RegionCLIP: Region-based Language-Image Pretraining

Yiwu Zhong[1*], Jianwei Yang[2], Pengchuan Zhang[2], Chunyuan Li[2], Noel Codella[3],
Liunian Harold Li[4], Luowei Zhou[3], Xiyang Dai[3], Lu Yuan[3], Yin Li[1], Jianfeng Gao[2]
[1]University of Wisconsin-Madison, [2]Microsoft Research, [3]Microsoft Cloud + AI, [4]UCLA
{yzhong52, yin.li}@wisc.edu,{jianwei.yang, penzhan, chunyl, ncodella,
luozhou, xidai, luyuan, jfgao}@microsoft.com, {liunian.harold.li}@cs.ucla.edu

Abstract

Contrastive language-image pretraining (CLIP) using image-text pairs has achieved impressive results on image classification in both zero-shot and transfer learning settings. However, we show that directly applying such models to recognize image regions for object detection leads to unsatisfactory performance due to a major domain shift: CLIP was trained to match an image as a whole to a text description, without capturing the fine-grained alignment between image regions and text spans. To mitigate this issue, we propose a new method called RegionCLIP that significantly extends CLIP to learn region-level visual representations, thus enabling fine-grained alignment between image regions and textual concepts. Our method leverages a CLIP model to match image regions with template captions, and then pretrains our model to align these region-text pairs in the feature space. When transferring our pretrained model to the open-vocabulary object detection task, our method outperforms the state of the art by ***3.8 AP50*** *and* ***2.2 AP*** *for novel categories on COCO and LVIS datasets, respectively. Further, the learned region representations support zero-shot inference for object detection, showing promising results on both COCO and LVIS datasets. Our code is available at https://github.com/microsoft/RegionCLIP.*

1. Introduction

The recent advances in vision-language representation learning has created remarkable models like CLIP [37], ALIGN [26] and Florence [59]. Such models are trained using hundreds of millions of image-text pairs by matching images to their captions, achieving impressive results of recognizing a large set of concepts without manual labels, and capable of transferring to many visual recognition tasks. Following their success on image classification, a natural question is whether these models can be used to reason about image regions, *e.g.*, for tasks like object detection.

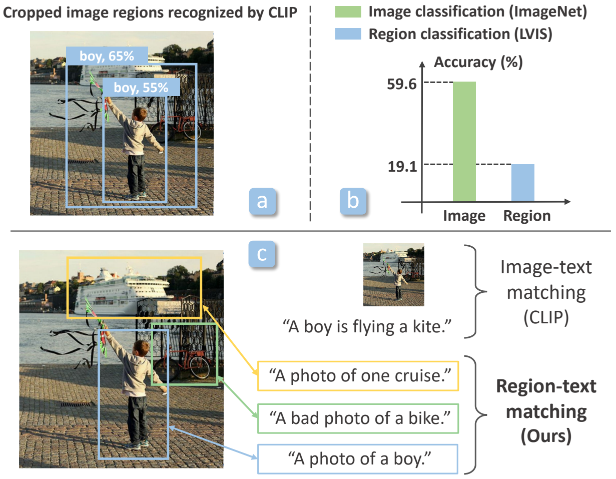
Cropped image regions recognized by CLIP
boy, 65%
boy, 55%
a
🟩 Image classification (ImageNet)
🟦 Region classification (LVIS)
Accuracy (%)
59.6
19.1
Image
Region
b
c
"A boy is flying a kite."
Image-text matching (CLIP)
"A photo of one cruise."
"A bad photo of a bike."
"A photo of a boy."
Region-text matching (Ours)

Figure 1. **(a)**. A pretrained CLIP model [37] failed to capture localization quality. **(b)**. A major drop on accuracy when using the same pretrained CLIP to classify image regions. **(c)**. Our key idea is learning to match *image regions* and their text descriptions.

