framework reaches 35.6 mAP₅₀^N, surpassing the current state-of-the-art method by 3.3 mAP₅₀^N. Code is released at <https://github.com/LutingWang/OADP>.

1. Introduction

Open-vocabulary object detection (OVD) [49] aims to endow object detectors with the generalizability to detect open categories including both *base* and *novel* categories where only the former are annotated in the training phase. Pretrained Vision-and-Language Models (PVLMs, e.g., CLIP [32] and ALIGN [19]) have witnessed great progress in recent years, and Knowledge Distillation (KD) [17] has led to a wave of unprecedented advances transferring the zero-shot visual recognition ability from PVLMs to detec-

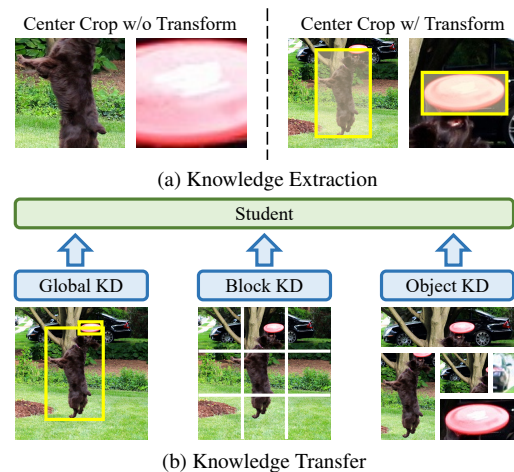


Figure 1. An overview of our OADP framework. (a) Directly applying center crop on proposals may throw informative object parts away, resulting in ambiguous image regions. In contrast, our OAKE module extracts complete objects and reduces the influence of surrounding distractors. (b) Our DP mechanism includes global, block, and object KD to achieve effective knowledge transfer.

tors [12, 25, 28, 29, 48, 53]. KD typically comprises two essential steps, i.e., *knowledge extraction* and then *knowledge transfer*. A common practice in OVD is to crop objects with class-agnostic proposals and use the teacher (e.g., CLIP visual encoder) to extract knowledge of the proposals. The knowledge is then transferred to the detector (e.g., Mask R-CNN [15]) via feature mimicking.

Despite significant development, we argue that conventional approaches still have two main limitations: 1) *Dilemma between comprehensiveness and purity during knowledge extraction*. As proposals have diverse aspect ratios, the fixed center crop strategy to square them may cut out object parts (fig. 1a). Enlarging those proposals via resizing function may alleviate this problem, but additional surrounding distractors may confuse the teacher to extract accurate proposal knowledge. 2) *Missing global scene understanding during knowledge transfer*. Conventional approaches merely concentrate on object-level knowledge transfer by directly mimicking the teacher’s features of indi-

*Corresponding author (liaoyue.ai@gmail.com)

vidual proposals. As a result, the student cannot fully grasp the contextual characteristics describing the interweaving of different objects. In light of the above discussions, we propose an Object-Aware Distillation Pyramid (OADP) framework to excavate the teacher’s knowledge accurately and effectively transfer the knowledge to the student.

To preserve the complete information of proposals while extracting their CLIP image embeddings, we propose an Object-Aware Knowledge Extraction (OAKE) module. Concretely, given a proposal, we square it with an adaptive resizing function to avoid destroying the object structure and involve object information as much as possible. However, the resizing process inevitably introduces environmental context, which may contain some distractors that confuse the teacher. Therefore, we propose to utilize an object token [OBJ] whose interaction manner during the forward process is almost the same as the class token [CLS] except that it only attends to patch tokens covered by the original proposal. In this way, the extracted embeddings contain precise and complete knowledge of the proposal object.

To facilitate complete and effective knowledge transfer, we propose a Distillation Pyramid (DP) mechanism (fig. 1b). As previous works only adopt object distillation to align the feature space of detectors and PVLMS, the relation between different objects is neglected. Therefore, we propose global and block distillation to compensate for the missing relation information in object distillation. For global distillation, we optimize the \mathcal{L}_1 distance between the detector backbone and the CLIP visual encoder so that the detector learns to encode rich semantics implied in the image scene. However, the CLIP visual encoder is prone to ignore background information, which may also be valuable for detection. Therefore, we take a finer step to divide the input image into several blocks and optimize the \mathcal{L}_1 distance between the block embeddings of the detector and the CLIP image encoder. Overall, the above three distillation modules constitute a hierarchical distillation pyramid, allowing for the transfer of more diversified knowledge from CLIP to the detectors.

We demonstrate the superiority of our OADP framework on MS-COCO [27] and LVIS [14] datasets. On MS-COCO, it improves the state-of-the-art results of mAP_{50}^N from 32.3 to 35.6. On the LVIS dataset, our OADP framework reaches 21.9 AP_r on the object detection task and 21.7 AP_r on the instance segmentation task, leading the former methods by more than 1.1 AP_r and 1.9 AP_r respectively.

2. Related Work

Knowledge Distillation for Object Detection. KD [17,37] is a technology that helps train compact student models under the supervision of powerful teacher models. Chen *et al.* [5] apply KD to object detection by implementing feature-based and response-based loss for Faster R-CNN.

Li *et al.* [23] apply \mathcal{L}_2 loss on features sampled by student proposals. FGFI [40] only distills foreground regions near the object anchors. DeFeat [13] distills the foreground and background regions simultaneously with different factors. GID [7] distills regions where the student and teacher perform differently. G-DetKD [46] proposes a general distillation framework for object detectors. FKD [54] distills the attention map to emphasize the changeable areas. FGD [43] proposes focal and global distillation for comprehensive knowledge transfer. Compared to these detection KD methods [9, 21, 30, 44, 50], our work concentrates on knowledge transfer from PVLMS to detectors to enable open-vocabulary detection.

