EdaDet: Open-Vocabulary Object Detection Using Early Dense Alignment

Cheng Shi, Sibei Yang[†]
School of Information Science and Technology, ShanghaiTech University
{shicheng2022,yangsb}@shanghaitech.edu.cn
Project page: https://chengshiest.github.io/edadet

Abstract

Vision-language models such as CLIP have boosted the performance of open-vocabulary object detection, where the detector is trained on base categories but required to detect novel categories. Existing methods leverage CLIP's strong zero-shot recognition ability to align object-level embeddings with textual embeddings of categories. However, we observe that using CLIP for object-level alignment results in overfitting to base categories, i.e., novel categories most similar to base categories have particularly poor performance as they are recognized as similar base categories. In this paper, we first identify that the loss of critical fine-grained local image semantics hinders existing methods from attaining strong base-to-novel generalization. Then, we propose Early Dense Alignment (EDA) to bridge the gap between generalizable local semantics and object-level prediction. In EDA, we use object-level supervision to learn the dense-level rather than object-level alignment to maintain the local fine-grained semantics. Extensive experiments demonstrate our superior performance to competing approaches under the same strict setting and without using external training resources, i.e., improving the $+8.4\%$ novel box AP50 on COCO and $+3.9\%$ rare mask AP on LVIS.

1. Introduction

Open-vocabulary object detection aims to localize and recognize objects of both *base categories* and *novel categories* when only the training labels on base categories are available. Beyond focusing on object detection on a closed set of categories [20, 3, 39, 30, 43], open-vocabulary detection requires generalizing well from base to all novel categories without annotations for each novel category.

One straightforward idea is to generate pseudo-proposals relevant to novel categories and train detectors with base and novel categories (see Figure 1b), adopted by [54, 28, 14, 38, 2]. They usually first extract concepts relevant to novel categories and then generate proposals of novel concepts from extra training resources. Among them, some works [54, 14, 38] follow the weak open-vocabulary setting [48], where the class names of novel categories directly corresponding to novel concepts are available in the training phase. Alternatively, some works [50, 28] extract novel concepts from captions or image-text pairs. Although these approaches have improved the performance of detecting novel categories, the need for additional training resources that heavily overlap with or are relevant to novel categories would limit them to practical applications.

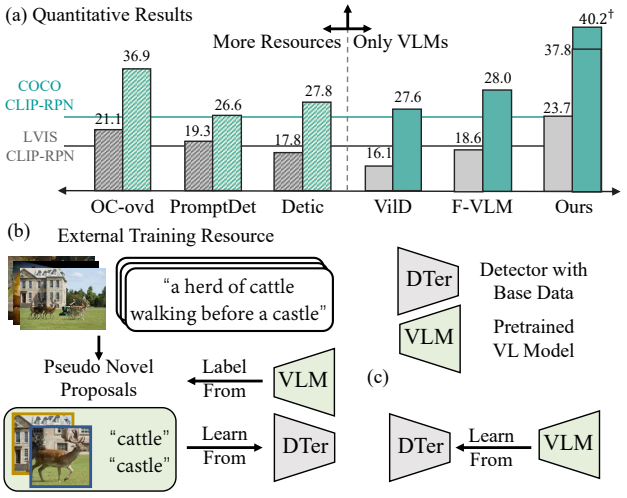


Figure 1: Different approaches to building an open-vocabulary detector: (a) their performance comparison. †: with self-training. (b) generate pseudo "novel" proposals from extra training resources and VLMs, or (c) generalize from VLMs

Recently, contrastive pre-training of vision-language models (VLMs) like CLIP [34] and ALIGN [22] have shown strong open-vocabulary image recognition ability. Some open-vocabulary detection works [16, 13, 50, 26, 47] explore utilizing VLMs to learn transferable object representations. Figure 2b shows a high-level abstraction of these open-vocabulary detection frameworks. Although they plug well-designed methods to close the gap between visual representation learning for objects and images, they

