Region-centric Image-Language Pretraining for Open-Vocabulary Detection

Dahun Kim   Anelia Angelova   Weicheng Kuo

Google DeepMind

Abstract. We present a new open-vocabulary detection approach based on region-centric image-language pretraining to bridge the gap between image-level pretraining and open-vocabulary object detection. At the pretraining phase, we incorporate the detector architecture on top of the classification backbone, which better serves the region-level recognition needs of detection by enabling the detector heads to learn from large-scale image-text pairs. Using only standard contrastive loss and no pseudo-labeling, our approach is a simple yet effective extension of the contrastive learning method to learn emergent object-semantic cues. In addition, we propose a shifted-window learning approach upon window attention to make the backbone representation more robust, translation-invariant, and less biased by the window pattern. On the popular LVIS open-vocabulary detection benchmark, our approach sets a new state of the art of 37.6 mask AP_r using the common ViT-L backbone and public LAION dataset, and 40.5 mask AP_r using the DataComp-1B dataset, significantly outperforming the best existing approach by $+3.7$ mask AP_r. On the COCO benchmark, we achieve very competitive 39.6 novel AP without pseudo labeling or weak supervision. In addition, we evaluate our approach on the transfer detection setup, where it demonstrates notable improvement over the baseline. Visualization reveals emerging object locality from the pretraining recipes compared to the baseline.[1]

1 Introduction

Understanding and localizing all objects in the visual world is a key problem in computer vision, crucial for applications like self-driving cars. Traditional object detectors rely on costly human annotations, which limits scalability. Open-vocabulary detection (OVD) [41] addresses this by pretraining on large-scale image-text data (*e.g.* CLIP [28]), enabling detectors to identify objects based on user-provided text queries unavailable during training.

Most existing open-vocabulary detection works assume the pretrained CLIP backbone is given, and focus on techniques such as knowledge distillation [5, 10], weak supervision [46], pseudo labeling [13, 29, 43, 44], and frozen backbone application [18], using the pretrained backbone. Consequently, during detection finetuning, the detector heads often need to be trained from scratch. This tends to result in sub-optimal generalization because the detector heads are trained on

---

[1] project page: github.com/google-research/google-research/tree/master/fvlm/dito
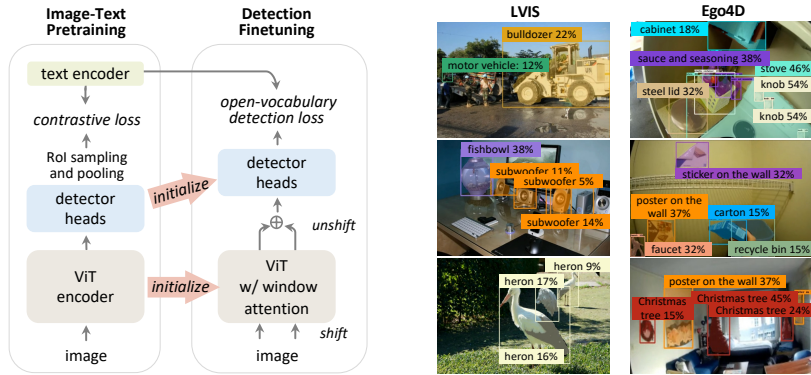
Fig. 1: DITO overview (left): Existing image-text pretraining methods for open-vocabulary detection [16,17] update only the vision transformer backbone and finetune detector heads from scratch. We propose to pretrain both backbone and detector heads directly from the large-scale image-text paired data, without a need for pseudo labeling or box annotations. In open-vocabulary detection finetuning, we introduce a simple shifted-window learning method to produce more robust representations from the pretrained vision transformer. **DITO prediction (right):** LVIS results only show the novel categories (*e.g.* , *bulldozer, fishbowl, subwoofer, heron*). While Ego4D is a real-world and out-of-distribution data, many unseen objects are detected (*e.g.* , *steel lid, sticker on the wall, recycle bin*). Best viewed with zoom in.

the limited vocabulary of detection datasets, while only the backbone contains the knowledge of open-vocabulary concepts.

Several studies have integrated detection models into CLIP pretraining. For instance, RegionCLIP [44] employs an off-the-shelf detector during CLIP pretraining to generate pseudo region-text labels. However, their pretraining only updates the image backbone, with the detector heads exclusively trained on detection data. These pseudo labeling-based pretraining methods [25, 44] require multi-stage training and handcrafted processing of pseudo labels, increasing complexity and training costs. Similarly, approaches such as GLIP [19,42], Grounding DINO [22], CoDet [24] and DetCLIP [38, 39] integrate detector architectures in CLIP training, but their joint training demands additional detection and visual grounding datasets and complex multitask learning setups.

In this paper, we propose a simple yet effective solution, Region-centric Pretraining approach, which incorporates detector modules into CLIP pretraining without the need for pseudo labeling or box annotations. This method involves generating random box regions across feature pyramid levels, followed by feature pooling over these regions. Subsequently, an image-text contrastive loss is applied, encouraging text-aligned region features to contribute more to the whole image representation. Our approach not only facilitates the warmstarting of detector heads in finetuning, but also leads to emergent representations with more localized semantic information compared to the baseline CLIP backbone, as demonstrated in our experiments. Compared to pseudo-labeling techniques [35,44–46], our approach can be viewed as an extension of contrastive

models to bypass the need for offline object proposal generation, which could be cumbersome for large-scale image-text data.