Open-Vocabulary Detection. OVD [4, 20, 29, 45] aims to train a model that can detect objects of arbitrary categories, even if the categories are not seen during training. OVR-CNN [49] is the seminal work that proposes this problem and achieves great performance using image captions as well as bounding box annotations. With the prevalence of PVLMS [19,32], ViLD [12] proposes to distill the open-vocabulary knowledge from CLIP to the detector. DetPro [8] improves upon ViLD with prompt optimization. RegionCLIP [53] develops a pretraining strategy to learn region-text alignment. Detic [55] adopts weak supervisions to jointly train the detector. GLIP [22] pretrains on massive image-text pairs in a self-training fashion by unifying the detection and grounding tasks. HierKD [28] proposes instance- and global-level distillation for one-stage detectors. OV-DETR [48] turns DETR into an open-vocabulary detector with conditional binary matching. VL-PLM [52] leverages pseudo labels on novel categories to augment the detector. PB-OVD [11] generates pseudo labels based on the image captions. PromptDet [10] establishes a scalable pipeline with regional prompt learning and self-training. In this paper, we propose an OADP framework focusing on comprehensive object knowledge extraction and effective knowledge transfer.

3. OVD Benchmarks

According to the training data, we summarize the existing OVD methods into four types of benchmarks: Vanilla OVD (V-OVD), Caption-based OVD (C-OVD), Generalized OVD (G-OVD), and Weakly Supervised OVD (WS-OVD). All benchmarks rely on instance-level annotations and large-scale image-text pairs to learn OVD. Some of them use more types of data, as shown in tab. 1. For clarity, we define base categories as those included in the instance-level annotations, and novel categories are the others.

V-OVD [4, 8, 12, 20, 22, 29, 45, 53] is a pure OVD benchmark setting, which requires the detector only to train in an object detection dataset with fixed categories set. Any information about the novel categories is unavailable, but unan-

Benchmark	Caption	Category Prior	Image Label
V-OVD			
C-OVD	✓		
G-OVD		✓	
WS-OVD	✓	✓	✓

Table 1. Summary of OVD benchmarks. ‘‘Caption’’: in-domain captions like COCO-Captions. ‘‘Category Prior’’: human priors on novel categories. ‘‘Image Label’’: image-level category labels.

notated data is allowed. A common practice for this benchmark is to learn open vocabulary knowledge from image-text pairs and transfer the knowledge to detectors through transfer learning or knowledge distillation. V-OVD is similar to ZSD [1, 33, 42, 57], except that V-OVD relies on large-scale image-text pairs to acquire open-vocabulary knowledge. Recently, V-OVD has attracted more and more researchers with the development of PVLMs.

C-OVD [3, 11, 28, 49] adds additional image caption annotation to the V-OVD benchmark. Note that by image caption data, we refer to the in-domain captions of the instance-level annotations, *e.g.*, COCO-Captions [6], instead of the large-scale image-text pairs, *e.g.*, CC3M [39] and CLIP400M [32]. The in-domain captions enrich the instance-level annotations and imply a distribution of potential novel categories. Compared with the V-OVD benchmark, C-OVD requires slightly more annotations and is expected to perform better.

G-OVD [10, 48, 52] introduces human priors on novel categories to the V-OVD benchmark. Intuitively, if some novel categories are far more likely to appear during inference, it would be beneficial to prepare for them during training. Most existing methods assume that all the dataset’s category names (including the novels) are known to the detectors during training. Therefore, the performance of G-OVD methods may not be fairly comparable with V-OVD and C-OVD methods. A typical solution is to generate instance-level pseudo annotations for the categories.

WS-OVD [55] further takes advantage of image-level category labels beyond G-OVD. Similar to Weakly Supervised Detection (WSD) [2, 47], the image-level category labels reflect the presence of the base and novel categories in each image. Thus, the annotation cost is far more than the benchmarks above. In this case, WS-OVD methods have the greatest potential to push the limit of OVD further.

4. Object-Aware Distillation Pyramid

We first briefly review the task definition of OVD and the architecture of Faster R-CNN in sec. 4.1. Then, we present the overview of our OADP framework in sec. 4.2. Sec. 4.3 and sec. 4.4 introduce the OAKE module and the DP mechanism in detail. Finally, in sec. 4.5, we demonstrate the procedure to generate pseudo labels based on OAKE.

4.1. Preliminaries

We represent traditional object detection datasets as $\mathcal{D} = \{(\mathbf{I}_i, \mathcal{O}_i)\}_{i=1}^{|\mathcal{D}|}$, where \mathbf{I}_i is the i^{th} image and $\mathcal{O}_i = \{o_{ij}\}_{j=1}^{|\mathcal{O}_i|}$ is the corresponding set of annotated objects. Each object o is a pair of object bounding box $b \in \mathbb{R}^4$ and category $y \in \mathcal{C}$, where \mathcal{C} is the category space of the dataset. We denote the training and validation datasets as \mathcal{D}^T and \mathcal{D}^V , respectively.

By the convention of OVD, we refer to the category space of \mathcal{D}^T and \mathcal{D}^V as \mathcal{C}^B and \mathcal{C} respectively. Normally, $\mathcal{C}^B \subset \mathcal{C}$. Categories in \mathcal{C}^B are called base categories, and those that only appear in \mathcal{D}^V are called novel categories. The novel category space is denoted as $\mathcal{C}^N = \mathcal{C} \setminus \mathcal{C}^B \neq \emptyset$. For each category $c \in \mathcal{C}$, we use a pretrained text encoder \mathcal{T} to encode its semantic embedding $t_c \in \mathbb{R}^d$. Specifically, we use a trainable embedding $t_{\text{bg}} \in \mathbb{R}^d$ to represent the background class `bg`.

Since our work is based on Faster R-CNN [36], we briefly recap its framework. Given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the backbone (including FPN [26]) encodes a set of hierarchical feature maps $\mathcal{F} = \{\mathbf{F}_2, \mathbf{F}_3, \dots, \mathbf{F}_6\}$ and the Region Proposal Network (RPN) generates a set of proposals $\mathcal{P} \subset \mathbb{R}^4$. Then the R-CNN head performs RoI Align on \mathcal{F} to extract proposal embeddings $\mathcal{E} = \{e_p\}_{p \in \mathcal{P}} \subset \mathbb{R}^d$. The logit of a proposal p being of category c can be defined as:

$$l(p, c) = \frac{e_p \cdot t_c}{\|e_p\| \cdot \|t_c\|}, \quad (1)$$

where \cdot is the dot product, t_c is the category embedding of c . For simplicity, we ignore the temperature τ in CLIP [32]. The probability of a proposal p belonging to category $c \in \mathcal{C} \cup \{\text{bg}\}$ is:

$$P_C(p, c) = \frac{\exp(l(p, c))}{\sum_{c' \in \mathcal{C} \cup \{\text{bg}\}} \exp(l(p, c'))}. \quad (2)$$

During training, each proposal p is assigned a category label $y_p \in \mathcal{C}^B \cup \{\text{bg}\}$. The R-CNN loss is defined as:

$$\mathcal{L} = - \sum_{p \in \mathcal{P}} \log P_{\mathcal{C}^B}(p, y_p). \quad (3)$$

For simplicity, we ignore the regression term in \mathcal{L} .