To answer this question, we construct a simple R-CNN style [16] object detector using a pretrained CLIP model, similar to adapting a convolutional network pretrained on ImageNet. This detector crops candidate object regions from an input image, and applies the CLIP model for detection by matching visual features of cropped regions to text embeddings of object categories. Fig. 1(a-b) shows the results on LVIS dataset [19]. When using object proposals [42] as the input regions, scores from CLIP often fail to capture the localization quality (Fig. 1a). Even with ground-truth object boxes, classification accuracy using CLIP drops significantly from 60% on ImageNet to 19% on LVIS, with a similar number of classes (Fig. 1b). There is thus a major performance degradation when applying a pretrained CLIP model for object detection. *How can we empower a vision-language pretrained model to reason about image regions?*

We believe the main gap lies in the training of these vision-language models. Many existing vision-language

[*] Work done as an intern at Microsoft Research.

models, including CLIP, are trained to match an image with its image-level text description. The training is unaware of the alignment between local image regions and text tokens. Thus, the models are unable to precisely ground a textual concept to an image region. Further, cropping local image regions and matching them to text tokens largely ignore the surrounding visual context that is critical for object recognition, not to mention the high computational cost, *e.g*. a few seconds per image on a modern GPU.

In this paper, we explore learning *region representations* for object detection via vision-language pretraining. Our key idea is to explicitly align image regions and text tokens during pretraining. However, two key challenges arise. First, the fine-grained alignment between image regions and text tokens is not available in image-text pairs and expensive to annotate. Second, the text description of an image is often incomplete, *i.e*. many image regions are not described by the text. To address these challenges, we propose to bootstrap from a pretrained vision-language model to align image regions and text tokens, and to fill in the missing region descriptions, as illustrated in Fig. 1c.

Specifically, our method starts with a pool of object concepts parsed from text corpus, and synthesizes region descriptions by filling these concepts into pre-defined templates. Given an input image and its candidate regions from either object proposals or dense sliding windows, a pretrained CLIP model is used to align the region descriptions and the image regions, creating "pseudo" labels for region-text alignment. Further, we combine "pseudo" region-text pairs and ground-truth image-text pairs to pretrain our vision-language model via contrastive learning and knowledge distillation. Although the "pseudo" region-text pairs are noisy, they still provide useful information for learning region representations, and thus help to bridge the gap in object detection, as validated by our experiments.

We pretrain our RegionCLIP model on image captioning datasets (*e.g*., Conceptual Caption [45]) and mainly evaluate our method on the benchmarks of open-vocabulary object detection (COCO [32] and LVIS [19] datasets). When transferred to open-vocabulary object detection, our pretrained model establishes new state of the art (SoTA) on COCO and LVIS. For instance, our method outperforms previous methods [18, 60] by at least **3.8 AP50** and **2.2 AP** for novel categories on COCO and LVIS. Moreover, our model supports zero-shot inference and outperforms a set of strong baselines by a clear margin.

Our contributions are summarized as follows: (1) We propose a novel method that aligns image regions and their text descriptions without manual annotation, thereby enabling vision-language pretraining for learning visual region representations. (2) A key technical innovation that facilitates our pretraining is a scalable approach using text prompts to align the object descriptions with image regions, without relying on human annotations nor limited to the text paired with an image. (3) Our pretrained model presents strong results when transferred to open-vocabulary object detection, and demonstrates promising capability on zero-shot inference for object detection.

2. Related Work

Representation learning for images. Early works on visual representation learning focused on training image classification models using labor-intensive human annotations [13, 22, 30, 46, 50]. The learned features can be transferred to recognition tasks [16], and the classifier can be used to label images for semi-supervised learning [36, 55, 57]. To reduce the annotation burden, self-supervised learning [5,6,17,20] has received considerable attention recently.

The most relevant work is learning visual representations from natural language, such as image tags [3, 8, 12, 25, 28] and text descriptions [11,23,43,53,62]. Leveraging millions of image-text pairs collected from the Internet, recent methods in vision-language pretraining [26, 37] learned to match images with text descriptions and demonstrated impressive performance on zero-shot inference and transfer learning for image classification. However, these works focus on global representation tailored for image classification. In this paper, we propose to learn visual representation for local image regions to enable zero-shot inference and transfer learning for region based reasoning (*e.g*., object detection).

