ProDet: Learning Prototype-Based Region-Word Alignment for Open-Vocabulary Object Detection

Anonymous ICCV submission

Paper ID 12956

Abstract

The open-vocabulary object detection (OVD) task aims to detect unseen categories that are not annotated during training. Existing methods typically rely on vision-language models pre-trained on large-scale image-text pairs, such as CLIP, to perform the detection of unseen categories. However, CLIP cannot model fine-grained region-word alignment knowledge, leading to inaccurate localization during detection. To this end, we propose a lightweight method for learning a set of prototypes that capture fine-grained region-word alignment for unseen categories. Our method first utilizes large-scale language-image generation (LLIG) models to generate example images for each unseen category. To further obtain pseudo-labels for training prototypes, we introduce the key-point-guided localization module, which utilizes cross-attention maps from LLIG models to extract bounding boxes for the target objects. These generated images and pseudo-labels enable us to learn a set of prototypes that represent common visual features of unseen categories. At inference time, we combine prototype-based scores and text-based CLIP scores to obtain the final prediction results. Our approach delivers highly competitive performance, achieving 40.6 mAP$_r$ on OV-LVIS and 44.9 AP$_{novel}$ on OV-COCO, without pre-training or joint training on additional datasets.

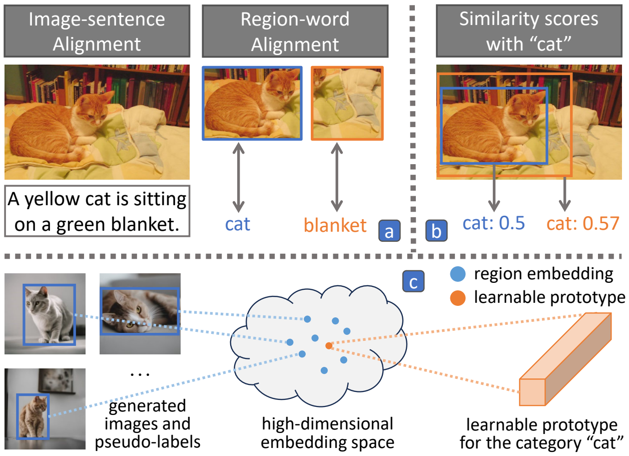
Image-sentence Alignment
Region-word Alignment
Similarity scores with "cat"
A yellow cat is sitting on a green blanket.
cat
blanket
a
b
cat: 0.5
cat: 0.57
c
region embedding
learnable prototype
...
generated images and pseudo-labels
high-dimensional embedding space
learnable prototype for the category "cat"

Figure 1. **(a) Image-sentence alignment vs. region-word alignment.** CLIP is trained to align complex images with descriptive sentences (image-sentence alignment), while the OVD task requires drawing bounding boxes for target objects and mapping each bounded region to a category name (region-word alignment). **(b) An example of the CLIP model struggling to accurately locate specific regions.** Since CLIP has not been trained for region-word alignment, it often assigns higher similarity scores to bounding boxes that contain more irrelevant information (orange), rather than to more precise and accurate bounding boxes (blue). **(c) Illustration of the proposed learnable prototypes.** We propose learnable prototypes that capture the fine-grained features of each unseen category from generated samples, thereby aiding CLIP in classifying unseen categories more effectively.

1. Introduction

The open-vocabulary object detection task(OVD)[32] has attracted extensive research in recent years to address the limitation of vocabulary in the traditional object detection task. In the OVD task, a set of base categories is annotated in all training images, and the detector is required to detect a set of unseen categories during inference time. To this end, the CLIP [20] model, which is pre-trained on 400 million image-text pairs covering a vast set of categories, has been introduced to the OVD models, serving as an encyclopedia for detecting unseen categories.

The central idea in existing works is to align region embeddings with text embeddings, both provided by the CLIP model. In this way, we can recognize unseen categories solely based on their category names. However, the CLIP model does not perform well in aligning regions with words. As shown in Figure 1(a), CLIP is trained to match complex images with descriptive sentences (image-sentence alignment), while the OVD task requires drawing bounding boxes for target regions and mapping each bounded region to a category name (region-word alignment). This misalignment causes CLIP to struggle with precisely localizing specific regions (Figure 1(b)).

To address the above problem, previous studies retrain the model by constructing explicit pseudo-labels in the form of bounding boxes for unseen categories [7, 28, 34, 35] or employ image-level supervision from additional human-annotated datasets to learn a large number of novel concepts [36]. However, these methods require high annotation costs and significant computational resources during training. Detailed comparisons are illustrated in Table 1.

In this paper, we propose a lightweight architecture to learn region-word alignment by training a set of prototypes on synthetic examples, which we call ProDet. Specifically, ProDet learns a common feature from various generated instances for each unseen category, termed a *prototype*. In this way, we can classify unseen objects by mapping their region proposals to the most similar prototype.

However, there are two main challenges in learning prototypes for unseen categories, including (1) how to obtain images of unseen categories and (2) how to obtain pseudo-labels of unseen objects for training. To overcome the first challenge, we repeatedly prompt the stable diffusion [23] model to generate sample images for unseen categories. Given generated sample images, we also need accurate bounding boxes and classification labels to learn a set of prototypes for unseen categories. The classification labels can be directly derived from the image generation prompts. Then, our key-point-guided localization module uses *image-text cross-attention maps* from the stable diffusion model to identify *key points* of target objects. Prompted by these key points, the SAM [14] model generates accurate segmentation masks for the target objects, from which we can derive their bounding boxes.

Given the automatically generated images and pseudo-labels, ProDet extracts common visual features for unseen categories by learning prototype-based region-word alignment. Compared to detecting objects based solely on category names, the proposed architecture introduces fine-grained visual reference information into the classification process of the detector, enhancing its accuracy.