4.2. Overview of OADP

To inject open-vocabulary concepts into Faster R-CNN, we propose an Object-Aware Distillation Pyramid (OADP) framework (fig. 2), which first extracts knowledge from CLIP [32] and then transfers it to the detector through knowledge distillation (KD) [17]. Specifically, we propose an Object-Aware Knowledge Extraction (OAKE) module, which inserts an `[OBJ]` token into the frozen CLIP visual encoder \mathcal{V} to extract informative knowledge from expanded region proposals selectively. For more effective knowledge transfer, we propose a Distillation Pyramid (DP) mechanism comprising an object distillation module \mathcal{M}^O , a block distillation module \mathcal{M}^B , and a global distillation module

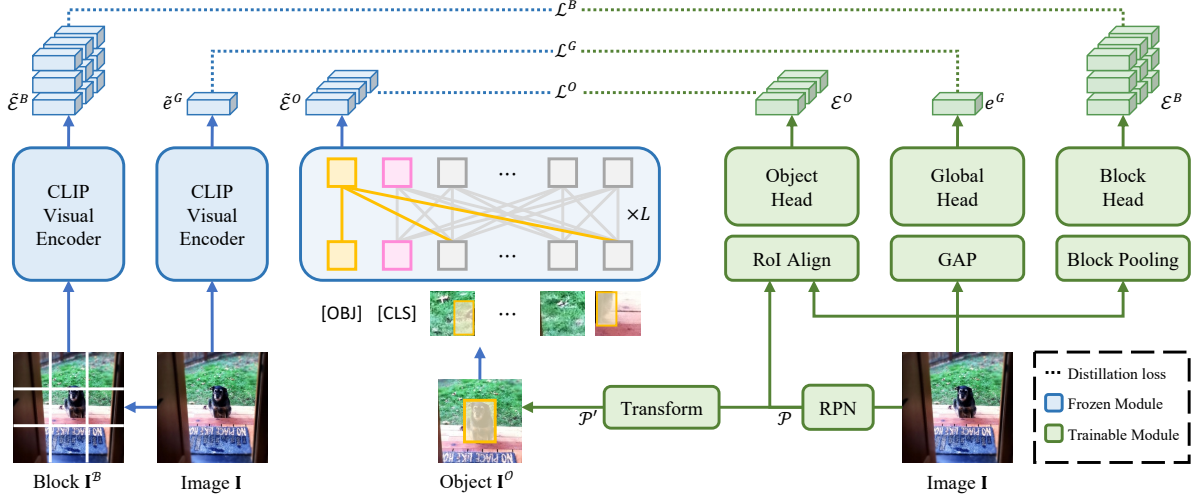


Figure 2. Illustration of our OADP training pipeline. We adopt a pyramid architecture comprising three distillation modules: global, block, and object. Given an image \mathbf{I} , RPN generates proposals \mathcal{P} . For object distillation, RoI Align and Object Head are applied for proposal embeddings \mathcal{E}^O . To extract complete and pure object knowledge from CLIP, we crop the image regions \mathbf{I}^O based on the transformed proposals \mathcal{P}' and feed them to L layers of masked attention, where an extra [OBJ] token (yellow) attends to the patches covered by the original proposal. For global and block distillation, GAP and block pooling are used before the corresponding heads to extract the global and block embeddings (\mathcal{E}^B and e^G). The teacher embeddings $\tilde{\epsilon}^B$ and $\tilde{\epsilon}^G$ are extracted via CLIP from \mathbf{I} and \mathcal{I}^B respectively.

\mathcal{M}^G . The losses of the three modules are denoted as \mathcal{L}^O , \mathcal{L}^B , and \mathcal{L}^G , respectively. The total training loss is:

$$\mathcal{L}^{\text{all}} = \mathcal{L} + w^O \cdot \mathcal{L}^O + w^B \cdot \mathcal{L}^B + w^G \cdot \mathcal{L}^G, \quad (4)$$

where \mathcal{L} is the R-CNN loss as defined in eq. (3); w^O , w^B , and w^G are loss weights.

We follow the inference pipeline of ViLD-ensemble [12] and use \mathcal{M}^O to calibrate $P_C(p, c)$. Similar to the R-CNN head, \mathcal{M}^O extracts the proposal embeddings $\mathcal{E}^O = \{e_p^O\}_{p \in \mathcal{P}} \subset \mathbb{R}^d$ and computes the logits:

$$l^O(p, c) = \frac{e_p^O \cdot t_c}{\|e_p^O\| \cdot \|t_c\|}, \quad (5)$$

$$P_C^O(p, c) = \frac{\exp(l^O(p, c))}{\sum_{c' \in \mathcal{C}} \exp(l^O(p, c'))}, \quad (6)$$

where t_c is the category embedding of c . The calibrated probability $P_C^{\text{cal}}(p, c)$ is:

$$P_C^{\text{cal}}(p, c) = \begin{cases} (P_C(p, c))^\lambda \cdot (P_C^O(p, c))^{(1-\lambda)}, & c \in \mathcal{C}^B \\ (P_C(p, c))^{(1-\lambda)} \cdot (P_C^O(p, c))^\lambda, & c \in \mathcal{C}^N \\ 1 - \sum_{c' \in \mathcal{C}} P_C(p, c'), & c = \text{bg} \end{cases} \quad (7)$$

where λ is set to 2/3. Note that the block and global distillation modules are not used during the inference phase, so the computation cost of our OADP framework is the same as ViLD-ensemble.