Representation learning for image regions. Many region based reasoning tasks, such as object detection [4, 41, 42, 52], rely on dense human annotations [14, 19, 29, 32]. Recently, semi-supervised learning was explored [48, 56, 66], where pretrained detectors are used to create pseudo labels of image regions. Beyond object labels, region representation learning benefits from from additional labels of object attributes [1, 29, 61], showing noticeable improvement on vision-language tasks [9, 31, 33, 51, 58, 63]. However, these works heavily rely on manual annotations and are limited to predefined categories. As a partial remedy, self-supervised learning was extended to region representations [24,40]. Inspired by CLIP [37] yet distinct from prior works, we propose to learn region representation via vision-language pretraining. Our learned representation enables the recognition of many visual concepts within image regions.

Zero-shot and open-vocabulary object detection. Zero-shot object detection aims at detecting novel object classes that are not seen during detector training [2, 18, 38, 39, 60, 65]. Bansal *et al*. [2] learned to match the visual features of cropped image regions to word embeddings [35] using max-margin loss. Rahman *et al*. [38] proposed polarity loss to model background category and to cluster categories with similar semantics. Zhu *et al*. [65] explored improving localization performance for novel categories by synthesizing

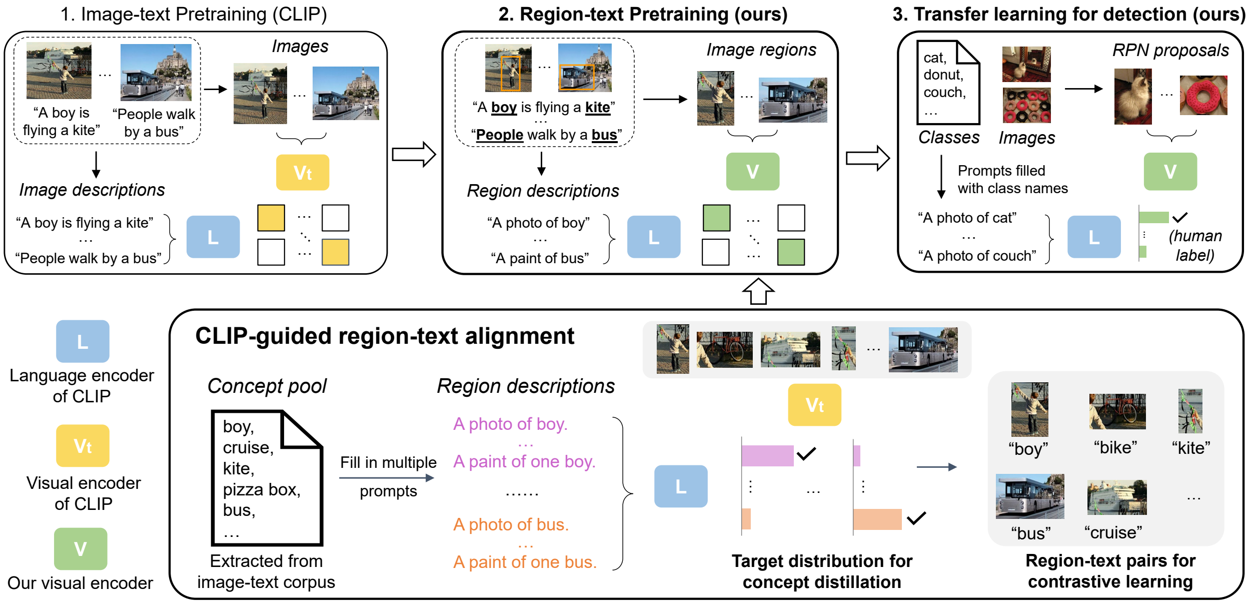

Figure 2. Method overview. We propose to learn visual representation for image regions via vision-language pretraining. Panel 1: With contrastive learning, CLIP is able to match images and their descriptions. Panel 2: Initialized by pretrained CLIP, our visual encoder learns visual region representation from the created region-text pairs. Specifically, as shown in the bottom row, we first create texts by filling the prompts with object concepts which are parsed from image descriptions, then use pretrained CLIP to align these texts and image regions proposed by RPN. Panel 3: When human annotation for image regions is available, we transfer our visual encoder for object detection.

visual features with a generative model.