[†]Corresponding author

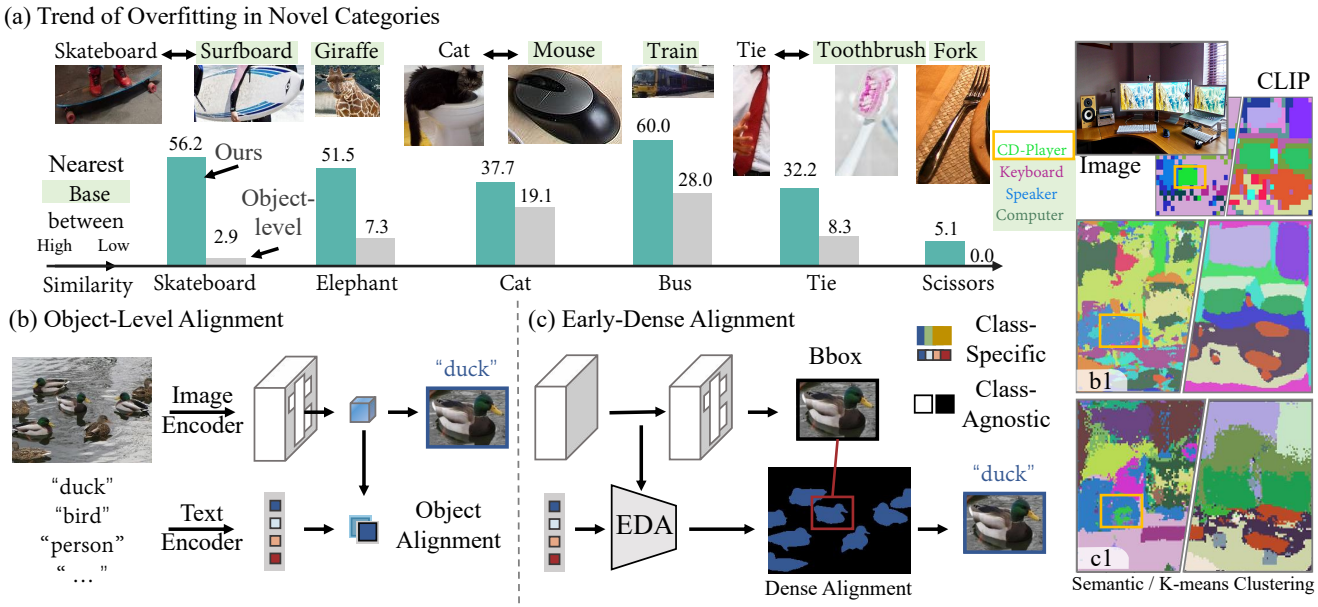


Figure 2: The comparison between Object-level Alignment and our Early Dense Alignment (Eda) on **(b)-(c)** architectures, **(b1)-(c1)** local image semantics and clustering results, and **(a)** box AP of novel categories similar to base categories. We list six novel categories most similar to base categories by calculating the average similarity between the randomly sampled thousands of novel objects' visual features and base categories' text embeddings. Our Eda: (1) successfully recognizes the fine-grained *novel CD-player* that is predicted to *base speaker* by object-level alignment; (2) better groups local image semantics into object regions compared with CLIP; (3) achieves a much higher novel box AP for predicting novel objects similar to base objects, showing that Eda can distinguish fine-grained details of similar novel and base categories. In contrast, object-level alignment overfits base categories.

only achieve comparable or slightly better performance than CLIP-RPN[1] on novel categories (right part of Figure 1a). We also follow the line of works that aim to generalize VLMs for object detection without generating pseudo-proposals relevant to novel categories (see Figure 1c). And we explore *how better to utilize VLMs for base-to-novel generalization in open-vocabulary object detection.*

In this paper, we start by discovering and analyzing the respective advantages of VLMs and existing open-vocabulary detection frameworks for object detection. First, we observe that *VLMs can predict local image semantics for novel categories while existing frameworks are easier to overfit base categories.* As shown in the "Semantic" of Figure 2 (CLIP), CLIP successfully recognizes the novel local regions of the "CD player", while the existing framework classifies the novel "CD player" as the "speaker" in base categories. The reason is that VLMs may have seen fine-grained image-text pairs describing local semantics during training. In contrast, existing frameworks directly align the object representations to the classifier of base categories, which loses the fine-grained details that distinguish novel objects from their similar base objects. Without fine-grained details, the object-level representations of novel objects and their similar base objects are similar, resulting in them being classified into base categories. Figure 2a (marked in grey) shows that the object-level alignment's prediction accuracy for novel objects similar to base objects is much lower. Therefore, we propose to *avoid direct object-level alignment and fully utilize VLMs' ability to distinguish fine-grained details for similar novel and base objects to preserve the recognition ability of novel categories.*

Second, we observe that *the existing framework can better group local image semantics into object regions than VLMs.* The reason is that its object-level supervision for object representations generated from local semantics improves local semantic consistency to objects. As shown in the "K-means clustering" of Figure 2-b1 and 2 (CLIP), the existing framework groups the "'keyboard" well (marked in yellow), while the corresponding two separate "keyboard" regions in CLIP's clustering map are mixed and indistinguishable. *Therefore, we propose to adopt object-level supervision for the dense alignment of local image semantics.*

Based on the above discoveries, we propose a simple but effective solution, named early dense[2] alignment (Eda), to combine the strengths of VLMs and the existing frameworks. To avoid overfitting to base categories caused by

[1]CLIP-RPN baseline simply utilizes CLIP to classify cropped proposals generated by region proposal network (RPN) trained on base categories.

[2]"early" means the use of features in the early stages of the backbone, and "dense" means per-pixel alignment to text for obtaining object-level prediction (Note that only object-level annotations are used).

object-level alignment, Eda directly predicts object categories from local image semantics to fully distinguish the fine-grained details of similar base objects and novel objects. To maintain the local semantics consistent for better grouping and localization, we use object-level supervision to learn the dense-level alignment. As shown in Figure 2c, Eda first aligns local image semantics to the CLIP's semantic space early and then predicts object-level labels based on the dense probabilities to categories. Our Eda enables dense-level alignment for local image semantics, which is much more generalizable than late object-level alignment to novel categories. Meanwhile, it can better group local semantics to object regions (see Figure 2c1). Also, Figure 2a (mark in green) shows that our Eda significantly improves the prediction accuracy for novel categories similar to base categories.

Finally, we propose EdaDet, a simple open-vocabulary detection framework by leveraging our early dense alignment (Eda). For object localization, we follow existing works [38, 54, 16, 26] to learn class-agnostic object proposals. To ensure an efficient end-to-end localization and recognition framework, we adopt a query-based proposal generation method like DETR [3] but revise it to be class-agnostic. For open-vocabulary recognition, we apply our generalizable Eda to predict the categories of class-agnostic proposals. In addition, for better generalization, EdaDet deeply decouples the object localization and recognition by separating the open-vocabulary classification branch from the class-agnostic proposal generation branch at a more shallow layer of the decoder.