4.3. Object Distillation

The object distillation module \mathcal{M}^O aims to transfer the object-level knowledge from CLIP [32] to the detector. For

each proposal $p \in \mathcal{P}$, \mathcal{M}^O motivates the detector to extract a proposal embedding e_p^O that resembles the corresponding embedding \tilde{e}_p^O extracted by the CLIP visual encoder \mathcal{V} :

$$\mathcal{L}^O = \mathcal{L}_1(\mathcal{E}^O, \tilde{\mathcal{E}}^O), \quad (8)$$

where $\tilde{\mathcal{E}}^O = \{\tilde{e}_p^O\}_{p \in \mathcal{P}} \subset \mathbb{R}^d$ denotes the proposal embeddings extracted by \mathcal{V} . Naturally, the quality of $\tilde{\mathcal{E}}^O$ affects the accuracy of $P_C^O(p, c)$ to a large extent. However, current approaches only yield sub-optimal $\tilde{\mathcal{E}}^O$ due to the dilemma between information comprehensiveness and less noise. For example, when non-square proposal regions are directly passed to CLIP, the center crop operation in \mathcal{V} will cut out informative parts of an object, leading to incomplete structural knowledge about the object. On the other hand, if the proposals are squared or enlarged, the proposal regions will contain more ambient contexts, which may corrupt the proposal embeddings.

To acquire more accurate $\tilde{\mathcal{E}}^O$, we propose an Object-Aware Knowledge Extraction (OAKE) module, where the proposals are first transformed and then encoded with a modified version of \mathcal{V} . Given a proposal $p \in \mathcal{P}$, the transformed proposal p' is a square with side length $s = \sqrt{r \times p_h \times p_w}$, where r is a constant scale ratio, p_h and p_w are the height and width of p . The center of p' is initially the same as p but may be translated if p' exceeds the image boundaries. All transformed proposals constitute \mathcal{P}' .

While \mathcal{P}' contributes to the comprehensiveness of object knowledge, it also includes surrounding distractors. To further suppress the contextual noise, we introduce an [OBJ] token into \mathcal{V} and substitute original attention layers with

masked attention layers. The modified version of \mathcal{V} is denoted as \mathcal{V}' . The input of \mathcal{V}' is a set of image regions $\mathbf{I}^O = \{\mathbf{I}_{p'}\}_{p' \in \mathcal{P}'}$, where p' is a transformed proposal and $\mathbf{I}_{p'}$ is the cropped image region of p' . Same as \mathcal{V} , each $\mathbf{I}_{p'}$ is first mapped to a sequence of tokens $\mathbf{X} \in \mathbb{R}^{N_x \times d_x}$, where $\mathbf{X}_{1:N_x-1}$ are the patch tokens of $\mathbf{I}_{p'}$ and \mathbf{X}_{N_x} is the [CLS] token. We then augment \mathbf{X} with an [OBJ] token:

$$\mathbf{X}' = [\mathbf{X}; x_{[\text{OBJ}]}] \in \mathbb{R}^{(N_x+1) \times d_x}. \quad (9)$$

Since the [OBJ] token serves a similar purpose as the [CLS] token, we initialize $x_{[\text{OBJ}]} = \mathbf{X}_{N_x}$. To regulate the interaction between [OBJ] and the other tokens, we construct a mask $m \in \mathbb{R}^{N_x}$, such that:

$$m_i = \begin{cases} \mathbb{1}\{\text{the } i\text{-th patch overlaps with } p\}, & i < N_x \\ 0, & i = N_x \end{cases} \quad (10)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Intuitively, eq. (10) means that [OBJ] only attends to the patch tokens that are covered by the original proposal p . To maintain the original attentions among \mathbf{X} in \mathcal{V} , our attention mask \mathbf{M} is constructed as follows:

$$\mathbf{M} = \begin{bmatrix} \mathbf{1}^{N_x \times N_x} & \mathbf{0}^{N_x} \\ m & 1 \end{bmatrix} \in \{0, 1\}^{(N_x+1) \times (N_x+1)}. \quad (11)$$

Suppose \mathcal{V} has L attention layers (ignoring FFNs), \mathcal{V}' is defined as follows:

$$\mathbf{X}^l = \begin{cases} \sigma(\log \mathbf{M} + \mathbf{Q}^l (\mathbf{K}^l)^\top) \mathbf{V}^l + \mathbf{X}^{l-1}, & 0 < l \leq L \\ \mathbf{X}', & l = 0 \end{cases} \quad (12)$$

where \mathbf{Q}^l , \mathbf{K}^l , and \mathbf{V}^l are linear transformations of \mathbf{X}^{l-1} , σ is Softmax function. Finally, instead of [CLS], we take the output of [OBJ] as the proposal embedding, *i.e.*, $\tilde{e}_p^O = \mathbf{X}_{N_x+1}^L$. Iterating over the proposals \mathcal{P} , we obtain a set of accurate proposal embeddings $\tilde{\mathcal{E}}^O$.

4.4. Global and Block Distillation

While the object distillation module \mathcal{M}^O aligns the proposal embeddings \mathcal{E}^O with the optimized CLIP embeddings $\tilde{\mathcal{E}}^O$, the detector lacks a comprehensive understanding of the relation between different proposals. Therefore, we propose a global distillation module \mathcal{M}^G , which transfers the knowledge of the entire image from the CLIP visual encoder \mathcal{V} to the detector:

$$\begin{aligned} e^G &= f(\text{GAP}(\mathbf{F}_6)) \in \mathbb{R}^d, \\ \tilde{e}^G &= \mathcal{V}(\mathbf{I}) \in \mathbb{R}^d, \\ \mathcal{L}^G &= \mathcal{L}_1(e^G, \tilde{e}^G), \end{aligned}$$

where $f(\cdot)$ is a linear transformation function, $\text{GAP}(\cdot)$ is the global average pooling, and \mathbf{I} is the input image.