In addition, we focus on the Vision Transformer (ViT) based CLIP models. These CLIP ViTs excel in zero-shot recognition, but are less commonly integrated into open-vocabulary detectors compared to ConvNets [18, 36, 44]. Pretraining ViT backbones on lower resolutions and then adapting them to higher-resolution detection images poses challenges, including increased computational demands and potential disruption of locality in pretrained features. Although the windowed attention technique of ViTDet [17, 20] reduces computation and preserves locality, it introduces bias from fixed window patterns. To address this, we propose Shifted-Window Learning (SWL) approach to enhance information mixing across fixed windows and mitigate grid pattern bias. Unlike the Swin Transformer [23], which applies shifted window layer by layer, SWL applies shifted windowed attention as a separate forward pass using the same ViT backbone. This simple strategy enhances the windowed attention representation when applied low-res trained ViT backbone to high-res images, ensuring compatibility with vanilla ViT backbones pretrained without shifted windows. SWL also better preserves the open-vocabulary knowledge of pretrained features compared to full-attention ViT perhaps due to its emphasis on local cues.

Incorporating region-centric pretraining and shifted-window learning, our approach DITO (Detection-aligned Image-Text pretraining for Open-vocabulary detection) serves to narrow the gap between image-text pretraining and open-vocabulary detection, improving generalization. The best DITO model achieves 37.6 mask AP_r on LVIS benchmark, surpassing the previous best by $+3.7$ AP_r. It achieves the state-of-the-art 40.5 mask AP_r when pretrained on the DataComp-1B dataset. On the COCO benchmark, DITO achieves a very competitive 39.6 novel AP without relying on pseudo-labels or joint training.

2   Related Work

Open-vocabulary detection.   Conventional closed-set object detectors perform well but are limited by vocabulary size. Open-vocabulary detection, inspired by the strong zero-shot capabilities of Vision-Language Models (VLMs) like CLIP [28], has made notable progress. Various techniques have been explored, including knowledge distillation [10], prompt optimization [5], and the use of pretrained CLIP backbones with new detection heads [17,18,26,36]. Many top-performing methods [6,8,19,25,44] rely on pseudo labeling to prevent catastrophic forgetting during detection finetuning, but these methods often require multi-stage training [35, 44] or extra steps to generate high-quality labels [25]. In contrast, our approach focuses on both upstream image-text pretraining and downstream open-vocabulary detection without using pseudo labeling, leading to significant improvements.

Region-aligned Vision-Language Models.   Building on advances in aligning image-text representations [28, 31], several studies have integrated region-level alignment into CLIP pretraining. For example, RegionCLIP [44] learns region-word alignment through pseudo region-text pairs but relies on an off-

the-shelf detector for pseudo labeling and focuses only on training the image backbone. GLIP [19, 42], Grounding DINO [22], DetCLIP [38, 39], CoDet [24] integrate detector architectures into CLIP training to align regions with words, but they require additional visual grounding annotations, leading to complex multitask learning. To our knowledge, our approach is the first to incorporate detector modules in image-text pretraining without relying on box annotations.

Vision Transformers in open-vocabulary detection. While ViT-based CLIP models excel in zero-shot recognition, their adaptation to open-vocabulary detection is less explored compared to ConvNet-based CLIP models. For instance, OWL-ViT [25, 26] finetunes ViT on higher-resolution detection images with global attentions, but the full attention is computationally intensive. RO-ViT [17] introduces region-aware positional embeddings to aid in detection finetuning and uses windowed attention from ViTDet [20]. Our paper proposes shifted-window learning to enhance information mixing across fixed windows and reduce window-induced bias while preserving the locality structure of pretrained features, thereby improving open-vocabulary detection.

Self-supervised pretraining for visual tasks. Self-supervised learning has emerged as a promising paradigm to learn object features for complex visual tasks such as detection, given the challenge of scaling up human annotations. Most relevant direction is contrastive learning, where the contrastive samples can take the forms of augmented images [3], sliding windows [37], object proposals [34], or point samples [1]. Some alternative strategies like pseudo-labeling [45] and pixel reconstruction [11] have also proven effective. While the majority of these methods have focused on learning from images without textual context, and applying to closed-vocabulary detection, DITO leverages large image-text data to tackle the more demanding open-vocabulary detection task, without a need for offline proposal generation [34, 45].

3   Method

We address the problem of open-vocabulary object detection. At training time, the model can access the class and box labels of base categories (C_B). At test time, the model is tasked with detecting objects from a set of novel categories (C_N) not present in the training set. To achieve this, we leverage pretrained vision and language models (VLMs) as in prior studies [10, 17, 18, 44].

More specifically, we leverage recent advances in ViT-based detectors [17, 20] for their promising results. However, instead of solely taking the pretrained ViT backbone, we demonstrate how to enhance the VLMs by integrating detector heads into the CLIP pretraining process, referred to as Region-centric image-text Pretraining (REP). Additionally, for detection finetuning, we propose a Shifted-Window Learning (SWL) to enhance the adaptation of the pretrained ViT to open-vocabulary detection task. By combining these approaches, DITO achieves significant improvements in open-vocabulary detection over prior arts.

Fig. 2: DITO method. Region-centric image-text pretraining (left): We train the detector heads (e.g. FPN [20, 21], Faster RCNN head [30], and RPN [30]) upon a ViT encoder backbone with multi-level image-text contrastive loss to bridge the gap between image-text pretraining and open-vocabulary detection. Shifted-window learning for detection (right): We roll the image and combines the shifted features with the original features to mitigate the bias of windowed attention grid [20], and produce more robust semantic representation.

3.1 Preliminaries

Baseline. Our baseline approach is RO-ViT [17], a state-of-the-art ViT-based method for open-vocabulary detection. RO-ViT introduces CLIP pretraining with a new positional embedding scheme called cropped positional embedding (CPE). CPE involves randomly cropping and resizing the standard whole-image positional embedding during pretraining to enhance the ViT's generalization onto region-level recognition task and higher resolution detection inputs downstream. For detection finetuning, it adopts ViTDet [20] architecture, initialized with the pretrained ViT backbone. In the following, we describe the image-text pretraining and the downstream open-vocabulary detection.