To evaluate the effectiveness of our EdaDet, we conduct experiments on LVIS [17], COCO [29], and Objects365 [42] benchmarks. In summary, our main contributions are as follows,

- We propose a novel and effective early dense alignment (Eda) for base-to-novel generalization in object detection without knowing the class names of novel categories and using extra training resources.
- We propose an end-to-end EdaDet framework, which deeply decouples the object localization and recognition by separating classification from the recognition at a more shallow layer of the decoder.
- Despite being simple, EdaDet achieves strong quantitative results, outperforming state-of-the-art methods with the strict setting on COCO and LVIS by 5.8% box AP50 and 2.0% mask AP on novel categories respectively. Moreover, EdaDet shows striking cross-dataset transferable capability.
- EdaDet shows impressive qualitative predictions on local image semantics and demonstrates efficient and effective performance improvement when scaling the model size thanks to our generalizable Eda and deeply decoupled detection framework.

2. Related Work

Transferable Representation Learning explores learning transferable representations from source tasks with large-scale data in the pre-training stage and then adapt the representations to a variety of target downstream tasks [1, 23]. Based on whether the data used in the pre-training is labeled or not, the pre-training can be divided into supervised and unsupervised. Supervised pre-training [19, 45] is commonly employed in computer vision community. For example, image classification on ImageNet [10] is often used as the pre-training task for the downstream visual recognition tasks [20, 3, 39, 30, 43, 40, 6, 32]. In contrast, unsupervised pre-training proposes self-supervised tasks for pre-training on unlabeled data, including the generative learning [35, 36, 18, 11] and contrastive learning [8, 4, 7, 9, 34]. The unsupervised pre-training enables learning generally transferable knowledge from many large-scale unlabeled website data [41, 5]. Among them, the contrastive vision-language models (VLMs) train a dual-modality encoder on large-scale image-text pairs to learn transferable visual representations with text supervision. In this paper, we aim to transfer the general knowledge to recognize open-vocabulary objects in images and therefore select the contrastive VLMs that can align pairs of image and text as our source models.

Visual Recognition from Generalizable VLMs. The contrastive VLMs such as CLIP [34] and ALIGN [22] pre-trained on large-scale image-text pairs have shown transferability to various visual recognition tasks, such as image classification [52, 53], semantic segmentation [37, 27], and object detection [16, 13, 50, 38]. With a handcraft prompt, VLMs can extract the category's text embeddings as the classifier for images. For image classification, zero-shot VLMs have already demonstrated strong zero-shot classification performance on various image classification tasks. CoOp [53] and CoCoOp [52] further model the context words of a prompt with learnable vectors, thereby eliminating the need for handcrafted design. Unlike image classification methods that can directly use the CLIP's image encoder, semantic segmentation approaches such as DenseCLIP [37] and LSeg [27] learn a segmenter from scratch by using the CLIP's text encoder as a frozen classifier for dense features. MASKCLIP+ [51] achieves better segmentation results on unseen categories by generating pseudo labeling of unseen categories and performing self-training with pseudo labels. In open-vocabulary detection, previous works [16, 13, 50, 38] also explore distilling knowledge from CLIP to detectors. However, they can only achieve a comparable performance of novel categories with zero-shot CLIP predicted on cropped class-agnostic proposals.

Therefore, we explore fully utilize CLIP's ability to distinguish fine-grained details for similar novel and base objects to preserve the recognition ability of novel categories.

Open-Vocabulary Object Detection. Depending on the availability of novel categories' vocabulary during training, open-vocabulary detection is divided into strict [16] (unknown) and weak [54] (known) settings. Under the weak open-vocabulary detection setting, one basic solution is to generate pseudo-proposals of novel categories. Previous works [54, 38, 14] all leverage additional image-level data to generate novel pseudo-proposals and train detectors with both base and novel categories. For the strict setting, recent works [47, 26, 13, 16] mainly focus on generalizing CLIP to detect objects of novel categories. They all learn class-agnostic proposals and then classify proposals by using category names' text embedding extracted from CLIP's text encoder as the classification weights but differ in adopting different strategies to improve the generalization performance. For example, DetPro [13] designs a learnable prompt token instead of a hand-craft prompt to achieve a better generalization performance. RegionCLIP [50] develops a region-text pre-training strategy to obtain fine-grained alignment between image regions and textual concepts, which is more suitable for object-level prediction. Furthermore, some methods such as ZSD-YOLO [46] and ViLD [16] align the proposal representations to that extracted from the CLIP's image encoder.

We follow the strict open-vocabulary detection setting. Instead of aligning object representations to CLIP's semantic space and relying on object-level alignment like previous works, we propose to align local image semantics to CLIP's space at the dense level to mitigate the overfitting issue.

3. Method

Our study is a first attempt to recognize open-vocabulary objects by utilizing dense-level alignment of local image semantics to CLIP's semantic space. We start with a brief introduction of CLIP (see Section 3.1) and a simple but necessary modification to its image encoder for dense-level prediction. Next, we introduce our overall detection framework (EdaDet), an end-to-end query-based object detection architecture, in Section 3.2. Moreover, the class-agnostic proposal generation is described in this section. Finally, we present the open-vocabulary object classification implemented by our early dense alignment (Eda) in Section 3.3.

3.1. Preliminary

CLIP [34] is trained on large-scale image-text pairs by image-level contrastive learning. Specifically, CLIP has a pair of image encoder $f(\cdot)$ and text encoder $g(\cdot)$. The image encoder $f(\cdot)$ can be presented into two parts: a visual backbone (*e.g.*, ResNet [21] or ViT [12]) denoted as $f_{\text{Backbone}}(\cdot)$ and the global feature aggregation layer (*e.g.*, the last global attention pooling layer for ResNet) denoted as $f_{\text{G-Pooling}}(\cdot)$. The global attention pooling layer $f_{\text{G-Pooling}}(\cdot)$ is a single layer of multi-head attention [44] that takes the globally average-pooled feature as a class token [cls] and concatenates it with patch tokens [patches] flattened from outputs of $f_{\text{Backbone}}(\cdot)$ as inputs. Given a text $\mathcal{T}$ and an image $\mathcal{I}$, CLIP computes the similarity $\mathcal{S}$ between $\mathcal{T}$ and $\mathcal{I}$ by:

$$\mathcal{S}_{\text{CLIP}} = \cos(f_{\text{G-Pooling}}(f_{\text{Backbone}}(\mathcal{I}))_{[\text{cls}]}, g(\mathcal{T})), \quad (1)$$

where $\cos(\cdot,\cdot)$ represents the cosine similarity, and class token's feature $f_{\text{G-Pooling}}(f_{\text{Backbone}}(\mathcal{I}))_{[\text{cls}]}$ is the global feature.