Recently, Zareian *et al.* [60] proposed OVR for open-vocabulary object detection, where a visual encoder is first pretrained on image-text pairs to learn object concepts and then transferred to zero-shot object detection setting. Another close work is ViLD [18] that focuses on learning object detectors by distilling visual features from a pretrained CLIP model [37], yet still requires object labels and boxes for training. Similar to OVR and ViLD, our detector also leverages the visual-semantic space learned from vision-language pretraining. Different from OVR, we propose to learn region region representations from our "pseudo" region-text pairs given by a pretrained CLIP model. Our method is thus not restricted to existing text descriptions of an image. Unlike ViLD, our work addresses the problem of region representation learning, and focuses on *pretraining* from region-text pairs. As a result, our learned representations support zero-shot inference, while ViLD can not.

3. Region-based Language-Image Pretraining

Our goal is to learn a regional visual-semantic space that covers rich object concepts so that it can be used for open-vocabulary object detection. Consider a text description t that describes the content of region r in an image I. In the visual-semantic space, the visual region representation $\mathcal{V}(I, r)$ extracted from r should be matched to text representation $\mathcal{L}(t)$. $\mathcal{V}$ is a visual encoder that takes image I and a region location r, and outputs a visual representation for this region. $\mathcal{L}$ is a language encoder that converts a text description in natural language to a semantic representation.

Disentanglement of recognition and localization. There are two key components for region based reasoning: localization and recognition. Inspired by [47], we disentangle these two components, use existing region localizers, and consider a recognition problem. Our focus is thus learning visual-semantic space to recognize image regions without human annotations.

Method overview. As shown in Fig. 2, we denote $\mathcal{V}_t$ and $\mathcal{L}$ as visual and language encoders pretrained to match images to their descriptions, such as CLIP. Our goal is to train a visual encoder $\mathcal{V}$ so that it can encode image regions and match them to region descriptions encoded by language encoder $\mathcal{L}$. To address the challenge of missing region descriptions, as shown at the bottom of Fig. 2, we construct a pool of object concepts, create the region descriptions by filling concepts into prompts, and leverage a teacher encoder $\mathcal{V}_t$ to align these text descriptions with the image regions proposed by an image region localizer. Given the created region-text pairs, our visual encoder $\mathcal{V}$ learns to match these pairs via contrastive learning and concept distillation. Once pretrained, our model supports zero-shot inference for region recognition, and can be transferred to train object detector when the human annotation is available. We now describe region-level visual and semantic representations, and the alignment between image regions and text descriptions.

3.1. Visual and Semantic Region Representation

Visual region representation. Image regions can be proposed by either off-the-shelf object localizers (*e.g*., RPN [42]) or dense sliding windows . By default, we use RPN pretrained on human-annotated object bounding boxes *without* object labels. We use RPN to propose image regions and obtain N image regions, denoted as $\{r_i\}_{i=1,...,N}$.

Given the proposed regions, the visual representation v_i of region r_i is extracted from our visual encoder $\mathcal{V}$ with a feature pooling method, such as RoIAlign [21]. RoIAlign pools regional visual features from the feature map of a full image by using interpolation. We note that our visual encoder $\mathcal{V}$ is initialized by the teacher $\mathcal{V}_t$ so that it can have a good starting point in visual-semantic space.

Semantic region representation. A single image usually contains rich semantics, covering one or more objects from thousands of categories. It is costly to annotate all these categories in the large-scale image-text datasets. To this end, we first build a large pool of concepts to exhaustively cover regional concepts. As shown at the bottom of Fig. 2, we create a pool of object concepts which are parsed from text corpus (*e.g*., the image descriptions collected from the Internet), by using off-the-shelf language parsers [27, 44].