The contributions of this paper are summarized below.

- We propose a new architecture to learn prototype-based region-word alignment from several generated examples of unseen categories.
- We design a novel method that uses cross-attention maps to automatically obtain bounding boxes and classification labels for generated images of unseen categories.
- Our method delivers highly competitive performance, achieving 40.6 mAP$_r$ on the OV-LVIS benchmark, alongside 44.9 AP$_{novel}$ and 56.7 AP$_{all}$ on the OV-COCO benchmark, without annotating large volumes of image-text pairs or joint training on additional human-annotated datasets.

2. Related Work

2.1. Open-vocabulary object detection

Open-vocabulary object detection (OVD) aims to detect unseen categories given arbitrary category names through free-form textual input. During training, detectors are provided with human-annotated training samples of base categories. During inference, the OVD task requires detectors to recognize both base and unseen categories. Recently, researchers have utilized pre-trained vision-language models like CLIP [20] to enhance image-text alignment. Some works use CLIP to acquire open-vocabulary recognition ability by knowledge distillation [8, 27] or directly freeze the CLIP model and fine-tune the localization capability of the detector [15, 32]. For instance, VilD [8] distills the knowledge from the CLIP model into a two-stage detector, while F-VLM [15] combines the outputs of the detector and the frozen CLIP for object detection.

2.2. Learning region-word alignment for open-vocabulary object detection

Although CLIP can provide open-vocabulary recognition knowledge for the detector to recognize unseen categories, directly integrating CLIP into an OVD detector presents challenges. The limitation of CLIP lies in its alignment of complex images and descriptive sentences during pre-training, which is not directly applicable to region-word alignment in object detection.

To address this issue, some methods construct bounding box labels for pre-training on additional image-text pairs [7, 28, 34, 35]. For example, RegionCLIP [35] aligns region proposals with captions derived from image-text pairs and pre-trains the CLIP model to map these region-word pairs into a shared feature space. Some other approaches incorporate supplementary human-annotated datasets into the training set for joint training [34, 36]. For instance, Detic [36] jointly trains the classifiers of a detector with image-level supervision from image classification datasets, while VL-PLM [34] generates pseudo-labels for unlabeled images and jointly trains the detector on the GT-labeled examples of base categories and the pseudo-labels of unseen categories. Although effective, these methods require high annotation costs and demand significant computational resources during training. Unlike previous methods, our approach uses LLIG models to generate a small set of sample images and automatically obtain their pseudo-labels. During training time, we train the detector only on the GT-labeled examples of base categories. Then, we freeze the detector and use the generated bounding box labels to fine-tune a set of prototypes for unseen categories. This design avoids retraining the detector when new categories are encountered while ensuring that our approach adheres to a strict open-vocabulary setting, wherein prior knowledge of new classes remains

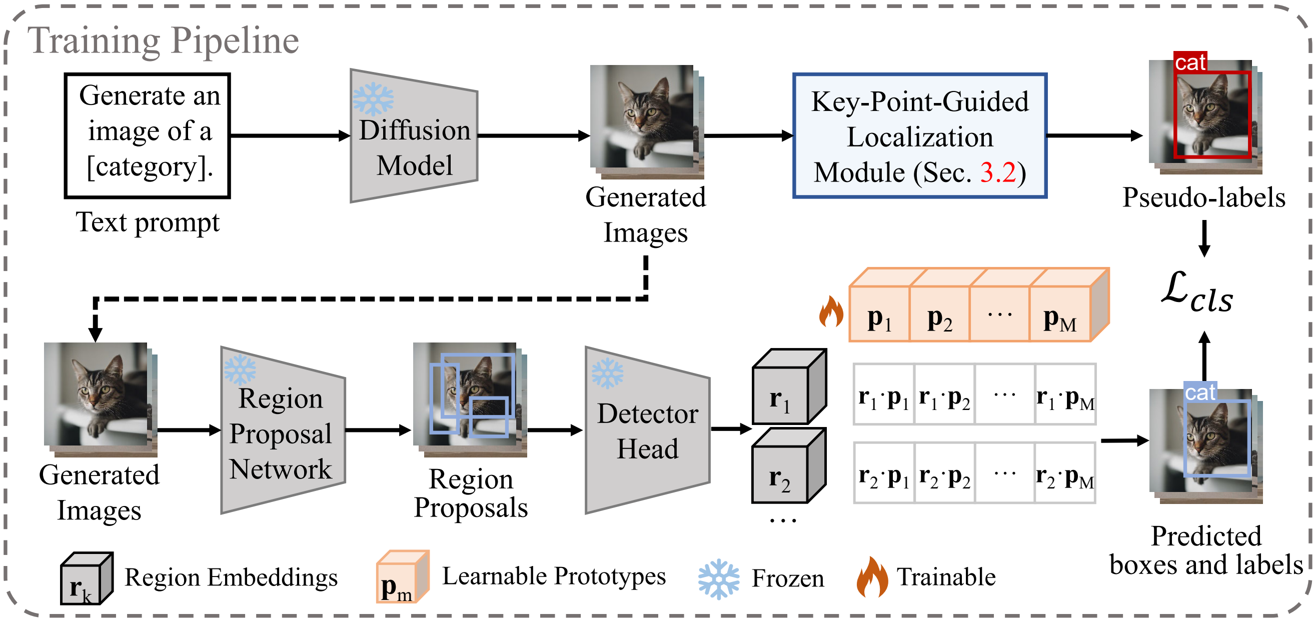
Training Pipeline
Generate an image of a [category].
Text prompt
Diffusion Model
Generated Images
Key-Point-Guided Localization Module (Sec. 3.2)
cat
Pseudo-labels
$\mathcal{L}_{cls}$
$\mathbf{p}_1$ $\mathbf{p}_2$ ⋯ $\mathbf{p}_M$
Generated Images
Region Proposal Network
Region Proposals
Detector Head
$\mathbf{r}_1$
$\mathbf{r}_2$
⋯
$\mathbf{r}_1 \cdot \mathbf{p}_1$ $\mathbf{r}_1 \cdot \mathbf{p}_2$ ⋯ $\mathbf{r}_1 \cdot \mathbf{p}_M$
$\mathbf{r}_2 \cdot \mathbf{p}_1$ $\mathbf{r}_2 \cdot \mathbf{p}_2$ ⋯ $\mathbf{r}_2 \cdot \mathbf{p}_M$
cat
Predicted boxes and labels
$\mathbf{r}_k$ Region Embeddings
$\mathbf{p}_m$ Learnable Prototypes
Frozen
Trainable