Modification of CLIP for Dense Prediction. As CLIP models for image-level prediction, it is not trivial to extract local patch prediction (*i.e.*, the actual similarities between patches $f_{\text{Backbone}}(\mathcal{I})$ and text $\mathcal{T}$) from CLIP. Similar to MaskCLIP [51] to reformulate the value-embedding layer of $f_{\text{G-Pooling}}(\cdot)$, we retain the global pooling layer but additionally conditioned on a diagonal mask $\mathcal{M}$ that prevents information exchange between patches. The modified formulation for dense prediction is expressed as,

$$\mathcal{S}_{\text{CLIP}} = \cos(f_{\text{G-Pooling}}(f_{\text{Backbone}}(\mathcal{I})|\mathcal{M})_{[\text{patches}]}, g(\mathcal{T})). \quad (2)$$

We adopt the modified pooling layer to produce the fine-grained dense prediction as shown in Figure 2 (CLIP).

3.2. Open-Vocabulary Object Detector (EdaDet)

Problem Setup. We share the same strict problem setup following previous works [16, 26, 47, 13, 31, 28]. Given an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, the detector predicts a set of bounding boxes with categories. It is trained with detection annotations of base categories $\mathbb{C}_{\text{train}}$ but needs to detect objects of both base and novel categories $\mathbb{C}_{\text{target}}$ where $\mathbb{C}_{\text{target}} \neq \mathbb{C}_{\text{train}}$. It means that the detector requires to be capable of localizing and recognizing objects belonging to novel categories $\mathbb{C}_{\text{novel}} = \mathbb{C}_{\text{target}} - \mathbb{C}_{\text{train}}$ that are not seen in training. In the strict problem setting, the vocabulary set of novel categories $\mathbb{C}_{\text{novel}}$ is unavailable in training.

The Overall EdaDet Architecture is shown in Figure 3a. Following previous works [16, 54, 26, 38], we break down open-vocabulary detection as two subsequent branches: (1) to generate class-agnostic object proposals and (2) to recognize open-vocabulary categories for these object proposals.

For the proposal generation, to ensure an efficient end-to-end detector and avoid hand-designed modules like anchor generation, we adopt a query-based proposal generation method like DETR [3] and retrofit it to class-agnostic. Specifically, for each object query, we predict a class-agnostic object confidence score $\mathcal{S}_{\text{obj}}$ to evaluate whether it is an object or background and regress its bounding box $\mathcal{B}$. We also use the bipartite matching loss for set prediction following DETR and denote the sum of confidence score loss and bounding box regression loss as $\mathcal{L}_{\text{box}}$. By the class-agnostic retrofit, the object proposals generated by our proposal generation branch have similar top-300 average recall to the RPN network of the existing Mask R-CNN detector.

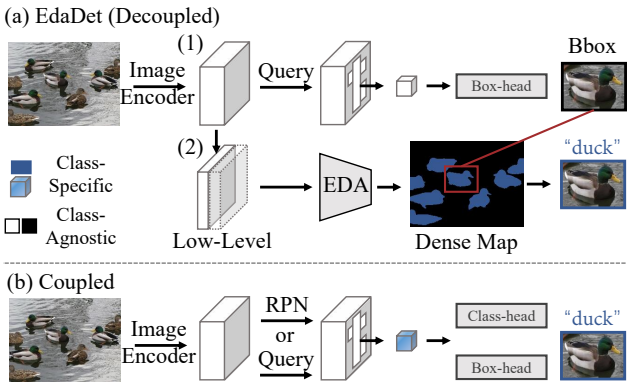


Figure 3: Architecture comparison between (a) our deeply-decoupled EdaDet and (b) existing open-vocabulary detection framework. EdaDet separates the open-vocabulary classification branch from the class-agnostic proposal generation branch at a more shallow layer of the decoder. EdaDet first individually generates object proposals and predicts dense probabilities to categories for local image semantics and then computes object proposals' categories based on the dense probabilities.

However, we observe that both the Mask R-CNN detector and our initial attempt predict the bounding box regression on the same feature space as classifying open-vocabulary proposal with only using two separate heads for box regression and classification, as shown in Figure 3b. We believe that the highly correlated predictions between the two branches hurt the class-agnostic proposal generation and validate this observation in Table 5. Also, OLN [24] shares a similar observation to us for learning open-world object proposals. Therefore, we separate the open-vocabulary classification from the proposal generation branch at a more shallow layer of the decoder, as shown in Figure 3a, to deeply decouple the two branches.

3.3. Early Semantic Alignment at Dense Level (Eda)

Our early dense alignment (Eda) performs open-vocabulary classification for class-agnostic proposals by leveraging the generalizable CLIP. As in previous methods [50, 16, 26, 54, 38, 28], we use text embeddings of categories extracted from frozen CLIP's text encoder $g(\cdot)$ as the classifier and denote the set of text embeddings of base categories $\mathbb{C}_{\text{train}}$ in training as E_{train}.