Due to the existence of Human Reporting Bias [31], CLIP is prone to ignore non-salient information in an image, *e.g.*, the background or prominent attributes of objects. Such information may be valuable for dense prediction tasks like detection [13]. Thus, we propose a block

distillation module \mathcal{M}^B to complement the missing knowledge in \mathcal{M}^G . We evenly divide the input image into several blocks $\mathcal{B} \subset \mathbb{R}^4$ via a partition function $g(\cdot)$ and denote the corresponding image regions as \mathbf{I}^B . The size of each block is fixed to $R \times R$, where R denotes the input resolution of \mathcal{V} . In this way, the resize and center crop operations in \mathcal{V} will not take effect, thus avoiding information loss when using \mathcal{V} to encode the block embeddings $\tilde{\mathcal{E}}^B = \{\tilde{e}_b^B\}_{b \in \mathcal{B}} \subset \mathbb{R}^d$.

On the student side, we apply block pooling and a block head to extract the block embeddings $\mathcal{E}^B = \{e_b^O\}_{b \in \mathcal{B}} \subset \mathbb{R}^d$. The block pooling is a combination of the block partition function $g(\cdot)$ and RoI Align. The loss of our proposed block distillation is defined as:

$$\mathcal{L}^B = \mathcal{L}_1(\mathcal{E}^B, \tilde{\mathcal{E}}^B). \quad (13)$$

Compared with \mathcal{L}^G , \mathcal{L}^B distills the knowledge of each block with the same weights. Therefore, the ignored information in \mathcal{M}^G can be compensated by \mathcal{M}^B . Note that neither \mathcal{M}^G nor \mathcal{M}^B is used during inference.

4.5. Pseudo Label Generation

To investigate the performance of our OADP framework under the G-OVD benchmark (refer to sec. 3), we propose to generate pseudo labels with our modified \mathcal{V}' . Given a proposal p , we first extract the proposal embedding \tilde{e}_p^O as described in sec. 4.3. Then, the probability of p belonging to category $c \in \mathcal{C}$ is given as:

$$l^{\text{PL}}(p, c) = \frac{\tilde{e}_p^O \cdot t_c}{\|\tilde{e}_p^O\| \cdot \|t_c\|}, \quad (14)$$

$$P_c^{\text{PL}}(p, c) = \frac{\exp(l^{\text{PL}}(p, c))}{\sum_{c' \in \mathcal{C}} \exp(l^{\text{PL}}(p, c'))}, \quad (15)$$

where t_c is the category embedding of c . Since $P_c^{\text{PL}}(p, c)$ does not reflect the localization quality of p [12], we define the confidence score $S_C(p, c)$ as:

$$S_C(p, c) = P_c^{\text{PL}}(p, c)^\gamma \cdot o_p^{(1-\gamma)}, \quad (16)$$

where \mathcal{C} includes both base and novel categories, $o_p \in [0, 1]$ is the objectness score of p , and γ is a constant balancing factor. $S_C(p, c)$ reflects the probability that p precisely locates an instance of category c . Finally, we apply class-wise NMS on the novel categories to obtain the pseudo labels.

Note that the Softmax operation in eq. (15) is performed over all categories \mathcal{C} , even though the pseudo labels do not include instances on base categories. Such a design effectively suppresses false positives in the pseudo labels.

The proposals to be labeled are extracted via an RPN model pretrained on \mathcal{D}^T , which only contains annotations of the base categories. ViLD [12] demonstrates that the generalization ability of \mathcal{P} is strong enough to recall most objects of the novel categories.

Benchmark	Method	mAP ₅₀ ^N	mAP ₅₀ ^B	mAP ₅₀
ZSD	SB [1]	0.3	29.2	24.9
	DELO [56]	3.4	13.8	13.0
	PL [34]	4.1	35.9	27.9
V-OVD	ViLD [12]	27.6	59.5	51.3
	RegionCLIP* [53]	14.2	52.8	42.7
	OADP (Ours)	30.0	53.3	47.2
C-OVD	OVR-CNN [49]	22.8	46.0	39.9
	HierKD [28]	20.3	51.3	43.2
	RegionCLIP [53]	26.8	54.8	47.5
	LocOV [3]	28.6	51.3	45.7
	PB-OVD [11]	29.1	44.4	40.4
G-OVD	OV-DETR [48]	29.4	61.0	52.7
	VL-PLM [52]	32.3	54.0	48.3
	OADP (Ours)	35.6	55.8	50.5
WSD	WSDNN [2]	19.7	19.6	19.6
	Cap2Det [47]	20.3	20.1	20.1
WS-OVD	Detic [55]	27.8	47.1	45.0

Table 2. Comparison with other state-of-the-art methods on the OV-COCO dataset. Methods are grouped by the benchmark they use. “ZSD” and “WSD” stand for Zero-Shot Detection and Weakly Supervised Detection. “V-OVD”, “C-OVD”, “G-OVD”, and “WS-OVD” are introduced in sec. 3. “RegionCLIP*” indicates a model without refinement using COCO-Captions.

Method	Object Detection				Instance Segmentation			
	AP _r	AP _c	AP _f	AP	AP _r	AP _c	AP _f	AP
ViLD [12]	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
DetPro [8]	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OV-DETR [48]	-	-	-	-	17.4	25.0	32.5	26.6
OADP (Ours)	21.9	28.4	32.0	28.7	21.7	26.3	29.0	26.6

Table 3. Comparison with other state-of-the-art methods on the OV-LVIS dataset.

5. Experiments

In this section, we first introduce the detailed experiment setup, including the datasets, evaluation metrics, and implementation details. We then evaluate the performance of our proposed OADP framework and analyze the results compared to the state-of-the-art approaches.

5.1. Datasets

Experiments are mainly conducted under the open-vocabulary COCO (OV-COCO) setting [49], where the MSCOCO 2017 dataset [27] is manually divided into 48 base categories and 17 novel categories. The training dataset contains 107,761 images, and the validation dataset contains 4,836 images. We report the mAP₅₀^N, mAP₅₀^B, and mAP₅₀ metrics, *i.e.*, the mAP at IoU threshold 0.5 for novel, base, and all categories. mAP₅₀^N is the main metric. Some experiments are conducted under the open-vocabulary LVIS (OV-LVIS) [12] setting, where the 337 rare categories in

LVIS [14] are treated as novel categories, and the other 866 are base categories. Metrics for the OV-LVIS setting are AP_r, AP_c, AP_f, and AP, *i.e.*, the mAP for rare (novel), common, frequent, and all categories. Both object detection and instance segmentation metrics are reported.