Figure 2. **The training pipeline of the proposed ProDet.** Our approach first uses the stable diffusion model to generate sample images for unseen categories. Simultaneously, the proposed key-point-guided localization module automatically obtains box labels for sample images. During training, the detector generates multiple region proposals and their region embeddings for the input images. We initialize a set of learnable prototypes and detect objects by matching each region proposal to the most similar prototype. Then, we update the learnable prototypes by minimizing classification loss L_{cls}.

unknown during the detector training phase.

2.3. Diffusion models (DMs)

Diffusion models [19, 21, 23] have demonstrated highly competitive generative quality in text-to-image tasks. Typically, they leverage pre-trained vision-language models like CLIP [20] to encode the text condition and use cascade diffusion structures like U-Net [24] to progressively denoise random Gaussian noise into a coherent image over multiple timesteps. Due to their ability to learn complex data distributions and generate high-quality synthetic data, diffusion models are useful for augmenting training datasets in cases of limited annotated data. Recent studies [11, 29, 31, 33] have explored their application in computer vision tasks like few-shot classification and instance segmentation.

3. Method

We propose a new architecture to learn region-word alignment by training a set of prototypes using automatically generated pseudo-labels, namely ProDet. In the subsequent sections, we first present the model architecture and training pipeline of ProDet (Sect. 3.1). We then introduce a novel key-point-guided localization module to automatically obtain pseudo-labels for prototype training (Sec. 3.2). Finally, we describe the inference pipeline of ProDet, proposing a prototype text ensemble module to help ProDet recognize a wider range of instance variation (Sect. 3.3).

3.1. Learning Prototype-Based Region-Word Alignment

In this study, we propose to learn a category-specific prototype for each category to enable precise recognition in open-vocabulary settings. Our approach begins by deriving region embeddings using a Region Proposal Network (RPN) and a detector head, as illustrated in Figure 2. Specifically, for an input image associated with a category c, a class-agnostic RPN first generates a set of K region proposals, each representing a potential target object. Subsequently, the detector head computes a corresponding region embedding $\mathbf{r}_k \in \mathbb{R}^d$ for the k-th region proposal, encapsulating the visual characteristics unique to that region. These embeddings enable us to derive a unified representation that captures the shared features of category c, designated as its *prototype*, which serves as a discriminative feature signature for category-level recognition in open-vocabulary settings.

Let $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M \in \mathbb{R}^d$ denote a set of learnable prototypes, where M represents the total number of prototypes. During training, we compute the cosine similarity between each region embedding $\mathbf{r}_k$ and the full set of prototypes as the detection score $\mathbf{z}_k$ of the detector.

$$\mathbf{z}_k = \text{Softmax}\left(\frac{1}{\tau}\left[\cos(\mathbf{r}_k, \mathbf{p}_1), \ldots, \cos(\mathbf{r}_k, \mathbf{p}_M)\right]\right), \quad (1)$$

where $\cos(\mathbf{r}_k, \mathbf{p}_m) = \frac{\mathbf{r}_k^{\mathrm{T}}\mathbf{p}_m}{||\mathbf{r}_k|| \, ||\mathbf{p}_m||}$, and τ is a fixed temperature. Then, we update the learnable prototypes by minimizing classification loss L_{cls} as follows.

$$L_{\text{cls}} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M} \alpha_m \left(1 - \mathbf{z}_{k,m}\right)^{\gamma} y_{k,m} \log(\mathbf{z}_{k,m}). \tag{2}$$

Instead of using traditional cross-entropy loss, we adopt the focal loss [17] as our classification loss. The focal loss integrates decaying factors α and γ to mitigate the dominance of easily classified samples during training, while simultaneously addressing the class imbalance between positive and negative instances. In Equation (2), y_k denotes the ground truth label of the k-th region proposal. For base categories, we directly leverage the ground-truth labels provided by the dataset to train their respective prototypes. For unseen categories, we first employ a generative approach to acquire sample images by utilizing the stable diffusion model [23]. Specifically, for each category name c, we formulate a standardized text prompt: "*generate a photo of a c*." This prompt is fed into the stable diffusion model, which then generates the content-related image I_c from a random Gaussian image noise over several time steps. Subsequently, we apply the proposed key-point-guided localization module, as detailed in section 3.2, to automatically produce bounding box labels for these generated images. This process facilitates the extraction of discriminative region embeddings, enabling effective learning of prototypes for unseen categories.