However, unlike previous methods that classify proposals by aligning their object-level visual features to the base classifier, we first align dense local image semantics early to the classifier by using object-level supervision. Then, we classify proposals according to the dense probabilities to categories to mitigate the overfitting to base categories $\mathbb{C}_{\text{train}}$ led by object-level alignment. Our dense-level alignment can persevere the fine-grained recognition ability to distinguish local semantics details for similar novel and base categories, which further helps to better generalize from similar base objects to novel objects.

Early Dense Semantic Alignment. Given the input image $\mathcal{I}$, we first extract its image feature map $\mathcal{F}_i(\mathcal{I})$ via our visual backbone $\mathcal{F}_i(\cdot)$, where feature at each spatial position of the feature map $\mathcal{F}_i(\mathcal{I})$ represents a local image semantic and i represents the feature of the i-th layer. Then, we calculate the probability map of each local image semantic belonging to each base category as follows,

$$\mathcal{S}_{\text{detector}} = \text{Softmax}(\cos(\mathcal{F}_i(\mathcal{I}), E_{\text{train}})/\tau), \tag{3}$$

where $\cos(\cdot,\cdot)$ represents the cosine similarity, and τ is the temperature coefficient. We also compute the CLIP's dense probability map $\mathcal{S}_{\text{CLIP}}(f_{\text{Backbone}}(\mathcal{I}), E_{\text{train}})$, *i.e.*, $\text{Softmax}(\cos(f_{\text{G-Pooling}}(f_{\text{Backbone}}(\mathcal{I})|\mathcal{M})_{[\text{patches}]}, E_{\text{class}})/\tau)$. Inspired by F-VLM [26] that fuse objects' CLIP scores and detection scores via geometric mean, we obtain the overall dense score map $\mathcal{S}$ as follows,

$$\mathcal{S} = \mathcal{S}_{\text{detector}}^{1-\lambda} \circ \mathcal{S}_{\text{CLIP}}^{\lambda}, \tag{4}$$

where $\lambda \in [0, 1]$ controls weights for the probability maps $\mathcal{S}_{\text{detector}}$ and $\mathcal{S}_{\text{CLIP}}$, and $\circ$ means element-wise product.

Next, we classify class-agnostic proposals generated from our proposal generation branch based on the overall dense score map $\mathcal{S}$. Given a proposal with predicted bounding box $\mathcal{B}$, we pool the box $\mathcal{B}$ into a fixed size score map $\text{RoIAlign}(\mathcal{B}, \mathcal{S})$ by performing RoIAlign [20] operation on the overall scores $\mathcal{S}$. Based on our observation that an object typically does not occupy the entire box, we independently average the top-k highest scores for each category in the fixed size score map $\text{RoIAlign}(\mathcal{B}, \mathcal{S})$ as the proposal's predicted score for that category, which is formulated as:

$$\mathcal{S}_{\text{proposal}} = \text{M-Pooling}(\text{RoIAlign}(\mathcal{B}, \mathcal{S}) \circ \mathcal{M}_k), \tag{5}$$

where $\text{M-Pooling}(\cdot)$ means the mean pooling layer. And mask $\mathcal{M}_k(i,j,c) = 1$ means that the score $\text{RoIAlign}(\mathcal{B}, \mathcal{S})(i,j,c)$ at spatial position (i,j) is among the top-k highest scores for the c-th category, and otherwise $\mathcal{M}_k(i,j,c) = 0$. And $\mathcal{S}_{\text{proposal}}$ represent the proposal's scores to each base category. Note that the mask $\mathcal{M}_k$ allows non-object parts to retain their original semantics and avoid their overfitting to the proposal's category.

During training, we minimize the cross-entropy loss $\mathcal{L}_{\text{cls}}$ with respect to each proposal's classification scores $\mathcal{S}_{\text{proposal}}$. During inference, we replace the text embeddings of base categories with that of target categories and predict the proposal's label as the category with the highest score.

Global Semantic Alignment. We observe that only local alignment leads to losing the global semantic information for the image. Therefore, we align the integrated local image semantics to CLIP's image encoder to improve the dense alignment. Specifically, given the global image

feature $f_{\text{G-Pooling}}(f_{\text{Backbone}}(\mathcal{I}))_{[\text{cls}]}$ extracted from CLIP, we apply L_1 loss to align the local image semantics $\mathcal{F}(\mathcal{I})$:

$$\mathcal{L}_{\text{g}} = \|\text{M-Pooling}(\mathcal{F}(\mathcal{I})) - f_{\text{G-Pooling}}(f_{\text{Backbone}}(\mathcal{I}))_{[\text{cls}]}\|_1. \quad (6)$$

4. Experiment

4.1. Benchmark and Implementation Detail

LVIS benchmark [17] consists of object detection and instance segmentation labels for 1203 object categories. The categories are split into three groups: frequent, common, and rare. Following ViLD [16], we treat frequent and common categories as the base categories during training and treat the 337 rare categories as the novel categories during testing. And the 337 rare categories are excluded from the training set. The mask mAP on novel categories is the key evaluation metric for LVIS.

COCO benchmark [29] is a common benchmark for numerous studies on open-vocabulary detection. The COCO vocabulary is partitioned into 48 base categories for training and 17 novel categories for testing. We conform to the standard protocol and report results under the generalized detection settings. The key evaluation metric is the box AP50 of novel categories.

Implementation Details. Our detection framework is based on DETR [3] following OV-DETR [47]. For fair comparison, we follow [54, 47] to train our detector for 10.2k iterations with batch size 32 and image size 800 $\times$ 800 and adopt AdamW optimizer with weight decay 1e-4 and initial learning rate 2e-4. We set the fixed size of the score map and k in Eq 5 as 14$\times$14 and 12$\times$12, respectively. The temperature coefficient τ=1e3 in Eq 3, and $\lambda = 0.25$ in Eq 4. We use standard CLIP's ImageNet prompts to extract text embeddings by CLIP-R50. We fuse feature maps from conv2 x and conv3 from the backbone to perform Early Dense Alignment, i.e. i=2,3 in Equ 3. Unless otherwise specified, the experiments are conducted under the same setting.