Image-text pretraining. We adopt dual-encoder CLIP pretraining widely used in existing works [28, 31]. The image embeddings $\{v\}$ and text embeddings $\{l\}$ are the average-pooled outputs from the image and text encoders, respectively. As in previous works, we compute the dot product of the embeddings in batch B, and scale it by a learnable temperature τ before applying the InfoNCE loss [27, 28]. Mathematically, the image-to-text (I2T) loss can be expressed as:

$$L_{\text{I2T}} = -\frac{1}{B} \sum_{i=1}^{B} \log\left(\frac{\exp(v_i l_i/\tau)}{\sum_{j=1}^{B} \exp(v_i l_j/\tau)}\right). \tag{1}$$

The text-to-image (T2I) loss is symmetrical by exchanging the inner/outer summation loops. The total contrastive loss L_{con} is obtained by $L_{con} = (L_{\text{I2T}} + L_{\text{T2I}})/2$. As mentioned above, we adopt the cropped positional embeddings (CPE) following [17].

Open-vocabulary detection finetuning.   At the fine-tuning stage, our detection finetuning recipe follows previous studies [10,17,18,41]. During the training phase, we use the RoI-Align [12] feature as the detection embedding for each detected region. We replace the fixed-size classifier layer with the text embeddings of base categories (C_B). The detection score p_i is determined by calculating the cosine similarity between the region embedding r_i and text embeddings of base categories (C_B) followed by a softmax operation. We prepend an additional background class embedding to C_B and use the term "background" to represent the background category. Any proposals that do not match to any base category annotations are treated as background during training. It is important that the text embeddings are computed from the same text encoder as from the image-text pretraining. During testing, we expand the text embeddings to include the novel categories ($C_B \cup C_N$), resulting in ($C_B \cup C_N + 1$) categories including the background. We calculate the detection scores (p_i) as the cosine similarity between the region embeddings (r_i) and the expanded text embeddings. Apart from the detection embedding (r_i), we extract the VLM embedding [18] of region i by RoI-Align at the last ViT backbone feature map. The VLM score (z_i) is calculated as the cosine similarity with the text embeddings of the combined categories ($C_B \cup C_N$).

To compute the open-vocabulary detection score (s_i^{ens}), we ensemble the detection and VLM scores by geometric means [10,18]. The formula is as follows:

$$s_i^{\text{ens}} = \begin{cases} z_i^{(1-\alpha)} \cdot p_i^{\alpha} & \text{if } i \in C_B \\ z_i^{(1-\beta)} \cdot p_i^{\beta} & \text{if } i \in C_N \end{cases} \tag{2}$$

Here, α, β are floats $\in [0,1]$ that control the weighting of base versus novel categories. The background score comes from the detection score (p_i) alone, because we observe the VLM score of "background" class is often less reliable.

3.2   Region-centric Image-Text Pretraining

Standard image-text pretraining uses classification architectures (*e.g.* ViT backbone followed by global pooling) as the language supervision occurs at the image level rather than the region level. Subsequently, for downstream detection, new detection heads are introduced and trained from scratch on a limited set of detection categories [16–18]. To fully utilize the knowledge embedded in large-scale image-text data, we propose Region-centric Pretraining (REP) which integrates the detector modules during the CLIP pretraining phase. Specifically, given access to image-text paired data but lacking box labels, our pretraining focuses on the *region-recognition pathway* of a detector, encompassing components like the backbone, FPN [20,21], RoI-Align [12], RPN-objectness [30], and Faster RCNN-classifier [30]. Consequently, the detector heads can be warm-started from the knowledge of large image-text data, thereby improving the generalization capability. To our knowledge, we are the first to integrate detector modules in image-text pretraining without box labels, and our experiments demonstrate clear benefits of our approach in open-vocabulary detection.

Detector head learning from random regions. Fig. 2 (left) illustrates our region-centric pretraining system. Following existing works [16,17], we adopt SimpleFPN [20] and Faster R-CNN [30] models to remove the detector differences and study the benefits of our region-centric pretraining. Specifically, the multi-level feature pyramid is computed from the ViT backbone. Then, RoI-Align [12] and Faster R-CNN classifier head are applied to these feature maps to match the classification pathway in pretraining with the region-recognition pathway in detection finetuning (see Table 4b for ablations).

For each level i of the feature pyramid, we randomly generate n_i box regions uniformly over the image by sampling the box size $h, w \sim Uniform(0.2, 0.5)$ and aspect ratio $h/w \sim Uniform(0.5, 2.0)$. The n_i value is set proportional to the size of the i-th feature map so that larger feature map would be covered by more regions. We extract the RoI-features of each region by RoI-Align, and feed them through the region classifier head [30] to obtain the RoI embeddings.

Multi-level image-text supervision. After computing the RoI embeddings across pyramid levels, we perform a max pooling over the RoI embeddings per-level to obtain an image embedding for each pyramid level. Intuitively, max pooling allows the representation to focus on salient regions and discriminative features, thereby learning region-level information without explicit supervision. Then we apply the standard image-text contrastive loss (see Eqn. (1)) on each feature level separately, which aids the learning of rich semantic and object knowledge within every feature map (see Table 4c for ablations). The losses from all levels are weighted equally and summed together. Without explicit region-level supervision, the max pooling over regions encourages the more salient, text-aligned region features to contribute more to the whole image representation in the contrastive loss. Our experiments show emergent region-text alignments from the multi-level training, where the feature maps possess more localized semantic information compared to the baseline CLIP [17] backbone (see Fig. 3).