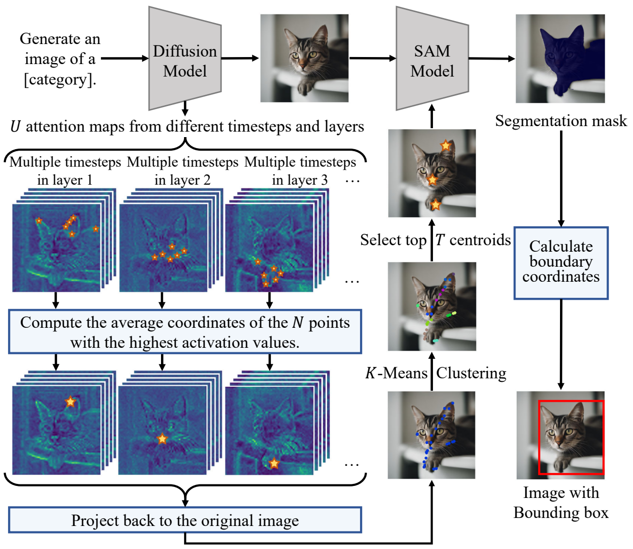
Generate an image of a [category].
Diffusion Model
SAM Model
Segmentation mask
U attention maps from different timesteps and layers
Multiple timesteps in layer 1
Multiple timesteps in layer 2
Multiple timesteps in layer 3
…
Select top T centroids
Calculate boundary coordinates
Compute the average coordinates of the N points with the highest activation values.
K-Means Clustering
Project back to the original image
Image with Bounding box

Figure 3. **Overview of the proposed key-point-guided localization module.** To provide the SAM model with precise coordinates, we propose a novel method to extract key points from multi-layer cross-attention maps. Specifically, given the cross-attention map from each layer and time step, we extract the top N locations with the highest activation values and compute the average coordinates of these N locations. Next, we apply K-means clustering to these average points and select the centroids of the top T clusters with the highest number of points. Finally, these centroids are utilized as key points to guide the SAM model in generating precise segmentation masks for the target objects, from which we can derive their bounding boxes.

Discussions on the advantages of learning prototypes. Iterative prototype learning facilitates the dynamic optimization of prototypes, enabling them to effectively capture the distinctive features of unseen categories. Through iterative training on synthetic data, our approach leverages this process to transfer open-vocabulary knowledge from large language-image-grounded (LLIG) models to the detector.

3.2. Key-Point-Guided Localization Module

Given the generated sample images of unseen categories, we also need to obtain bounding box labels for the target objects. To achieve high-quality bounding boxes, we propose utilizing an off-the-shelf segmentation model, SAM [14], to segment the target objects within the generated images and subsequently compute the boundary coordinates of the resulting masks to derive the bounding boxes. Specifically, for an image I_c and a set of n coordinates $\{(x_i, y_i)\}_{i=1}^{n}$ corresponding to the target object, the SAM model leverages these input coordinates to produce a segmentation mask for the object. The quality of this segmentation mask is directly dependent on the quality of the provided coordinates.

To provide the SAM model with precise coordinates, we propose a key point recognition algorithm designed to extract a set of coordinates that most effectively represent the target object within the input image I_c. Notably, the image I_c is generated by the stable diffusion model based on the text prompt "*generate a photo of a c*." We propose utilizing the multi-layer cross-attention maps of the stable diffusion model to pinpoint the key points of the target object. This approach is motivated by the fact that cross-attention maps in the stable diffusion model reflect the interactions between the text prompt and the image throughout the generation process. Within these maps, regions exhibiting higher attention scores to the category name demonstrate stronger associations with the target object, thus providing critical clues for pinpointing key points.

In this study, we utilize cross-attention maps from different layers to extract multi-scale information about the target object, while also employing cross-attention maps from various time steps to capture information at different levels of precision. As illustrated in Figure 3, let U denote the total number of the cross-attention maps from different layers and time steps. For each u-th cross-attention map, we focus exclusively on the channel corresponding to the category name, denoted as $A_u \in \mathbb{R}^{H \times W}$, to identify the key points associated with the target objects.

Specifically, for each cross-attention map A_u, we calculate the average coordinates of the top N points with the highest attention scores, obtaining an average point $P_u \in \mathbb{R}^2$. We consider this average point as the most focused point in A_u. By doing so, we obtain a list of average points $Coords$ for all U cross-attention maps. After the image generation, we use K-means to cluster the average points and obtain K clusters: $\{\mathcal{C}_k\}_{i=1}^{\mathrm{K}}$. The centroids of these clusters offer a reliable approximation of the key points of target objects. However, due to the potential influence of outlier values, these centroids alone are insufficient to serve directly as their key points. To address this, we apply centroid filtering by ranking the clusters in descending order based on their cardinality and selecting the centroids of the top T clusters as the final key points, denoted as $\{\mathbf{c}_t\}_{t=1}^{T}$. Algorithm 1 delineates the workflow of our proposed method.