4.2. Comparison with State-of-the-Art Detectors

We evaluate EdaDet on various open-vocabulary object detection and instance segmentation benchmarks. Results in Table 1 show that EdaDet achieves a stronger base-to-new generalization on both COCO and LVIS benchmarks. Despite being trained in a strict open-vocabulary setting and without using any additional training resources, EdaDet consistently outperforms all prior methods, even these under more relaxed experimental setups.

COCO Benchmark. Table 1 and Table 4 (w self-training) show that EdaDet consistently and significantly outperforms state-of-the-art methods under both weak and strict open-vocabulary settings. Specifically, EdaDet outperforms OV-DETR [47], which shares the same setting (strict and without using any additional training resources) and detection framework with us by $+\mathbf{8.4}$ box AP50 of novel categories. Compared to the strict best-performing VLDet [28] that is trained with additional caption resources, we improve box AP50 of novel categories by $+\mathbf{5.8}$. Moreover, following the same weak open-vocabulary setting as best-performing OC-ovd [38], our EdaDet (without using external training resources) can further boost the box AP50 of all categories and the novel categories to 57.1 and 40.2 (see Table 4), which outperforms all state-of-the-art methods by $+\mathbf{5.6}$ and $+\mathbf{3.3}$, respectively. Different from LVIS, we observed significant overfitting to the base classes on the COCO dataset. Therefore, for COCO, we took the specific approach of using an RPN trained on the base classes to additionally extract class-agnostic proposals.

LVIS Benchmark. Table 1 demonstrates that EdaDet achieves the new state-of-the-art mask AP50 of novel categories and is very competitive under different experimental settings. Compared to approaches under the same setting with us, EdaDet offers better performance, *i.e.*, $+\mathbf{3.9}$ mask AP_{Novel} improvement compared with best-performing DetPro [13]. Compared to the leading method VLDet [28] which leverages a large image-text dataset CC3M to expand vocabulary, our method directly transfers knowledge from VLMs and improves the mask AP_{Novel} by $+\mathbf{2.0}$. Moreover, in contrast with ViLD [16], our EdaDet still preserves the performance of base categories when improving the novel classes. Even compared with the ensemble version of ViLD-ens, EdaDet still boosts the performance by $+\mathbf{2.0}$ and $+\mathbf{7.1}$ on mask AP_{All} and mask AP_{Novel}, respectively.

Transfer Detection Benchmark. We conduct a transfer detection experiment to assess the effectiveness of EdaDet as a universal detector for various data sources. Considering the base categories of the LVIS dataset contain almost all the COCO categories, and COCO shares the same image sources as LVIS, we do not conduct the transfer detection from LVIS to COCO like ViLD [16]. Instead, following [28], we conduct a more challenging transfer experiment that evaluates the COCO-trained model on LVIS and the LVIS-trained model on Objects365 [42]. Specifically, we simply replace the COCO-based classifier (80 categories) with the LVIS-based classifier (1203 categories) without fintuning and report the box AP as the evaluation metric on LVIS of all categories. The transfer from LVIS to Objects365 follows a similar protocol. The results are shown in Table 2. For the transfer from COCO to LVIS, we observe that the method [28] under strict open-vocabulary setting usually outperforms those under weak setting [38, 54] due to the better generalization ability of the former. Our EdaDet outperforms the best-performing strict VLDet [28] by $+\mathbf{2.8}$ box AP_{50}. For objects365, EdaDet outperforms state-of-the-art methods [26, 13, 16] by $+\mathbf{1.7}$ and $+\mathbf{2.0}$ in terms of AP and AP_{75}, respectively.

Methods	Publication	Training Resources for Novel Proposal Generation	COCO Detection			LVIS Segmentation			
			$AP50^{box}$	$AP50^{box}_{base}$	$AP50^{box}_{novel}$	AP^{mask}	AP^{mask}_{f}	AP^{mask}_{c}	AP^{mask}_{novel}
Weak open-vocabulary Setting									
Detic [54]	ECCV2022	IN-O	45.0	47.1	27.8	26.8	31.6	26.3	17.8
PromptDet [14]	ECCV2022	LAION	50.6	-	26.6	21.4	25.8	18.3	19.3
VL-PLM [49]	ECCV2022	-	48.3	54.0	32.3	-	-	-	-
PB-OVD [15]	ECCV2022	CAP-L	42.1	46.1	30.8	-	-	-	-
OC-ovd [38]	NeurIPS2022	IN-O + COCO CAP	51.5	56.6	36.9	25.9	29.1	25.0	21.1
LocOv [2]	GCPR2022	COCO CAP	45.7	51.3	28.6	-	-	-	-
Strict open-vocabulary Setting									
CLIP-RPN	-	-	27.8	28.3	26.3	17.7	16.0	18.8	18.9
OVR-CNN [48]	CVPR2021	COCO CAP	39.9	46.0	22.8	-	-	-	-
ViLD [16]	ICLR2022	-	51.3	59.5	27.6	22.5	28.3	20.0	16.1
ViLD-Ens. [16]	ICLR2022	-	-	-	-	25.5	30.3	24.6	16.6
DetPro [13]	CVPR2022	-	-	-	-	25.9	28.9	25.6	19.8
RegionCLIP [50]	CVPR2022	COCO CAP + CC3M	50.4	57.1	31.4	28.2	34.0	27.4	17.1
OV-DETR [47]	ECCV2022	-	52.7	61.0	29.4	26.6	32.5	25.0	17.4
OWL-ViT † [38]	ECCV2022	-	-	-	-	19.3	-	-	16.9
VLDet [28]	ICLR2023	COCO CAP + CC3M	45.8	50.6	32.0	30.1	34.3	29.8	21.7
F-VLM [26]	ICLR2023	-	39.6	-	28.0	24.2	26.9	24.0	18.6
Ours (RN50 backbone)	ICCV2023	-	52.5	57.7	**37.8**	27.5	29.1	27.5	**23.7**