Given the concept pool, the semantic representations for regions are created by two steps: (1) a short sentence for each concept is created by filling it to prompt templates (*e.g*., prompts of CLIP [37]), *e.g*., the "kite" concept is converted to "A photo of a *kite*"; (2) the resulting text descriptions are further encoded into semantic representations by using the pretrained language encoder $\mathcal{L}$. Finally, all regional concepts are represented by their semantic embeddings $\{l_j\}_{j=1,...,C}$ and C denotes the size of concept pool.

While our region descriptions are built on existing image descriptions, our method is not constrained by the particular text descriptions that pair with images. Importantly, using a powerful language encoder $\mathcal{L}$ trained with hundreds of millions of text descriptions containing tens of thousands of words allows us to easily customize and scale up our concept pool. Such a capacity is deemed difficult to achieve using human annotations. In addition, the disentanglement of visual recognition and localization makes our method flexible to adopt different ways of extracting candidate regions.

3.2. Visual-Semantic Alignment for Regions

Alignment of region-text pairs. We leverage a teacher visual encoder $\mathcal{V}_t$ to connect image regions and our created texts (represented as semantic embeddings). Again, visual representation v_i^t of region r_i is extracted from teacher encoder $\mathcal{V}_t$ by pooling features from a local image region with RoIAlign. A matching score $S(v, l)$ between v_i^t and each concept embedding l_j is then computed by

$$S(v, l) = \frac{v^T \cdot l}{||v|| \cdot ||l||}. \tag{1}$$

The object concept with highest matching score, denoted as l_m, is selected and linked to region r_i. Finally, we obtain a pseudo label for each region, forming the pairs of $\{v_i, l_m\}$.

Our pretraining scheme. Our pretraining leverages both created region-text pairs and the existing image-text pairs. Given the aligned region-text pairs ($\{v_i, l_m\}$), we design a contrastive and a distillation loss based on the regions across different images to pretrain our visual encoder. Inspired by [34], the contrastive loss is computed as

$$L_{cntrst} = \frac{1}{N} \sum_i - \log(p(v_i, l_m)), \tag{2}$$

where $p(v_i, l_m)$ is given by

$$p(v_i, l_m) = \frac{\exp(S(v_i, l_m)/\tau)}{\exp(S(v_i, l_m)/\tau) + \sum_{k \in \mathcal{N}_{r_i}} \exp(S(v_i, l_k)/\tau)}. \tag{3}$$

Here τ is a predefined temperature, and $\mathcal{N}_{r_i}$ represents a set of negative textual samples for region r_i, *i.e*., the object concepts that are not matched to region r_i but matched to other regions in the batch.

Since positive pairs in the contrastive loss are inevitably "noisy", we also consider knowledge distillation for image regions. Knowledge distillation learns from a soft target and helps to handle the noise in those pseudo region-text pairs. This distillation loss is defined as

$$L_{dist} = \frac{1}{N} \sum_i L_{KL}(q_i^t, q_i), \tag{4}$$

where L_{KL} is the KL divergence loss; both q_i^t and q_i are probabilities over all object concepts. q_i^t is a soft target from teacher model computed as $softmax(S(v_i^t, l_1)/\tau, ..., S(v_i^t, l_C)/\tau)$. q_i is similarly computed from our student model.

Given image-text pairs collected from the Internet, our region-level contrastive loss L_{cntrst} can naturally extend to image-level contrastive loss $L_{cntrst-img}$. It can be considered as a special case where (1) the visual representation is extracted for a single global box that covers the whole image, (2) the corresponding text from the Internet describes the full image, and (3) negative samples are the text descriptions associated with other images. Finally, our overall loss function is given by

$$L = L_{cntrst} + L_{dist} + L_{cntrst-img}. \tag{5}$$

Zero-shot inference. Once pretrained, our visual encoder can be directly applied to region reasoning tasks. For example, given region proposals from RPN, region representations extracted from our visual encoder can be used to match the embeddings of target object concepts, and thus recognize the concepts within local image regions, thereby enabling zero-shot inference for object detection.