Algorithm 1 Key point recognition algorithm

```
Input:
$\{A_u\}_{u=1}^{U}$ - cross-attention maps where $A_u \in \mathbb{R}^{H \times W}$
$N$ - number of top coordinates for averaging
$K$ - number of k-means clusters
$T$ - number of key points to select, where $T \leq K$
Output: Key points $\{\mathbf{c}_t\}_{t=1}^{T}$ for the SAM model
Initialize: $Coords \leftarrow [\,]$                ▷ Coordinate storage
for $u = 1$ to $U$ do
    $P_u \leftarrow \{(x, y) \mid A_u(x, y) \in \text{Top}_N(A_u)\}$      ▷ Coordinate extraction
    $\bar{\mathbf{p}}_u \leftarrow \frac{1}{N} \sum_{(x,y) \in P_u} (x, y)$      ▷ Spatial averaging
    $Coords$.append($\bar{\mathbf{p}}_u$)
end for
$\{\mathcal{C}_k\}_{k=1}^{K} \leftarrow$ KMeans($Coords, K$) ▷ Perform K-means clustering on coordinates
Sort clusters by cardinality in descending order: $|\mathcal{C}_{\sigma(1)}| \geq |\mathcal{C}_{\sigma(2)}| \geq \cdots \geq |\mathcal{C}_{\sigma(K)}|$ ▷ Renumber clusters as $\mathcal{C}_{\sigma(k)}$
for $t = 1$ to $T$ do
    $\mathbf{c}_t \leftarrow \frac{1}{|\mathcal{C}_{\sigma(t)}|} \sum_{(x,y) \in \mathcal{C}_{\sigma(t)}} (x, y)$  ▷ Compute centroid of cluster $\mathcal{C}_{\sigma(t)}$
end for
return $\{\mathbf{c}_t\}_{t=1}^{T}$
```

Subsequently, these key points are input into the SAM model to generate segmentation masks, from which boundary coordinates are derived to define the bounding boxes for the target objects. To mitigate the impact of poorly generated masks, we implement a mask-filtering strategy. Specifically, we exclude masks with areas below 5% or above 95% of the entire image, as such masks are highly prone to segmentation errors. In this way, we have developed a simple yet efficient method to automatically obtain bounding box labels for target objects. By performing this generation process in batches, we can create a small-scale synthetic dataset with images and pseudo-labels for unseen categories—about 100 images per category—while minimizing computational resources and time.

3.3. Prototype-Text Ensemble Module

After training, ProDet can rely solely on the trained prototypes for object detection. However, in real-world scenarios, different instances can exhibit significant visual variation of the same category. Since the prototypes for unseen categories are trained on a limited number of generated images (approximately 100 per category), the diversity of these instances is constrained, which restricts the ability of the trained prototypes to recognize a broad range of instance variations. As illustrated in Figure 4, to improve the accuracy of our model when detecting unseen categories in real-world settings, we follow the approach of CFM-ViT [12] by using the frozen CLIP model to extract CLIP region embeddings for auxiliary classification. Specifically, for the k-th region proposal of an input image, we use the frozen CLIP visual encoder and RoI-Align to extract the CLIP region embedding $\mathbf{v}_k$. Next, for each category name c, we first use a simple template to create a text prompt: "*A photo of a c*". Then we input this prompt into the frozen CLIP text encoder to obtain a set of text embeddings $[\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_M]$. Subsequently, we compute the cosine similarity between CLIP region embeddings and text embeddings to obtain the CLIP score $\mathbf{w}_k$.

$$\mathbf{w}_k = \text{Softmax}\left(\frac{1}{\mu}\left[\cos(\mathbf{v}_k, \mathbf{t}_1), \ldots, \cos(\mathbf{v}_k, \mathbf{t}_M)\right]\right) \quad (3)$$

μ denotes the fixed temperature. Finally, we ensemble the detection scores and the CLIP scores by geometric means to compute the final classification scores s_k.

$$\mathbf{s}_k = \begin{cases} (\mathbf{z}_k)_i^{\alpha} \cdot (\mathbf{w}_k)_i^{1-\alpha} & i \in C_B \\ (\mathbf{z}_k)_i^{\beta} \cdot (\mathbf{w}_k)_i^{1-\beta} & i \in C_N \end{cases} \quad (4)$$

where α, $\beta \in [0,\, 1]$ are floating-point parameters that regulate the weighting of base and unseen categories, while C_B and C_N denote the sets of base categories and unseen categories, respectively. For each region proposal, we select the category with the highest score as its classification result. At last, the model selects all region proposals with classification scores above a certain threshold, along with their corresponding categories, as the final output.

4. Experiments

4.1. Benchmark Setup.

We evaluate our approach on the **OV-COCO** and **OV-LVIS** datasets. The **OV-COCO** dataset originates from the famous COCO benchmark [16] for evaluating zero-shot detection [2] and open-vocabulary detection methods