Region proposal network learning. Fig. 3 shows that the learnt multi-level representations exhibit localized semantics well-aligned with the text query, *i.e.* the salient regions have higher similarity with the text relative to the back-ground. Motivated by this, we employ the multi-level visual-text similarity as a supervisory signal for training the Region Proposal Network (RPN) [30]. Specif-ically, we compute the cosine similarity between each RoI embedding and the text, and use it as the target RPN score for the center location of each RoI. Any negative dot product value is mapped to zero to keep the target score in range $[0, 1]$. We use L1 regression loss and set the loss weight equal to the multi-level contrastive loss (see supplementary). The losses are propagated only through the RoI centers and other pixels are ignored. Note that the box regression of the RPN is *not* trained here but learnt later through the detection finetuning, as we only use the image-text paired data without any box annotations.

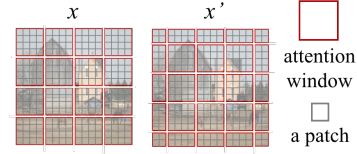### 3.3 Shifted-Window Learning for Detection

The CLIP ViT backbones are typically pretrained on lower resolutions (*e.g.* 224×224) and later adapted to higher resolution detection images (*e.g.* 1024×1024).

Algorithm 1 Shifted Window backbone

---

x: image patch tokens + positional embeddings.
M: attention window size.
s = M // 2
y = `forward_vit_with_win_attn`(x, M)
x' = `np.roll`(x, `shift=`[s, s], `axis=`[0, 1])   # shift
y' = `forward_vit_with_win_attn`(x', M)
y'' = `np.roll`(y', `shift=`[-s, -s], `axis=`[0, 1])   # unshift
return (y + y'') / 2   # SWL backbone output

While detection tasks benefit from global information, directly applying the pretrained ViT to high-resolution inputs is computationally intensive and can disrupt the locality structure of the pretrained lower-res features. Windowed attention in ViTDet [20] helps reduce computation but introduces bias from the fixed-size grid pattern. To address this, we propose the Shifted-Window Learning (SWL) approach, which improves information mixing across fixed windows and mitigates grid pattern bias.

Network architecture. Fig. 2 (right) and Algorithm 1 describe the SWL algorithm. The standard ViT consists of a patchifying layer and a set of transformer layers. After feeding the image through the patchifying layer, we obtain a feature map x of shape (h, w, d). This feature map x is fed through the rest of the ViT with windowed attention layers on a grid $K \times K$, and L global attention layers evenly spaced throughout the ViT (where $L = 4$) following [20], resulting in output y. In parallel, we create another copy of x, which is rolled along both h and w axes by s pixels. The elements that roll beyond the last position are reintroduced from the first. We carefully design the attention masks such that the rolled around patches would not attend to the patches on the other side of the image (see the right figure of Algorithm 1). The shift size s is set as the half of the attention window size M (*i.e.* $s = M//2$). Empirically, the window size $M = 16$ equals the image size (*e.g.* 1024) divided by the product of patch size (*e.g.* $P = 16$) and the grid size (*e.g.* $K = 4$). The shifted feature map x' is then processed through the rest of the ViT in the same manner, resulting in output y' in the same shape (h, w, d) as y. We then unshift y' and combine it with y by averaging. We apply the above shifted window operations in detection finetuning and inference times.

Comparison with Swin Transformer. Compared to the Swin Transformer [23], we apply the shifted-window ideas as separate forward passes, while Swin Transformer applies similar ideas in an alternating manner layer by layer. Our approach requires no change to the vanilla transformer architecture and is compatible with any ViT backbones pretrained without shifted windows (*e.g.* [28]), whereas Swin Transformer requires specialized pretraining on the same architecture. Compared to the full-attention ViT [4], we observe that shifted-window ViT taps more effectively into the semantic knowledge of pretrained backbone than full global attention, perhaps because the window helps the model focus more on local cues and ignore the noises farther away.

3.4 Distillation from Frozen ViT Backbone

While the ViT backbone adapts to the detection tasks, it may lose some of its pretrained open-vocabulary knowledge. Therefore, we propose a simple distillation approach which uses a separate frozen ViT to teach the finetuned ViT backbone during the detection finetuning. We use a cosine distance loss that aligns the RoI-Align embeddings extracted from the feature maps of both backbones. The cosine distance is computed for each RoI then averaged over all RoIs. Empirically, we find it advantageous to add a 1×1 Conv projection layer to the finetuned ViT backbone before the RoI-Align, allowing some flexibility in distillation when jointly trained with other detection objectives. The auxiliary distillation loss is then added to other detection losses, with a loss weight $\gamma = 1$ (see supplementary). At inference, the ViT backbone features after the projection are used to compute the region VLM score z_i (Sec. 3.1). It is worth noting that the frozen ViT backbone is only used during training for distillation purposes and is removed at inference.

While previous studies [2, 10] have utilized knowledge distillation from the CLIP models, their teacher CLIP models mostly operate on image crops in an offline process, thus needing multiple forward passes through the backbone for all RoIs. In contrast, our distillation operates efficiently on RoIs cropped from dense feature maps in a single forward pass during the detection finetuning with minimal overhead.