Table 1: Open-vocabulary object detection results on COCO and LVIS datasets. All the methods share the RN50 backbone except † with ViT-B/32. The IN-O is the ImageNet21k [10] (IN-21K) dataset's subset with 997 overlapping categories with LVIS [17]. LAION [41] is a large-scale image-text dataset, CC3M represents the Conceptual Caption 3M [5], and CAP-L consists of COCO Caption [29], Visual-Genome [25] and SBU Caption [33].

Method	COCO → LVIS				LVIS → Objects365		
	AP	AP_{50}	AP_{75}		AP	AP_{50}	AP_{75}
OC-ovd†	5.6	8.5	6.0	VilD	11.8	18.2	12.6
Detic†	5.5	8.5	5.8	DetPro	12.1	18.8	12.9
VLDet	-	10.0	-	F-VLM	11.9	19.2	12.6
Ours	**9.1**	**12.8**	**9.6**	Ours	**13.6**	**19.8**	**14.6**

Table 2: Transfer detection of EdaDet. We evaluate COCO-trained model on LVIS and LVIS-trained model on Objects365 without finetuning. †: evaluate with official code and checkpoint.

Method	Size	AP^{mask}_{novel}	#Iters	Epochs
ViLD-EN-B7	-	26.3	180k	460
OWL-ViT-Large	1216	31.2	70k	180
F-VLM-RN50x64	812	32.8	46.1k	118
VLDet-ViT-Base	345	26.3	90k	13
EdaDet-ViT-Base	345	29.9	42k	10
EdaDet-ViT-Base	345	35.6	126k	30

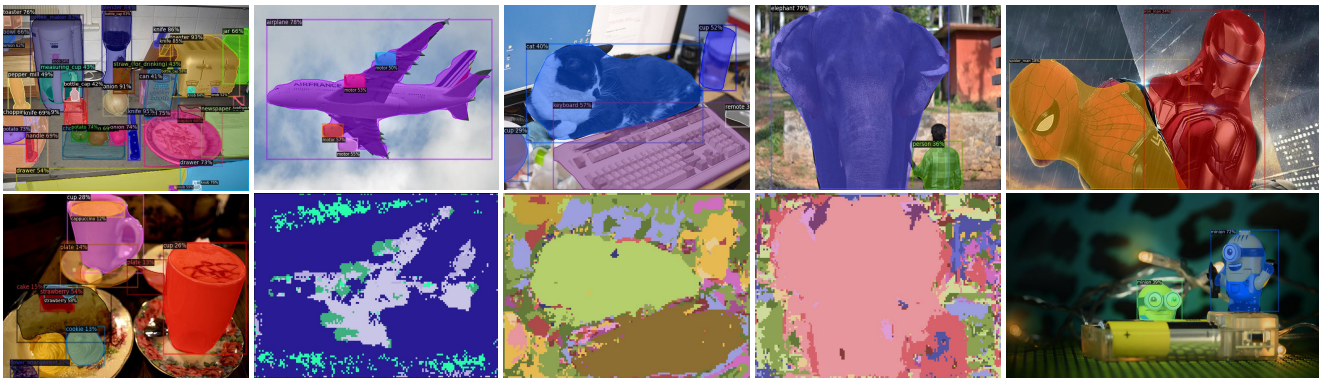
Table 3: Comparison of scalability and training efficiency on LVIS. We report AP^{mask}_{novel} to show the trade-off between performance and training costs under relatively fair model sizes in MB.

Scalability and Training Efficiency Benchmark. Table 3 summarizes the performance, iteration and epoch of different methods with the strict setting using large backbone networks. Notice that since different methods adopt different backbones and batch sizes, Table 3 is not for an apple-to-apple comparison but just to illustrate the scalability and training efficiency for different methods. With a relatively smaller backbone ($1/4$ of OWL-ViT), EdaDet achieves a 35.6 (**+4.4**) AP^{mask}_{novel} with a shorter training epochs.

4.3. Ablation Study

Roadmap to build a strong open-vocabulary detector is shown in Table 4. We start by training a conventional deformable DETR [55] detector on base categories, which achieves 61.7 AP50 of base categories. (1) To equip the base model with open-vocabulary detection ability, we replace its classifier with text embeddings of categories extracted from CLIP-R50. (2) Then, we deeply decouple the proposal generation and classification branches. To train the proposal generation branch on more objects beyond base objects in images, we use the annotated base objects and the generated proposals with high confidence scores and without overlapping with annotated objects as the box supervision. Note that we do not generate pseudo novel proposals from external training resources but only extend the base objects with class-agnostic proposals with high confidence scores. (3) Next, we ensemble text prompts to obtain a better classifier. So far, the model can achieve 15.1 AP50 of novel categories, while the AP50 of base categories yields a performance drop by -4.2.

Furthermore, we stepwisely develop our early dense alignment (Eda). (4) By replacing the object-level alignment with our dense-level alignment (4.1.1), the AP50 of novel categories significantly improves to 30.1, demonstrating the effectiveness of preserving the fine-grained details to distinguish the similar novel and base objects. (5) We



(a) LVIS (b) COCO (c) Out-of-distribution images

Figure 4: Qualitative results of EdaDet. We visualize our detection results and semantic maps on (a) LVIS [17], (b) COCO [29] and (c) out-of-distribution images from the open-source website.