5.2. Implementation Details

Training is conducted on 8 V-100 GPUs with batch size 16 in total. We use stochastic gradient descent (SGD) optimizer with 0.02 initial learning rate, 0.9 momentum, and 2.5×10^{-5} weight decay. The student backbone is ResNet-50 [16]. Following DetPro [8], we adopt the ViT-B/32 CLIP [32] as the teacher and initialize the student backbone using SoCo [41]. The loss weights w^O , w^B , and w^G are set to 0.5, 0.25, and 0.25 respectively. Under the OV-COCO setting, we train the detector for 40,000 iterations. At the 32,000th iteration, the learning rate is divided by 10. For OV-LVIS, we use 2x (24 epochs) training schedule, where the learning rate is divided by 10 at the 16th and 22th epochs.

5.3. Main Results

We compare our OADP framework with the other state-of-the-art OVD methods. As described in sec. 3, we categorize existing OVD methods by the benchmark they belong to. For completeness, we include two related benchmarks: Zero-Shot Detection (ZSD) [1, 18, 24, 35, 51] and Weakly Supervised Detection (WSD) [2, 47].

Our method mainly focuses on the V-OVD benchmark and the G-OVD benchmark. As shown in tab. 2, our OADP framework achieves 30.0 mAP₅₀^N on the V-OVD benchmark. RegionCLIP* [53] uses CLIP [32] as the pretrained weight, thus adhering to the V-OVD benchmark. Some V-OVD methods [4, 20, 22, 29, 45] are not included because they rely on large-scale detection and image-text datasets and cannot be compared fairly. While C-OVD is not our primary concern, we include the corresponding methods for reference and report their performance when only COCO Captions [6] is available. Under such constraint, the performance of the caption-based methods is relatively lower than the V-OVD methods, even if additional caption data is used. Moreover, our OADP framework is perpendicular to the caption-based methods and has the potential to achieve higher performance using captions.

For the G-OVD benchmark, we generate pseudo labels for novel categories as described in sec. 4.5. The pseudo labels are then merged with the instance-level annotations for base categories. Training our OADP framework on the mixed dataset yields 35.6 mAP₅₀^N, surpassing the previous SOTA method VL-PLM [52] by 3.3 mAP₅₀^N. Since Prompt-Det [10] relies on an external dataset (LAION-400M [38]) and uses smaller image size (640 × 640), the result 26.6 mAP₅₀^N is not listed in tab. 2 for fairness.

Global Distillation	Block Distillation	Object Distillation	mAP	Novel mAP ₅₀	mAP ₇₅	mAP	Base mAP ₅₀	mAP ₇₅	mAP	All mAP ₅₀	mAP ₇₅
			13.32	24.99	12.35	31.87	50.29	34.03	27.02	43.67	28.36
✓			13.51	25.72	12.36	32.82	51.89	35.31	27.77	45.04	29.31
	✓		14.57	27.25	13.17	34.45	53.60	37.20	29.25	46.71	31.06
		✓	15.49	27.23	15.25	35.99	55.96	38.57	30.63	48.45	32.47
✓	✓		13.50	26.49	12.50	32.19	51.25	33.94	27.30	44.78	28.33
✓		✓	15.47	28.80	14.62	34.08	54.29	36.28	29.21	47.62	30.61
	✓	✓	15.92	29.01	15.64	35.30	55.45	37.88	30.23	48.53	32.06
✓	✓	✓	16.21	29.95	15.47	33.33	53.26	35.47	28.85	47.17	30.24

Table 4. Ablation study of the Global, Block, and Object Distillation modules in the OADP framework. The baseline is our re-implemented ViLD-ensemble model.

Tab. 3 shows the comparison between our method and the other state-of-the-art methods on the OV-LVIS dataset. Most of the methods under the OV-LVIS setting adhere to the V-OVD benchmark, so we conduct experiments on the V-OVD benchmark only. For the object detection task, our OADP framework achieves 21.9 AP_r, surpassing DetPro [8] by 1.1 AP_r. We also report the performance of the instance segmentation task, which achieves 21.7 AP_r and is 1.9 AP_r higher than the previous SOTA method.

5.4. Ablation Study

We conduct ablation studies on the OV-COCO dataset to evaluate the effectiveness of each component in our proposed OADP framework.

OADP. Tab. 4 shows the effectiveness of the distillation modules in our OADP framework. The first row is our re-implemented ViLD-ensemble [12]. Due to the expensive training cost of ViLD, the performance 24.99 mAP₅₀^N is far below the official 27.60 mAP₅₀^N. Nevertheless, with our proposed distillation pyramid, we can eventually surpass ViLD. The 2nd to 4th row in tab. 4 adds \mathcal{M}^G , \mathcal{M}^B , and \mathcal{M}^O to the baseline respectively. The global distillation module brings a 0.73 mAP₅₀^N gain, while the other two bring 2.26 mAP₅₀^N and 2.24 mAP₅₀^N gain. Note that by adding \mathcal{M}^O to the baseline, we remove the original image head in ViLD-ensemble. Therefore, the 2.24 mAP₅₀^N gain results from the OAKE module instead of the distillation operation. The 5th row in tab. 4 adds \mathcal{M}^G and \mathcal{M}^B together. While the 26.49 performance is higher than sole \mathcal{M}^G , it is slightly lower than the 27.25 mAP₅₀^N of \mathcal{M}^B . However, along with the object distillation module \mathcal{M}^O , \mathcal{M}^G , and \mathcal{M}^B achieves 28.80 mAP₅₀^N and 29.01 mAP₅₀^N, suggesting that \mathcal{M}^G and \mathcal{M}^B have a similar function in transferring the global scene knowledge from CLIP to the detector. Using all three modules together, we achieve 29.95 mAP₅₀^N.

OAKE. We demonstrate the effectiveness of our OAKE module in tab. 5. Given the ground truth bounding boxes, we use different strategies to crop their image regions. (1) Baseline: 1× crop; (2) ViLD*: 1× and 1.5× crop; (3)

Method	Macro Precision		Weighted Precision	
	w/o OAKE	w/ OAKE	w/o OAKE	w/ OAKE
Baseline	58.08	-	62.04	-
ViLD*	63.36	-	65.91	-
MBS	61.70	63.83	64.81	65.82
Fixed	49.07	64.53	51.49	69.75
Adaptive	51.64	66.09	55.85	68.68

Table 5. Ablation study of OAKE module. “ViLD*” indicates our re-implementation of multi-scale region embedding. “MBS”, “Fixed”, and “Adaptive” are three transforming strategies.