4 Experimental Results

Baseline reproduction. As discussed in Sec. 3.1, our baseline method is RO-ViT [17], a leading ViT-based approach for open-vocabulary detection. RO-ViT introduces cropped positional embedding (CPE) in CLIP pretraining to enhance the generalization of the pretrained ViT onto the downstream detection. Additionally, it adopts global average pooling of the ViT features instead of CLS-token pooling, which simplifies the adaptation onto higher resolution inputs, and the extraction of region features (*e.g.* RoI-Align) at the final layer. We reproduce the CLIP pretraining of RO-ViT [17] using the widely adopted LAION-2B dataset [31], as the OpenAI CLIP's WIT [28] and ALIGN [14] datasets are not publicly available. We trained the CLIP models from scratch, following the same pretraining protocol and hyperparameters, including 500k iterations and 16k batch size (*i.e.* 8B samples seen in training), 224×224 image size, global average pooling, and cropped positional embedding. We use the standard InfoNCE loss (Eqn. (1)) instead of their focal constrastive loss [17]. The following compares our reproduced CLIP with OpenAI CLIP [28] and OpenCLIP [31]:

method	dataset	# samples seen	backbone	ImageNet Top-1 Acc.
CLIP [28]	WIT-400M	13B	ViT-L/14	75.5
OpenCLIP [31]	LAION-2B	32B	ViT-L/14	75.2
RO-ViT CLIP [17] *our repro.*	LAION-2B	8B	ViT-L/16	73.9

On the zero-shot ImageNet classification benchmark, the LAION-2B pretraining matches or slightly underperforms the OpenAI CLIP with WIT-400M.

method	pretraining model	pretraining data	detector backbone	w/ pseudo box labels	**mask** $\mathbf{AP}_r$	mask AP
ConvNet based:						
OV-DETR [40]	ViT-B/32	CLIP-400M	R-50	-	17.4	26.6
Kaul *et al.* [15]	ViT-B/32	CLIP-400M	R-50	-	19.3	30.6
DetPro-Cascade [5]	ViT-B/32	CLIP-400M	R-50	-	20.0	27.0
Rasheed [29]	ViT-B/32	CLIP-400M	R-50	-	21.1	25.9
BARON [35]	ViT-B/32	CLIP-400M	R-50	-	22.6	27.6
CoDet [24]	R-50	CLIP-400M + CC3M	R-50	-	23.4	30.7
EdaDet [32]	R-50	CLIP-400M	R-50	-	23.7	27.5
VL-PLM [43]	ViT-B/32	CLIP-400M	R-50	✓	17.2	27.0
PromptDet [6]	ViT-B/32	CLIP-400M	R-50	✓	21.4	25.3
OADB [33]	ViT-B/32	CLIP-400M	R-50	✓	21.7	26.6
RegionCLIP [44]	R-50x4	CLIP-400M + CC3M	R-50x4	✓	22.0	32.3
CORA [36]	R-50x4	CLIP-400M	R-50x4	✓	22.2^{bx}	-
Detic-CN2 [46]	ViT-B/32	CLIP-400M + INet-21K	R-50	WS	24.6	32.4
ViLD-Ens [10]	EffNet-B7	ALIGN-1.8B	EffNet-B7	-	26.3	29.3
F-VLM [18]	R-50x64	CLIP-400M	R-50x64	-	32.8	34.9
ViT based:						
OWL-ViT [26]$^{\text{O+VG}}$	ViT-L/14	CLIP-400M	ViT-L/14	-	25.6^{bx}	34.7^{bx}
OWL-ViT v2 [25]$^{\text{O+VG}}$	ViT-L/14	WebLI-10B	ViT-L/14	✓	45.9^{bx}	50.4^{bx}
RO-ViT [17]	ViT-B/16	ALIGN-1.8B	ViT-B/16	-	28.0	30.2
RO-ViT [17] †	ViT-L/16	LAION-2B	ViT-L/16	-	32.4	32.9
CFM-ViT [16]	ViT-B/16	ALIGN-1.8B	ViT-B/16	-	28.8	32.0
CFM-ViT [16]	ViT-L/16	ALIGN-1.8B	ViT-L/16	-	33.9	36.6
CFM-ViT [16] *	ViT-L/16	LAION-2B	ViT-L/16	-	33.8	36.4
DITO (ours)	ViT-S/16	LAION-2B	ViT-S/16	-	**26.2**	28.8
DITO (ours)	ViT-B/16	LAION-2B	ViT-B/16	-	**31.5**	32.4
DITO (ours)	ViT-L/16	LAION-2B	ViT-L/16	-	**37.6**	36.2
DITO (ours)	ViT-L/16	DataComp-1B	ViT-L/16	-	**40.5**	38.0

Table 1: LVIS open-vocabulary detection (mask AP). DITO outperforms the best existing approach by +3.7 mask AP_r. WS: uses weak supervision from ImageNet-21K. †: reports LAION-2B results in arXiv version. *: our reproduced result using LAION-2B. O+VG: uses extra Objects365 and Visual Genome data.

Pretraining setup. After the above-mentioned baseline CLIP training, we apply our region-centric pretraining (REP) where we freeze the image and text encoders trained in the first phase and introduce the detector heads. We use the Simple FPN [20], and classification layers of Faster R-CNN and RPN [30], where we replace the batch normalization with layer normalization. At the i-th pyramid level $i \in \{2, 3, 4, 5\}$, we randomly sample $n_i \in \{400, 200, 100, 50\}$ box regions and compute their RoI-Align features. We use a short training cycle of 30k iterations, 4k batch size, 256×256 image size, AdamW optimizer with an initial learning rate of 1e-4 with linear decay. For both phases of image-text pretraining, we use the publicly available LAION-2B [31] dataset.