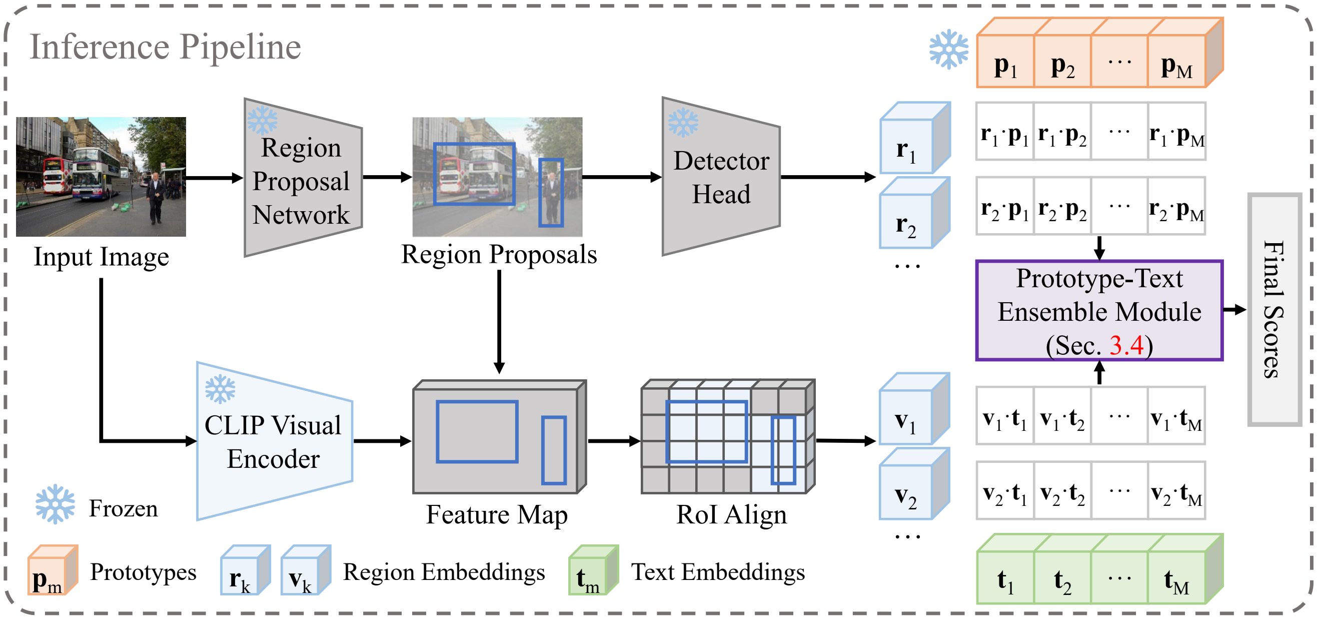
Inference Pipeline
Input Image
Region Proposal Network
Region Proposals
Detector Head
$\mathbf{r}_1$
$\mathbf{r}_2$
$\cdots$
$\mathbf{p}_1$ $\mathbf{p}_2$ $\cdots$ $\mathbf{p}_M$
$\mathbf{r}_1 \cdot \mathbf{p}_1$ $\mathbf{r}_1 \cdot \mathbf{p}_2$ $\cdots$ $\mathbf{r}_1 \cdot \mathbf{p}_M$
$\mathbf{r}_2 \cdot \mathbf{p}_1$ $\mathbf{r}_2 \cdot \mathbf{p}_2$ $\cdots$ $\mathbf{r}_2 \cdot \mathbf{p}_M$
Prototype-Text Ensemble Module (Sec. 3.4)
Final Scores
CLIP Visual Encoder
Feature Map
RoI Align
$\mathbf{v}_1$
$\mathbf{v}_2$
$\cdots$
$\mathbf{v}_1 \cdot \mathbf{t}_1$ $\mathbf{v}_1 \cdot \mathbf{t}_2$ $\cdots$ $\mathbf{v}_1 \cdot \mathbf{t}_M$
$\mathbf{v}_2 \cdot \mathbf{t}_1$ $\mathbf{v}_2 \cdot \mathbf{t}_2$ $\cdots$ $\mathbf{v}_2 \cdot \mathbf{t}_M$
$\mathbf{t}_1$ $\mathbf{t}_2$ $\cdots$ $\mathbf{t}_M$
Frozen
$\mathbf{p}_m$ Prototypes
$\mathbf{r}_k$ $\mathbf{v}_k$ Region Embeddings
$\mathbf{t}_m$ Text Embeddings

Figure 4. **The inference pipeline of ProDet.** To enhance the accuracy for unseen categories, we use a frozen CLIP model for additional classification support. The final classification scores are obtained by ensembling prototype-based scores and text-based CLIP scores.

[8, 15, 18, 35]. It divides the 80 categories of COCO into 48 base categories and 17 unseen(novel) categories, with 15 categories removed due to the lack of synsets in the WordNet [5] hierarchy. Following standard practices, we report results under the generalized detection setting. The primary evaluation metric on OV-COCO is AP50 of novel categories. The **OV-LVIS** dataset is constructed on the larger-scale LVIS dataset [9], which encompasses a diverse collection of 1203 categories. Following the methodology of ViLD [8], we designate the 866 common and frequently occurring classes in the LVIS dataset as base categories, while treating the remaining 337 rare classes as unseen categories. The primary evaluation metric for OV-LVIS is the mAP of rare categories, denoted as mAP$_r$.

4.2. Implementation Details.

In the 3.2 section, we set the number of points selected for calculating the average point N to 1, 10, and 100, respectively, to obtain three average points for each cross-attention map. Subsequently, in the centroid filtering process, we set the number of final key points T to 3, selecting only the top three centroids with the highest number of points within the cluster as the final key points. For image generation, we use SDXL from stable diffusion series to generate sample images for unseen categories. On a single NVIDIA GeForce RTX 4090 GPU, generating 100 sample images takes 25 minutes, generating key points for these 100 images requires an additional 12 minutes, and performing key-point-guided localization on the 100 images takes 4 minutes.

In the 3.1 module, we choose Faster R-CNN [22] with ResNet50 [10] backbone for experiments on OV-COCO. For experiments on OV-LVIS, we use ViT-L/14 [4] as the backbone. In contrast to other methods for constructing pseudo-labels (e.g., VL-PLM[34], CORA[30]), which jointly train detectors using ground-truth base-category annotations combined with additional pseudo bounding box labels, our approach trains the detector exclusively on ground-truth-labeled examples of base categories. Subsequently, we freeze the detector and use the generated box labels to fine-tune a set of prototypes for unseen categories. This design avoids retraining the detector when new categories are encountered while ensuring that our approach adheres to a strict open-vocabulary setting, wherein prior knowledge of new classes remains unknown during the detector training phase. During inference, we use the pre-trained ViT-B-16 variant of CLIP as the frozen CLIP model. We set the temperature $\tau = 1.0$ in Equation (1) and $\mu = 0.01$ in Equation (3). For the final classification score, we use $\alpha = 1.0$ and $\beta = 0.7$ in Equation (4).