MBS: the minimum bounding square of the original bounding box; (4) Fixed: 224 × 224 bounding square; (5) Adaptive: adaptively enlarge the bounding square. For the above strategies, we use CLIP to directly extract embeddings for their image regions (“w/o mask”). Alternatively, we can utilize the modified CLIP visual encoder mentioned in sec. 4.3 (“w/ mask”). Finally, we classify these embeddings by calculating their similarity with category embeddings. To evaluate the performance, we compute “Macro Precision” (precision for each category independently with equal weights) and “Weighted Precision” (weights depending on the number of bounding boxes in each class).

As shown in the 1st and 3rd columns, “Fixed” and “Adaptive” strategies bring performance drops as they crop a larger bounding square compared to other strategies (*e.g.*, “MBS”) which may introduce additional surrounding distractors that confuse the CLIP visual encoder. However, with our object-aware CLIP visual encoder, the performances of the two strategies boost significantly and suppress others. It validates that the CLIP visual encoder can focus on the proposal object with our mask attention mechanism to extract accurate knowledge.

Pseudo Label. We follow VL-PLM [52] to adopt the COCO-ZS setting for our ablation studies of the pseudo label. Both mAP and the average per-image number of PLs (#PL) on novel categories are used as metrics to evaluate the quality of the pseudo label. The baseline method directly uses [CLS] token of CLIP visual encoder to extract

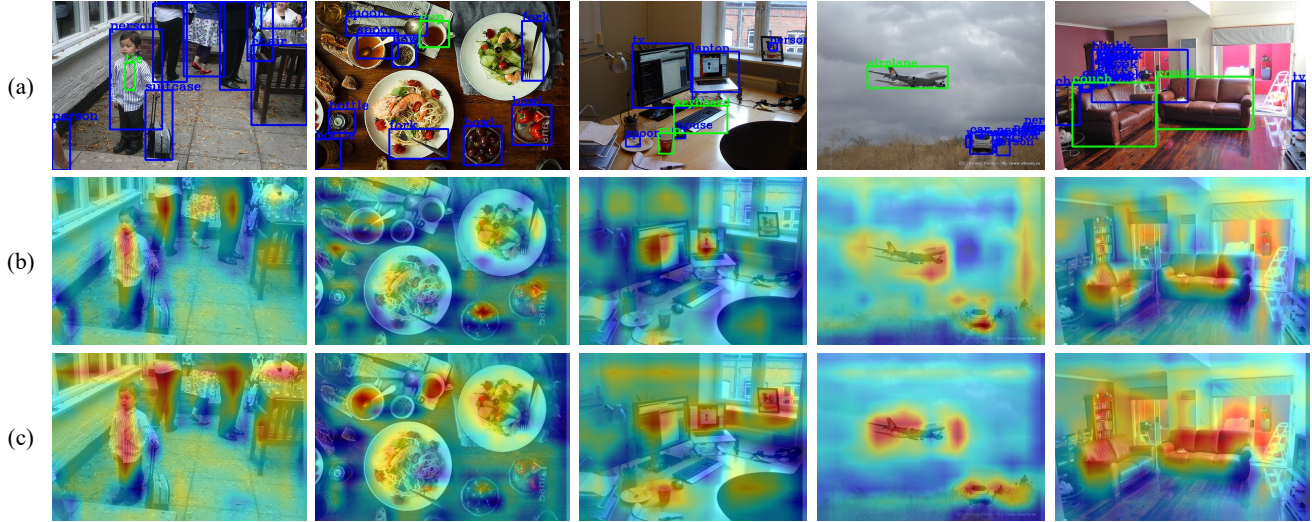


Figure 3. Visualization of activation patterns from different detectors. (a) Pseudo labels (green) and ground truth annotations (blue) for each image. (b) The baseline detector. (c) Our OADP. The intensity of the feature response increases from blue to red.

Method	mAP ^{PL}	mAP ₅₀ ^{PL}	mAP ₇₅ ^{PL}	#PL
Baseline	5.0	12.1	3.1	100
ViLD*	17.8	28.1	18.6	100
Ours	19.0	29.9	19.9	100
Baseline	3.9	9.3	2.5	6.53
VL-PLM [52]	-	25.3	-	4.26
Ours	17.4	26.5	18.6	4.14

Table 6. Ablation study of pseudo labels. “#PL” is the number of pseudo labels per image.

proposal embeddings from the original proposal, and it relies merely on the classification score to sort proposals. The poor detection accuracy in 1st and 4th rows of tab. 6 show that without the objectness score, the baseline method can not accurately localize objects. We re-implement multi-scale region embedding of ViLD [12] with a geometric mean of CLIP classification score and objectiveness score, *i.e.*, “ViLD*”. We adopt an adaptive transform strategy for proposals and regard the output of [OBJ] as proposal embedding. The score fusion strategy is described in sec. 4.5, where γ is 0.3. Our method achieves the highest mAP when the number of PLs is sufficient. VL-PLM [52] adopts a multi-scale region embedding method similar to ViLD [12] except for an arithmetic mean of classification score and objectiveness score. When the pseudo labels are filtered with a higher confidence threshold, our method still has a significant advantage compared to VL-PLM [52] (26.50 mAP₅₀^{PL} compared to 25.30 mAP₅₀^{PL}).

5.5. Visualization

Fig. 3 (a) visualizes our generated PLs on novel categories in green with ground truth boxes of base categories in

blue. We try our best to ensure the accuracy of PLs as much as possible to be fewer and more precise. These green PLs demonstrate that our proposal embeddings can clearly distinguish novel objects from base ones. Correspondingly, we also show activation maps from baseline (b) and our detector (c) in fig. 3. Taking the 2nd column as an example, the activation map of our detector accurately highlights more area of novel objects, *i.e.* “cup”, with our distillation pyramid mechanism. Therefore, the backbone of OADP generates more informative feature maps, which further help detect novel objects.