Detection finetuning setup. As noted in Sec. 3.2, we adopt the ViTDet [20] with SimpleFPN as our detector following the baseline works [16,17]. We follow the same finetuning settings of [17]. Specifically, we train the detector with image size 1024×1024 and use windowed attention in the backbone with grid size 4×4. The learning rate for the backbone is set lower as 0.6× of the detector head layers. We use α=0.3, β=0.65 for score combination in Eqn. (2). The text embedding of each category is calculated as the average over the CLIP prompt templates. We use the batch size 128, the SGD optimizer with momentum 0.9. The initial learning rate and iterations are set to 0.18 and 36.8k for LVIS, and 0.02 and 11.3k for COCO datasets.

method	pretraining model	pretraining data	detector backbone	w/ pseudo box labels	novel AP	AP
ConvNet based:						
OVR-CNN [41]	R-50	CLIP-400M + COCO-Cap	R-50	-	22.8	39.9
ViLD [10]	ViT-B/32	CLIP-400M	R-50	-	27.6	51.3
F-VLM [18]	R-50	CLIP-400M	R-50	-	28.0	39.6
OV-DETR [40]	ViT-B/32	CLIP-400M	R-50	-	29.4	52.7
CoDet [24]	R-50	CLIP-400M + COCO-Cap	R-50	-	30.6	46.6
PromptDet [6]	ViT-B/32	CLIP-400M	R-50	✓	26.6	50.6
XPM [13]	R-50	CLIP-400M	R-50	✓	27.0	41.2
OADB [33]	ViT-B/32	CLIP-400M + COCO-Cap	R-50	✓	30.0	47.2
VL-PLM [43]	ViT-B/32	CLIP-400M	R-50	✓	34.4	53.5
RegionCLIP [44]	R-50x4	CLIP-400M + CC3M + COCO-Cap	R-50x4	✓	39.3	55.7
EdaDet [32]	R-50	CLIP-400M	R-50	✓	40.2	52.5
CORA [36]	R-50x4	CLIP-400M	R-50x4	✓	41.7	43.8
Detic-CN2 [46]	ViT-B/32	CLIP-400M + INet-21K	R-50	WS	27.8	45.0
ViT based:						
RO-ViT [17]	ViT-B/16	ALIGN-1.8B	ViT-B/16	-	30.2	41.5
RO-ViT [17]	ViT-L/16	ALIGN-1.8B	ViT-L/16	-	33.0	47.7
RO-ViT [17] *	ViT-L/16	LAION-2B	ViT-L/16	-	33.3	47.9
CFM-ViT [16]	ViT-B/16	ALIGN-1.8B	ViT-B/16	-	30.8	42.4
CFM-ViT [16]	ViT-L/16	ALIGN-1.8B	ViT-L/16	-	34.1	46.0
CFM-ViT [16] *	ViT-L/16	LAION-2B	ViT-L/16	-	34.3	46.4
DITO (ours)	ViT-S/16	LAION-2B	ViT-S/16	-	**32.3**	44.4
DITO (ours)	ViT-B/16	LAION-2B	ViT-B/16	-	**36.6**	48.8
DITO (ours)	ViT-L/16	LAION-2B	ViT-L/16	-	**39.6**	54.4
DITO (ours)	ViT-L/16	DataComp-1B	ViT-L/16	-	**40.2**	54.6

Table 2: COCO open-vocabulary detection (box AP50). DITO demonstrates a very competitive novel category AP without using pseudo labeling or weak supervision (WS). *: our reproduced result using LAION-2B.

method	backbone	AP	AP_{50}	AP_{75}
Supervised [10]	R-50	25.6	38.6	28.0
ViLD [10]	R-50	11.8	18.2	12.6
DetPro [5]	R-50	12.1	18.8	12.9
BARON [35]	R-50	13.6	21.0	14.5
F-VLM [18]	R-50x16	16.2	25.3	17.5
F-VLM [18]	R-50x64	17.7	27.4	19.1
RO-ViT [17]	ViT-L/16	17.1	26.9	18.5
CFM-ViT [16]	ViT-L/16	18.7	28.9	20.3
DITO (ours)	ViT-L/16	**20.0**	**31.8**	**21.5**

Table 3: Zero-shot transfer detection from LVIS$_{base}$ to Objects365 (box AP). All models are tested on Objects365 dataset following the setup of [10].

4.1 Main Results

LVIS Benchmark. In Table 1, we report the comparison with existing methods on the challenging LVIS benchmark. The 'frequent' and 'common' classes of the dataset belong to the base categories C_B, and the 'rare' classes are the novel categories C_N. The primary metric is the mask AP on rare classes (mask AP_r). The DITO model achieves the performance of 37.6 mask AP_r, which significantly outperforms the state-of-the-art approach RO-ViT [17] with the same ViT-L backbone by +5.1 points using the same pretraining data LAION-2B [31]. We also outperform the state-of-the-art CFM-ViT [16] by +3.6 points. Our best performance sets a new state-of-the-art 40.9 mask AP_r when using DataComp-1B [7] in pretraining. With the ViT-B backbone, DITO maintains a healthy margin of around +2.5 AP_r above existing ViT-B based approaches.

method	AP_r	AP
RO-ViT [17]	32.4	32.9
RO-ViT *our repro.*	32.2	33.0
w/REP	34.8 (+2.6)	34.9
w/SWL	35.0 (+2.8)	35.2
w/REP+SWL	36.3 (+4.1)	35.8
w/REP+SWL+FD	37.6 (+5.4)	36.2

(a) **DITO framework**.

pretraining method	AP_r
RO-ViT *our repro.*	32.2
w/FPN	33.2
w/FPN+Head	34.2
w/FPN+Head+RPN	34.8

(b) **Detector components**.

RoI embedding	AP_r
global avg/lvl	34.0
global max/lvl	33.7
multi RoIs, avg/lvl	33.8
multi RoIs, max/lvl	34.8
multi RoIs, max all	34.1

(c) **RoI embedding**.

Table 4: **Ablation on overall DITO framework and Region-centric Pretraining (Sec. 3.2).** REP: Region-centric Pretraining, SWL: Shifted-Window Learning, FD: Frozen backbone Distillation. Best setting is in gray.

backbone	AP_r	AP
fully global attn.	33.4	33.8
baseline window attn. [17]	32.2	33.0
Swin [23] style	31.3	33.1
shifted window	35.0	35.2

(a) **Shifted-window learning**.