Ablation	$AP50^{box}_{base}$	$AP50^{box}_{novel}$
Supervised from base	61.7	0.0
(1) Replace classifier with CLIP-R50	56.6	7.6
(2) Deeply decouple from (1)	57.3	14.2
(3) Add prompt ensembling from (2)	57.5	15.1
(4) Early Dense Alignment from (3)		
(4.1) Dense Alignment		
(4.1.1) $\lambda = 0$, w/o $\mathcal{M}_k$	56.7	30.1
(4.1.2) $\lambda = 0$, w $\mathcal{M}_{k=7\times7}$	55.5	29.3
(4.1.3) $\lambda = 0$, w $\mathcal{M}_{k=12\times12}$	57.3	33.1
(4.1.4) $\lambda = 0.25$, w $\mathcal{M}_{k=12\times12}$	57.2	35.7
(4.2) Global Alignment from (4.1.4)	57.7	37.8
(4.3) Early Dense Alignment in con2_x	56.4	35.9
(4.4) Early Dense Alignment in con3_x	56.7	36.2
(5) Self-training from (4.2)	57.1	40.2

Table 4: Ablation study of EdaDet on the COCO dataset.

Methods	AR	AR_S	AR_M	AR_L
Mask-RCNN	0.31	0.19	0.39	0.53
Def-Detr	0.33	0.21	0.41	0.51
Decomp	0.36	0.24	0.46	0.57

Table 5: Ablation study of proposal generation. We report top-300 average recall (AR) of all categories on LVIS. The AR_S, AR_M and AR_L denote AR for the small, medium and large objects, respectively. All methods are trained on LVIS base categories.

further add the mask $\mathcal{M}_k$ to allow non-object local image semantics in the object box to not directly participate in computing object categories (4.1.2 and 4.1.3), which helps mitigate overfitting base categories. The AP50 of novel categories is increased to 33.1. (6) Moreover, we integrate CLIP's probability map in our dense probability map (4.1.4), slightly improving the AP50 of novel categories by 2.6 while maintaining the AP50 of base categories to 57.2. (7) We plug the global alignment into Eda (4.2), and the performance gain is $+\mathbf{2.1}$ AP50 on novel categories. In summary, our Eda, including the dense and global alignments, brings in $+\mathbf{22.7}$ improvement in AP50 of novel categories, which illustrates its effectiveness.

How much does deeply decoupling of two branches help proposal generation. Compared to Deformable DETR [55] and Mask-RCNN [20], the consistent improvement on all metrics indicates the superiority of deeply decoupling label prediction and box regression (see Table 5).

4.4. Qualitative Visualization

We visualize EdaDet's detection results and semantic maps $\mathcal{S}_{\text{detector}}$ (Eq 3) in Figure 4. For LVIS with diverse target categories and complex scenes, our EdaDet performs well on crowded prediction. Figure 4c shows that EdaDet even successfully detects the film characters *iron man*, *spider man* and the animation character *minion*, which demonstrates EdaDet's open-vocabulary capacity and the importance of generalizing pretrained VLMs on open-vocabulary detection.

5. Conclusion

We propose a simple but effective open-vocabulary detection method (EdaDet) that generalizes the pretrained VLMs to achieve a strong base-to-novel detection ability. Experiments on various datasets show that EdaDet consistently outperforms state-of-the-art methods in open-vocabulary object detection and instance segmentation. **Ethics Statement:** Since our open-vocabulary capability is solely derived from VLMs, biases, stereotypes and controversies that may exist in the image-text pairs for training VLMs may be introduced into our models.

Acknowledgment: This work was supported by the National Natural Science Foundation of China (No.62206174), Shanghai Pujiang Program (No.21PJ1410900), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 3

[2] M. Bravo, S. Mittal, and T. Brox. Localized vision-language matching for open-vocabulary object detection. In *German Conference on Pattern Recognition (GCPR) 2022*, 2022. 1, 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1, 3, 4, 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3, 7

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1597–1607. PMLR, 13–18 Jul 2020. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 7

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[13] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 1, 3, 4, 6, 7

[14] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 1, 4, 7

[15] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Towards open vocabulary object detection without human-provided bounding boxes. *arXiv preprint arXiv:2111.09452*, 2021. 7

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 3, 4, 5, 6, 7

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3, 6, 7, 8

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[19] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 3

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3, 5, 8

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[22] 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, 3

[23] Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. Transferability in deep learning: A survey. *arXiv preprint arXiv:2201.05867*, 2022. 3

[24] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 2022. 5

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 7

[26] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 1, 3, 4, 5, 6, 7

[27] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[28] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 1, 4, 5, 6, 7

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 6, 7, 8

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1, 3

[31] 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. 4

[32] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 3

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 7

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 4

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[37] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[38] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *36th Conference on Neural Information Processing Systems (NIPS)*, 2022. 1, 3, 4, 5, 6, 7

[39] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 3

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[41] 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. 3, 7

[42] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3, 6

[43] 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, 3

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[45] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95, 2002. 3

[46] Johnathan Xie and Shuai Zheng. Zero-shot object detection through vision-language embedding alignment. *arXiv e-prints*, pages arXiv–2109, 2021. 4

[47] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 106–122. Springer, 2022. 1, 4, 6, 7

[48] 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, 7

[49] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pages 159–175. Springer, 2022. 7

[50] 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, 3, 4, 5, 7

[51] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3, 4

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[54] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand

classes using image-level supervision. In *ECCV*, 2022. 1, 3, 4, 5, 6, 7

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 7, 8