4.3. Benchmark Results.

Table 1 presents our results on the OV-LVIS and OV-COCO datasets. On the OV-LVIS benchmark, ProDet outperforms SoTA methods in mAP$_r$, without pre-training or joint training on additional human-annotated data. On the OV-COCO benchmark, ProDet achieves the second-best performance in AP$_{novel}$ and the best performance in AP$_{all}$ among existing methods. Compared with the leading method

Method	Detector backbone (# Params.)	Require pre-training	Extra human-annotated data			OV-LVIS	OV-COCO	
			Captions	Image-text pairs	Bbox annotations	mAP$_r$	AP$_{novel}$	AP$_{all}$
ViLD [8]	RN50 (25M)	✗	✗	✗	✗	16.6	27.6	51.3
Rasheed [1]	RN50 (25M)	✗	✗	✗	✓	21.1	36.9	51.5
PromptDet [6]	RN50 (25M)	✗	✗	✓	✗	21.4	26.6	50.6
BARON [26]	RN50 (25M)	✗	✗	✗	✓	22.6	42.7	51.7
OVR-CNN [32]	RN50 (25M)	✓	✓	✗	✗	-	22.8	39.9
Detic [36]	RN50 (25M)	✗	✓	✗	✗	24.6	27.8	45.0
MEDet [3]	ViT-B/32 (86M)	✗	✓	✗	✗	22.4	32.6	49.4
RegionCLIP [35]	RN50x4 (102M)	✓	✗	✓	✗	22.0	39.3	55.7
CORA [30]	RN50x4 (102M)	✗	✓	✗	✗	28.1	43.1	56.2
F-VLM [15]	RN50x64 (420M)	✗	✗	✗	✗	32.8	28.0	39.6
CFM-ViT [13]	ViT-L/16 (307M)	✓	✗	✓	✗	33.9	34.1	46.0
CoDet [13]	EVA02-L (304M)	✗	✗	✓	✗	37.0	30.6	46.6
DITO [12]	ViT-L/16 (307M)	✓	✗	✓	✗	38.4	40.8	50.3
DITO [12]	ViT-L/16 (307M)	✓	✗	✓	✓	-	**46.1**	54.2
ProDet (Ours)	ViT-L/14 (303M)	✗	✗	✗	✗	**40.6**	44.9	**56.7**

Table 1. **Comparison with state-of-the-art methods on the OV-LVIS and OV-COCO benchmarks.** Our method outperforms SoTA methods in mAP$_r$ and AP$_{all}$ without annotating large volumes of image-text pairs or joint training on additional human-annotated datasets.

Frozen CLIP	Prototype	AP$_{novel}$	AP$_{all}$
✓	✗	23.9	50.6
✗	✓	41.5	55.6
✓	✓	**44.9**	**56.7**

Table 2. **Ablation study on the contribution of the frozen CLIP model and the prototypes on OV-COCO.** This table evaluates the respective contributions of the frozen CLIP model and the learned prototypes during inference. Utilizing prototypes alone substantially outperforms directly using the frozen CLIP model. Furthermore, integrating predictions from the frozen CLIP model with the learned prototypes results in even greater performance improvements.

DITO [12], the primary reason for the underperformance of ProDet lies in its lack of external box annotations during the detector training phase. In fact, ProDet outperforms the version of DITO that was not trained on additional Objects365 [25] box annotations with a clear margin of +4.1 in AP$_{novel}$ and +6.4 in AP$_{all}$. It is noteworthy that most other methods either construct bounding box labels for pre-training on additional image-text pairs or incorporate extra human-annotated datasets into the training set for joint training. However, these methods require high annotation costs and demand significant computational resources. In contrast, ProDet requires only ground-truth base-category annotations to train the detector. During the fine-tuning phase, ProDet can rapidly adapt to new categories by fine-tuning a set of prototypes for unseen categories on a small set of automatically generated samples. Since the generated data is highly lightweight, our method significantly reduces the consumption of training resources.

4.4. Visualization and Analysis

In Figure 5, we visualize the detection results of ProDet on the COCO dataset. Objects within red boxes belong to unseen categories, while those within green boxes belong to base categories. As shown in Figure 5, ProDet detect unseen and base categories with high-quality box predictions.

4.5. Ablation Study

Ablation study on the contribution of each module. As shown in Table 2, we observe that using prototypes alone significantly outperforms directly using the frozen CLIP model. Additionally, combining the predictions from the frozen CLIP model with the learned prototypes yields even better results. We further adjusted the weighting of the detection scores and the CLIP scores in Equation (4) by varying the values of α and β. Experimental results demonstrate that the model achieves optimal AP$_{novel}$ performance when $\alpha = 1.0$ and $\beta = 0.7$. We hypothesize that for base categories, where prototypes are trained on ground-truth annotations, relying exclusively on base prototypes yields significantly better outcomes than utilizing the frozen CLIP model alone. Conversely, for unseen categories, where prototypes are trained on synthetic data with limited diversity, incorporating text guidance from the frozen CLIP model enhances performance.