3.3. Transfer Learning for Object Detection

Our pretraining leverages region-text alignment created by the teacher model. Such alignment does not require human efforts, yet is not very accurate. When strong supervision for image regions is available (*e.g.*, the human-annotated detection labels), our visual encoder can be further fine-tuned by replacing the region descriptions with human annotations, as shown in Panel 3 of Fig. 2.

Specifically, we transfer our pretrained visual encoder to object detectors by initializing their visual backbones. To detect image objects, same as our pretraining, we use off-the-shelf RPN to localize object regions and recognize these regions by matching their visual region representation with the semantic embeddings of target object classes (*e.g.*, the object classes in detection dataset).

4. Experiments

Our main results are reported on transfer learning of our model for open-vocabulary object detection. Further, we evaluate our model on fully supervised object detection, as well as the zero-shot inference for object detection. Finally, we conduct ablations to study our model components.

Datasets. For pretraining, we consider Conceptual Caption dataset (CC3M) [45] with 3 millions of image-text pairs from the web. We also use a smaller dataset COCO Caption (COCO Cap) [7] when conducting ablation studies. COCO Cap contains 118k images, each associated with 5 human annotated captions. The parser from [27] is adopted to extract triplets (*e.g.*, man-play-ball) from captions in COCO Cap/CC3M dataset. Object concepts whose frequency are lower than 100 are discarded, leading to 4764/6790 concepts on COCO Cap/CC3M.

For transfer learning of open-vocabulary object detection, we train detectors with base categories of COCO detection dataset [32] and LVIS dataset (v1) [19], respectively. On COCO, We follow the data split of [2] with 48 base categories and 17 novel categories which are subsets of COCO object classes. We use the processed data from [60] with 107,761 training images and 4,836 test images. On LVIS, following [18], we use the training/validation images for training/evaluation and adopt the category split with 866 base categories (common and frequent objects) and 337 novel categories (rare objects).

Evaluation protocol and metrics. We evaluate object detection performance on COCO and LVIS for both transfer learning and zero-shot inference. The standard object detection metrics are used, including Average Precision (AP) and AP50 (AP at an intersection over union of 0.5).

Implementation details. *During pretraining*, the default student model and teacher model were ResNet50 [22] from pretrained CLIP. RPN used in pretraining was trained with the base categories of LVIS dataset. Our default model was pretrained on CC3M dataset with the concepts parsed from COCO Cap. SGD was used with the batch size 96, initial learning rate 0.002, maximum iteration of 600k, and 100 regions per image. The temperature τ was 0.01.

For transfer learning of object detection, our detectors were developed on Detectron2 [54] using Faster RCNN [42] (ResNet50-C4). RPN used in transfer learning was trained by the base categories of target dataset (*e.g.*, the transfer learning on COCO used the RPN trained on COCO). SGD was used with batch size 16, initial learning rate 0.002, and 1x schedule. Moreover, we applied class-wise weighted cross-entropy loss. (1) For base categories, we used focal scaling with the weight for a base category as $(1 - p^b)^\gamma$, where p^b is probability after softmax for this base category and $\gamma = 0.5/0.0$ on COCO/LVIS. Empirically, focal scaling helps to alleviate the forgetting of previously learned object concepts in pretraining, and thus is beneficial for novel categories. (2) For background category, we used a fixed all-zero embedding and a predefined weight (0.2/0.8 on COCO/LVIS) to background regions following [60].

For zero-shot inference of object detection, RPN was the same as pretraining stage and NMS threshold was set to 0.9. Inspired by [47, 64], we fused RPN objectness scores and category confidence scores by geometry mean. Empirically, fusing RPN scores significantly improves zero-shot results.

4.1. Transfer to Open-Vocabulary Object Detection

Setup. We evaluate our models on two benchmarks for open-vocabulary object detection, including COCO and LVIS. On COCO, we report AP50 and follow the evaluation settings in [60]: (1) only predicting and evaluating novel categories (Novel), (2) only predicting and evaluating base categories (Base), (3) a generalized setting that predicts and evaluates all categories (Generalized). On LVIS, we follow the benchmark of [18] where the rare objects are defined as novel categories. We report AP for novel categories (APr), base categories (APc, APf) and all categories (mAP), respectively. The detectors are trained by base categories and evaluated on base and novel categories (*e.g.*, 48/866 base categories and 17/337 novel categories on COCO/LVIS). To compare with ViLD [18], all experiments on LVIS additionally consider mask annotation.