# global attn. layer	base	w/ SWL
0	30.7	34.6 (+3.9)
4	32.2	35.0 (+2.8)
12	32.4	34.2 (+1.8)
24 (all layers)	33.4	33.4 (+0.0)

(b) **SWL w.r.t. # global attention layers**.

Table 5: **Ablation on Shifted-Window Learning (SWL - Sec. 3.3).**

COCO Benchmark. We present the comparison on COCO benchmark in Table 2. The main metric is AP50 of novel categories (novel AP). Without using pseudo labeling [6, 13, 43, 44], weak supervision [46], or externally trained detector modules [29, 35], our model demonstrates competitive results of 39.6 novel AP with LAION-2B and 40.2 with DataComp-1B. Among the ViT-based methods, DITO outperforms recent works RO-ViT [17] and CFM-ViT [16] by a clear margin of +6.3 and +5.3 points, respectively.

Transfer detection. We further evaluate DITO in the transfer detection setting, where the open-vocabulary detector trained on one dataset (LVIS$_{base}$) is tested on another dataset (Objects365) without any finetuning. By simply replacing the text embeddings, Table 3 shows that DITO achieves 20.0 AP, outperforming previous methods using ConvNet or ViT backbones of similar size.

4.2 Ablation Studies

For ablation study, we use the ViT-L/16 model pretrained with LAION-2B, and evaluate on the LVIS benchmark and report mask AP_r.

DITO overall framework. Table 4a summarizes the benefits of each DITO components. The region-centric pretraining (REP) improves the contrastive model baseline by +2.6 AP_r and shifted window learning (SWL) improves the baseline by +2.8 points. Combining both strategies brings a significant gain of +4.1 AP_r. Lastly, we observe that incorporating the frozen backbone distillation leads to an additional boost of +1.2 points. In the following, we provide ablations for each components.

Region-centric Pretraining. In Table 4b, we ablate our region-centric image-text pretraining by progressively adding the FPN, Faster R-CNN head, and RPN into the pretraining. Our 'baseline' is the contrastive image-text pretraining with cropped positional embedding [17]. On top of this, 'w/ FPN' introduces the FPN into the pretraining, where each pyramid level (whole image) map is mean-pooled

backbone	AP	AP_r	AP_c	AP_f
GT boxes given (region classification):				
before det. training	48.7	56.8	51.5	42.1
after det. training	53.7	54.2	53.7	53.5
after det. training w/FD	54.8	57.7	54.3	53.2

(a) GT region recognition.

method	AP_r	AP
RO-ViT *our repro.*	32.2	33.0
RO-ViT *our repro.*+FD	34.0 (+1.8)	33.6
w/ REP+SWL	36.3	35.8
w/ REP+SWL+FD	37.6 (+1.3)	36.2

(b) Effect of FD on OVD.

Table 6: Ablation on frozen backbone distillation (FD - Sec. 3.4).

into an image embedding, followed by the image-text contrastive loss per level. It improves the baseline by $+1.0$ AP_r. Adding the Faster R-CNN head further improves the alignment between the pretraining and detection finetuning, showing a gain of $+2.0$ AP_r. Incorporating all components *i.e.* the FPN, Faster R-CNN and RPN heads achieves the best 34.8 AP_r, a significant gain of $+2.6$ points over the baseline.

Table 4c shows that both global avg- and max-pooling are sub-optimal due to the lack of saliency map (avg-pool), and the limited capacity of a single pixel to represent semantic concepts for contrastive learning (max-pool), respectively. Our approach combines the best of both worlds, by first avg-pooling within each RoI, and then max-pooling over these RoI embeddings. Each embedding represents a proper RoI and the global representation captures the saliency map through max-pooling. Pooling per pyramid level (*i.e.* multi-level image-text supervision) outperforms pooling over all levels.

"a man riding a cart pulled by a donkey"   "a crane operates amidst piles of rubble"   "a person wearing blue rubber gloves is loading a syringe with vegetables"

"a mother and her two children sit on a green bench"   "cows on the street next to people that are on the sidewalk"   "small white bird sits on the back of a horse in an open field"

Fig. 3: Visual-text similarity map. For each example, we show the paired image (left) and text (bottom) input, and the visual-text similarity map using the contrastive model baseline [17] backbone features (middle) or our region-centric pretraining features (right). We use Flickr30K (top row) and COCO Captions (bottom row) datasets.

Shifted-Window Learning for detection. The CLIP ViT backbones are initially pretrained on lower resolutions and then adapted to higher resolution detection images. In Table 5a, we assess the efficacy of the shifted-window backbone in open-vocabulary detection training. Although the fully global attention model improves the detection task compared to the baseline windowed attention model [17], directly applying the pretrained ViT on high-resolution inputs may not be optimal, potentially compromising the locality structure of the pretrained lower-resolution features. Our shifted-window learning approach achieves 35.0 AP_r by preserving the locality structure from windowed attention as well as integrating information across fixed windows, outperforming the fully global

attention model (33.4 AP_r) which is computationally intensive. Notably, OWL-ViT [25,26] adopts the fully global attention model. In addition, naively applying the layer-alternating shifted window as in Swin [23] leads to a performance drop (see Sec. 3.3). Table 5b delves deeper into the behavior of SWL. The advantage of SWL diminishes steadily with increasing number of global attention layers in the windowed attention backbone, validating its better information mixing enabled by the SWL.