6. Conclusion

In this paper, we reconsider the way of knowledge extraction and knowledge transfer in existing KD-based OVD methods and propose an Object-Aware Distillation Pyramid (OADP) framework. To preserve complete and purified object representation in proposals during knowledge extraction, we propose an Object-Aware Knowledge Extraction (OAKE) module to adaptively transform proposals and extract precise object knowledge. A Distillation Pyramid (DP) mechanism is proposed to transfer contextual knowledge about the relation of different objects for better scene understanding. Experiments show that our OADP outperforms previous methods on two popular OVD benchmarks.

Acknowledgement. This research is supported by the National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (NO. 62122010), Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F02022F02, Key Research and Development Program of Zhejiang Province under Grant 2022C01082.

References

- [1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-Shot Object Detection. In *ECCV*, 2018. 3, 6
- [2] Hakan Bilen and Andrea Vedaldi. Weakly Supervised Deep Detection Networks. In *CVPR*, 2016. 3, 6
- [3] Maria A. Bravo, Sudhanshu Mittal, and Thomas Brox. Localized Vision-Language Matching for Open-vocabulary Object Detection. In *GCPR*, 2022. 3, 6
- [4] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-DETR: A Versatile Architecture for Instance-wise Vision-Language Tasks. In *ECCV*, 2022. 2, 6
- [5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, 2017. 2
- [6] Xinlei Chen, Hao Fang, Tsung-yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 6
- [7] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General Instance Distillation for Object Detection. In *CVPR*, 2021. 2
- [8] Yu Du, Fangyun Wei, Zihe Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. In *CVPR*, 2022. 2, 6, 7
- [9] Zhixing Du, Rui Zhang, Ming Chang, Xishan Zhang, Shaoli Liu, Tianshi Chen, and Yunji Chen. Distilling Object Detectors with Feature Richness. In *NeurIPS*, 2021. 2
- [10] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. PromptDet: Expand Your Detector Vocabulary with Uncurated Images. In *ECCV*, 2022. 2, 3, 6
- [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 *ECCV*, 2022. 2, 3, 6
- [12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, 2022. 1, 2, 4, 5, 6, 7, 8
- [13] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling Object Detectors via Decoupled Features. In *CVPR*, 2021. 2, 5
- [14] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*, 2019. 2, 6
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. *TPAMI*, 42(2), 2020. 1
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6
- [17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *NeurIPS*, 2014. 1, 2, 3
- [18] Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. Robust Region Feature Synthesizer for Zero-Shot Object Detection. In *CVPR*, 2022. 6
- [19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, 2021. 1, 2
- [20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *ICCV*, 2021. 2, 6
- [21] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-Conditional Knowledge Distillation for Object Detection. In *NeurIPS*, 2021. 2
- [22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-training. In *CVPR*, 2022. 2, 6
- [23] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking Very Efficient Network for Object Detection. In *CVPR*, 2017. 2
- [24] Zihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-Shot Object Detection with Textual Descriptions. *AAAI*, 33(1), 2019. 6
- [25] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*, pages 20091–20100, 2022. 1
- [26] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017. 3

- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2, 6
- [28] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation. In *CVPR*, 2022. 1, 2, 3, 6
- [29] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple Open-Vocabulary Object Detection with Vision Transformers. In *ECCV*, 2022. 1, 2, 6
- [30] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant. *AAAI*, 34(04), 2020. 2
- [31] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *CVPR*, 2016. 5
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 2021. 1, 2, 3, 4, 6
- [33] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive Learning for Zero-Shot Object Detection. In *ICCV*, 2019. 3
- [34] Shafin Rahman, Salman Khan, and Nick Barnes. Improved Visual-Semantic Alignment for Zero-Shot Object Detection. *AAAI*, 34(07), 2020. 6
- [35] Shafin Rahman, Salman H. Khan, and Fatih Porikli. Zero-Shot Object Detection: Joint Recognition and Localization of Novel Concepts. *IJCV*, 128(12), 2020. 6
- [36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *TPAMI*, 39(6), 2017. 3
- [37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. In *ICLR*, 2015. 2
- [38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6
- [39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018. 3
- [40] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling Object Detectors With Fine-Grained Feature Imitation. In *CVPR*, 2019. 2
- [41] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning Pretraining for Detection via Object-Level Contrastive Learning. In *NeurIPS*, volume 27, 2021. 6
- [42] Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-Guided Contrastive Network for Zero-Shot Object detection. *TPAMI*, 2022. 3
- [43] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and Global Knowledge Distillation for Detectors. In *CVPR*, 2022. 2
- [44] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked Generative Distillation. In *ECCV*, 2022. 2
- [45] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection. In *NeurIPS*, 2022. 2, 6
- [46] Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-DetKD: Towards General Distillation Framework for Object Detectors via Contrastive and Semantic-guided Feature Imitation. In *ICCV*, 2021. 2
- [47] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2Det: Learning to Amplify Weak Caption Supervision for Object Detection. In *ICCV*, 2019. 3, 6
- [48] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-Vocabulary DETR with Conditional Matching. *ECCV*, 2022. 1, 2, 3, 6
- [49] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *CVPR*, 2021. 1, 2, 3, 6
- [50] Peizhen Zhang, Zijian Kang, Tong Yang, Xiangyu Zhang, Nanning Zheng, and Jian Sun. LGD: Label-Guided Self-Distillation for Object Detection. *AAAI*, 36(3), 2022. 2

- [51] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang. GT-Net: Generative Transfer Network for Zero-Shot Object Detection. *AAAI*, 34(07), 2020. 6
- [52] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B. G, Anastasis Sathopoulos, Manmohan Chandraker, and Dimitris Metaxas. Exploiting Unlabeled Data with Vision and Language Models for Object Detection. In *ECCV*, 2022. 2, 3, 6, 7, 8
- [53] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based Language-Image Pretraining. In *CVPR*, 2022. 1, 2, 6
- [54] Chunting Zhou, Graham Neubig, and Jiatao Gu. Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors. In *ICLR*, 2021. 2
- [55] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*, 2022. 2, 3, 6
- [56] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't Even Look Once: Synthesizing Features for Zero-Shot Detection. In *CVPR*, 2020. 6
- [57] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Zero Shot Detection. *TCSVT*, 30(4), 2020. 3