Baselines. We consider several strong baselines:

- **Zero-shot object detectors** (SB [2], DELO [65], PL [38]): Zero-shot object detection is the closest area to open-vocabulary object detection. These detectors usually rely on the pretrained word embeddings of object classes for generalization to novel categories.
- **Open-vocabulary object detectors** (OVR [60], ViLD [18]): These detectors leverage pretrained vision-language models that have learned a large vocabulary from image-text pairs. OVR is our close competitor in the sense that we both pretrain visual encoders and

Visual Encoder Pretraining			Detector Training		COCO				
					Novel	Base	Generalized (17+48)		
Method	Dataset	Backbone	Method	Backbone	(17)	(48)	Novel	Base	All
Cls-ResNet [22]	ImageNet	RN50	FR-CNN [42]	RN50-C4	-	54.5	-	-	-
Cls-IncRN [49]	ImageNet	IncRNv2	SB [2]	IncRNv2	0.70	29.7	0.31	29.2	24.9
Cls-DarkNet [41]	ImageNet	DarkNet19	DELO [65]	DarkNet19	7.60	14.0	3.41	13.8	13.0
Cls-ResNet [22]	ImageNet	RN50	PL [38]	RN50-FPN	10.0	36.8	4.12	35.9	27.9
OVR [60]	COCO Cap	RN50	OVR [60]	RN50-C4	27.5	46.8	22.8	46.0	39.9
OVR [60]	CC3M	RN50	OVR [60]	RN50-C4	16.7	43.0	-	-	34.3
CLIP [37]	CLIP400M	ViT-B/32	ViLD* [18]	RN50-FPN	-	-	27.6	**59.5**	**51.3**
CLIP [37]	CLIP400M	RN50	Ours	RN50-C4	22.5	53.1	14.2	52.8	42.7
Ours	COCO Cap	RN50	Ours	RN50-C4	30.8	55.2	26.8	54.8	47.5
Ours	CC3M	RN50	Ours	RN50-C4	**35.2**	**57.6**	**31.4**	57.1	50.4
Ours	CC3M	RN50x4	Ours	RN50x4-C4	**43.3**	**61.9**	**39.3**	**61.6**	**55.7**

Table 1. Open-vocabulary object detection results on COCO dataset. Initialized by our pretrained visual encoder, our detector outperforms previous works on all metrics by a remarkable margin, and outperforms the recent work ViLD* on novel categories. ViLD* trains the detector with data augmentation of large-scale jittering (LSJ) [15] and a much longer training schedule (16x). Notations: Cls denotes the image classification pretraining on ImageNet [10], RN50 means ResNet50, IncRNv2 is Inception-ResNet-V2.

Visual Encoder Pretraining			Detector Training				LVIS			
Method	Dataset	Backbone	Method	Backbone	Training Strategy	Supervision	APr	APc	APf	mAP
-	-	-	Mask RCNN [21]	RN50-FPN	16x+LSJ [15]	Base+Novel	13.0	26.7	**37.4**	28.5
Cls-ResNet [22]	ImageNet	RN50	Mask RCNN [21]	RN50-C4	1x+Standard	Base+Novel	11.9	22.0	29.7	23.3
CLIP [37]	CLIP400M	ViT-B/32	ViLD* [18]	RN50-FPN	16x+LSJ [15]	Base	16.7	26.5	34.2	27.8
Ours	CC3M	RN50	Ours	RN50-C4	1x+Standard	Base	17.1	27.4	34.0	28.2
CLIP [37]	CLIP400M	ViT-B/32	ViLD* [18]	RN152-FPN	16x+LSJ [15]	Base	19.8	27.1	34.5	28.7
Ours	CC3M	RN50x4	Ours	RN50x4-C4	1x+Standard	Base	**22.0**	**32.1**	36.9	**32.3**

Table 2. Open-vocabulary object detection results on LVIS dataset. Without sophisticated training strategy, our detector still outperforms ViLD* on most metrics. Using same training strategy, our open-vocabulary detector beats the fully-supervised Mask RCNN for all metrics.

use them as the detector initialization. ViLD is a recent work that focuses on detector training by distilling visual features of a pretrained model from CLIP. ViLD specially uses the data augmentation of large-scale jittering (LSJ) [15] with 16x training time.