Frozen backbone distillation.   Table 6a studies the region classification capability of the CLIP ViT backbone before and after the detection finetuning. We use the ground truth boxes and measures AP scores to evaluate the zero-shot region classification of the base ('frequent' + 'common') and novel ('rare') categories. We observe that the finetuned backbone indeed shows a notable drop of -2.6 AP_r on novel classes, while overfitting to the base classes. The frozen backbone distillation ('w/ FD') leads to a significant improvement of +3.5 AP_r even surpassing the frozen pretrained backbone (before detection finetuning), while maintaining performance on the base classes (AP_c and AP_f). Table 6b presents the open-vocabulary detection results where the frozen backbone distillation brings a clear gain of +1.3~1.8 AP_r. These results highlight the efficacy of our frozen backbone distillation in detection finetuning (Sec. 3.4), as it effectively preserves the pretrained open-vocabulary knowledge while acquiring explicit region-text alignment through detection supervision.

4.3   Visualization

In Fig. 3, we visualize the similarity maps between image features and query text embeddings. We compare the baseline [17] contrastive model's backbone features with our region-centric pretrained features. We select pyramid level 4 which has the same resolution as the backbone features and apply the Faster R-CNN head in a sliding window manner to obtain the dense feature map. We observe that region-centric pretraining captures more localized semantic

Fig. 1 visualizes DITO outputs on LVIS novel categories and real-world out-of-distribution data from Ego4D [9]. Using the same DITO detector trained on $LVIS_{base}$, we find that DITO effectively identifies many novel and unseen objects, even under significant domain shifts.

5   Conclusion

We introduce DITO, a region-centric approach for open-vocabulary detection using large-scale image-text pairs. By integrating detection architecture onto the image backbone in CLIP pretraining, it learns locality-sensitive information without requiring pseudo labeling or box annotations. Furthermore, we propose a shifted-window learning method to mitigate the bias of the window attention pattern in CLIP ViT detectors. Experiments show that DITO outperforms the state-of-the-art by large margins on the LVIS benchmark, and is very competitive on the COCO benchmark and transfer detection. We hope this work would inspire the community to explore region-centric image-language pretraining for open-vocabulary localization tasks.

References

1. Bai, Y., Chen, X., Kirillov, A., Yuille, A., Berg, A.C.: Point-level region contrast for object detection pre-training. In: CVPR. pp. 16061–16070 (June 2022) 4
2. Chen, J., Zhu, D., Qian, G., Ghanem, B., Yan, Z., Zhu, C., Xiao, F., Elhoseiny, M., Culatana, S.C.: Exploring open-vocabulary semantic segmentation without human labels. arXiv preprint arXiv:2306.00450 (2023) 9
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020) 4
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 8
5. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: CVPR (2022) 1, 3, 10, 11
6. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: European Conference on Computer Vision. pp. 701–717. Springer (2022) 3, 10, 11, 12
7. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023) 11
8. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: ECCV (2022) 3
9. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022) 14
10. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022) 1, 3, 4, 6, 9, 10, 11
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) 4
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 6, 7
13. Huynh, D., Kuen, J., Lin, Z., Gu, J., Elhamifar, E.: Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7031 (2022) 1, 11, 12
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 9
15. Kaul, P., Xie, W., Zisserman, A.: Multi-modal classifiers for open-vocabulary object detection. arXiv preprint arXiv:2306.05493 (2023) 10
16. Kim, D., Angelova, A., Kuo, W.: Contrastive feature masking open-vocabulary vision transformer. ICCV (2023) 2, 6, 7, 10, 11, 12
17. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14
18. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. ICLR (2023) 1, 3, 4, 6, 10, 11

19. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022) 2, 3, 4
20. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: ECCV (2022) 3, 4, 5, 6, 7, 8, 10
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 5, 6
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 2, 4
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CVPR (2021) 3, 8, 12, 14
24. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024) 2, 4, 10, 11
25. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2023) 2, 3, 4, 10, 14
26. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. In: ECCV (2022) 3, 4, 10, 14
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 5
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 3, 5, 8, 9
29. Rasheed, H., Maaz, M., Khattak, M.U., Khan, S., Khan, F.S.: Bridging the gap between object and image-level representations for open-vocabulary detection. arXiv preprint arXiv:2207.03482 (2022) 1, 10, 12
30. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 5, 6, 7, 10
31. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) 3, 5, 9, 10, 11
32. Shi, C., Yang, S.: Edadet: Open-vocabulary object detection using early dense alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15724–15734 (2023) 10, 11
33. Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., Liu, S.: Object-aware distillation pyramid for open-vocabulary object detection. In: CVPR (2023) 10, 11
34. Wei, F., Gao, Y., Wu, Z., Hu, H., Lin, S.: Aligning pretraining for detection via object-level contrastive learning. In: NeurIPS (2021) 4
35. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15254–15264 (2023) 2, 3, 10, 11, 12

36. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7031–7040 (2023) 3, 10, 11

37. Xiao, T., Reed, C.J., Wang, X., Keutzer, K., Darrell, T.: Region similarity representation learning. In: ICCV. pp. 10539–10548 (October 2021) 4

38. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23497–23506 (2023) 2, 4

39. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., , Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In: arXiv:2209.09407 (2022) 2, 4

40. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary DETR with conditional matching. In: ECCV. pp. 106–122. Springer (2022) 10, 11

41. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR (2021) 1, 6, 11

42. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. Advances in Neural Information Processing Systems **35**, 36067–36080 (2022) 2, 4

43. Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Stathopoulos, A., Chandraker, M., Metaxas, D., et al.: Exploiting unlabeled data with vision and language models for object detection. In: ECCV (2022) 1, 10, 11, 12

44. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining. In: CVPR (2022) 1, 2, 3, 4, 10, 11, 12

45. Zhong, Y., Wang, J., Wang, L., Peng, J., Wang, Y.X., Zhang, L.: Dap: Detection-aware pre-training with weak supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4537–4546 (2021) 2, 4

46. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022) 1, 2, 10, 11, 12