Ablation study on the number of generated images and filtering strategies. As shown in Table 3, detection ac-

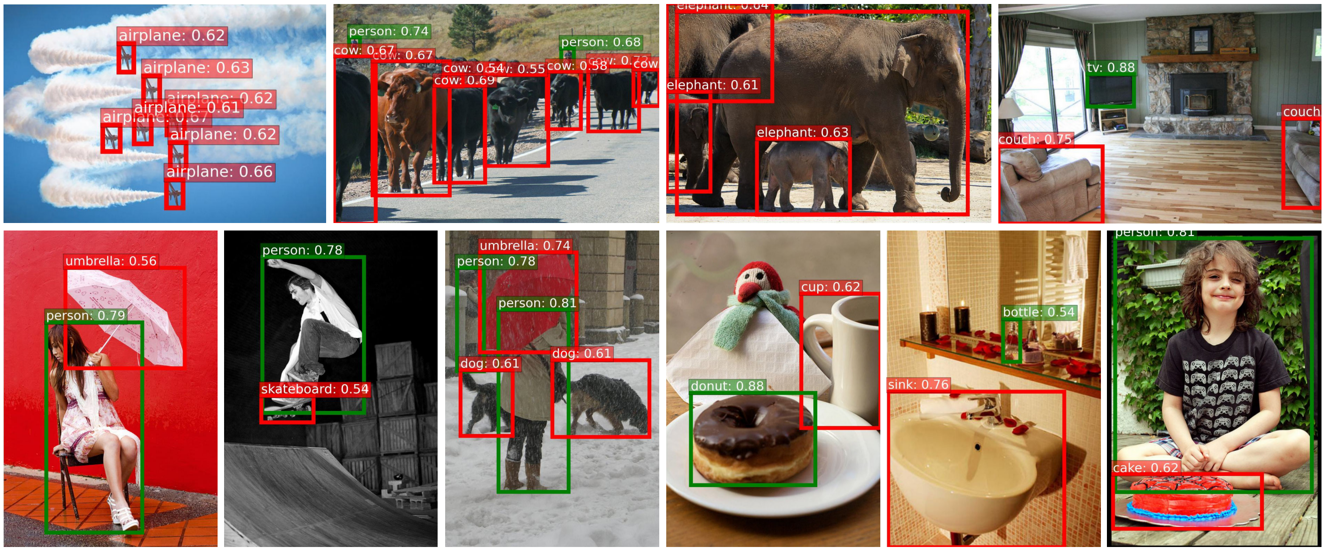

Figure 5. Visualization of the detection results of ProDet on OV-COCO. Objects within red bounding boxes belong to unseen categories, while those within green bounding boxes belong to base categories.

Image number per unseen category	Centroid filtering	Mask filtering	AP_{novel}
10	✗	✗	35.3
100	✗	✗	39.0
400	✗	✗	39.2
100	✓	✗	42.3
100	✗	✓	41.6
100	✓	✓	**44.9**

Table 3. **Ablation study on the number of generated images and filtering strategies.** The top three rows illustrate the effect of varying the number of generated images per unseen category. The bottom three rows demonstrate that ProDet achieves the highest AP_{novel} when employing both centroid and mask filtering.

Classification loss	AP_{novel}	AP_{all}
Cross entropy loss	40.1	55.6
Focal loss	**44.9**	**56.7**

Table 4. **Ablation study on classification loss.** By employing focal loss, ProDet improved the AP_{novel} by 4.8 compared to using cross-entropy loss.

curacy reaches a saturation point when the number of generated images per category approaches approximately 100. Moreover, the application of centroid filtering and mask filtering, as detailed in Section 3.2, enhances AP_{novel} by 3.1 and 2.4, respectively. Centroid filtering mitigates the influence of outlier data by ensuring that key points are positioned in critical regions of the target objects, while mask filtering prevents poorly generated masks from adversely impacting prototype learning.

Ablation study on the classification loss. As illustrated in Table 4. By adopting focal loss, we obtained superior results compared to those achieved with cross-entropy loss. The focal loss integrates decaying factors α and γ to mitigate the dominance of easily classified samples during training, while simultaneously addressing the class imbalance between positive and negative instances.

5. Conclusion

We present ProDet, a new method that learns prototype-based region-word alignment for open-vocabulary detection. Firstly, we use the stable diffusion model to generate sample images for each unseen category. Then we propose the key-point-guided localization module to automatically obtain bounding box labels for the target objects in the generated images. Finally, with these automatically generated images and bounding box labels, we can train a set of prototypes from various instances of unseen categories. Our method delivers highly competitive performance, on the OV-LVIS benchmark, ProDet outperforms SoTA methods in mAP_r, without pre-training or joint training on additional human-annotated data. On the OV-COCO benchmark, ProDet achieves the second-best performance in AP_{novel} and the highest performance in AP_{all} among existing methods. We hope that the simple design and strong performance make ProDet an alternative approach for detecting unseen categories instead of annotating large volumes of image-text pairs for pre-training or joint training on additional datasets.

References

[1] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 7

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

[3] Peixian Chen, Kekai Sheng, Mengdan Zhang, Mingbao Lin, Yunhang Shen, Shaohui Lin, Bo Ren, and Ke Li. Open vocabulary object detection with proposal mining and prediction equalization. *arXiv preprint arXiv:2206.11134*, 2022. 7

[4] 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. 6

[5] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 6

[6] 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 *European conference on computer vision*, pages 701–717. Springer, 2022. 7

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

[8] Xiuye Gu, Tsung-Yi Lin, WenChuan Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *Learning,Learning*, 2021. 2, 6, 7

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[11] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation, 2023. 3

[12] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Detection-oriented image-text pretraining for open-vocabulary detection. *arXiv preprint arXiv:2310.00161*, 2023. 5, 7

[13] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15602–15612, 2023. 7

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4

[15] 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. 2, 6, 7

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[18] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36, 2024. 6

[19] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[20] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv,Cornell University - arXiv*, 2021. 1, 2, 3

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1137–1149, 2017. 6

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[24] 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

[25] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7

[26] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15254–15264, 2023. 7

[27] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2

[28] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6117–6125, 2024. 2

[29] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models, 2024. 3

[30] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching, 2023. 6, 7

[31] Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *International Journal of Computer Vision*, pages 1–20, 2024. 3

[32] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 7

[33] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 3

[34] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, pages 159–175. Springer, 2022. 2, 6

[35] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6, 7

[36] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision, 2022. 2, 7