- **Fully supervised detectors**: On COCO, we include the supervised baseline from OVR which is a Faster RCNN [42] trained by the base categories with 1x schedule. On LVIS, we include the supervised baseline from ViLD which is a Mask RCNN [21] trained by base and novel categories with special data augmentation as ViLD. We additionally report a Mask RCNN trained in standard 1x schedule from Detectron2 [54].
- **Our detector variants**: We consider initializing our detector with different pretrained visual encoders, including CLIP and our model pretrained on COCO Cap.

Results. Table 1 and Table 2 show the results on COCO and LVIS datasets, respectively.

On COCO dataset, initialized by our pretrained backbone, our detector significantly outperforms previous method OVR [60] on all metrics (*e.g.*, 31.4 vs. 22.8 on novel categories). Compared with the CLIP backbone from which we start our region-based pretraining, our model brings a remarkable gain across all metrics, particularly +17.2 AP50 on novel categories. When compared with ViLD, a recent SoTA method with sophisticated training strategy, our model is still comparable on Base and All, while substantially better on Novel (*e.g.*, 31.4 vs. 27.6) which is the main focus in open-vocabulary detection. On LVIS dataset, with comparable backbone size (RN50x4-C4 of ours: 83.4M, RN152-FPN of ViLD: 84.1M), our detector outperforms ViLD by a large margin (*e.g.*, +2.2 APr and +3.6 mAP). Note that these superior detection results on COCO and LVIS are achieved by using a single pretrained backbone, with standard data augmentation and 1x training schedule. These results suggest that our region-based vision-language pretraining has learned better alignment between image regions and object concepts, and thus facilitates open-vocabulary object detection.

4.2. Transfer to Fully Supervised Object Detection

We further report results of fine-tuning our model with full supervision, following standard detection benchmark.

Setup. Detection annotation of all object categories are used during training and evaluation. Again, all experiments on LVIS additionally use mask annotation to train detector.

Baselines. We consider the following baselines: (1) Faster RCNN [42] initialized by ImageNet pretrained backbone: This is a common object detector in the community [54]. (2) Our detector initialized by pretrained CLIP. This baseline is

Visual Encoder Pretraining			Detector Training		COCO Train: 80, Test: 80		LVIS Train: 1203, Test: 1203			
Method	Dataset	Backbone	Method	Backbone	AP50	mAP	APr	APc	APf	mAP
Cls-ResNet [22]	ImageNet	RN50	FR-CNN [42]	RN50-C4	55.9	35.7	11.9	22.0	29.7	23.3
CLIP [37]	CLIP400M	RN50	Ours	RN50-C4	56.3	36.4	16.0	25.0	32.0	26.2
Ours	CC3M	RN50	Ours	RN50-C4	59.8	38.8	18.6	27.8	34.8	29.0
Ours	CC3M	RN50x4	Ours	RN50x4-C4	**64.4**	**42.7**	**24.5**	**32.0**	**36.5**	**32.5**

Table 3. Fully supervised object detection results on COCO and LVIS datasets. Our detector initialized by our pretrained visual encoder converges faster and significantly outperforms the petrained backbones of ImageNet and CLIP on all metrics at 1x schedule.

Visual Encoder Pretraining			Region	COCO	LVIS
Method	Dataset	Backbone	Proposals	All	mAP
OVR [60]	COCO Cap	RN50	GT	44.5	-
CLIP [37]	CLIP400M	RN50	GT	58.3	42.2
Ours	CC3M	RN50	GT	61.4	44.4
Ours	CC3M	RN50x4	GT	**65.5**	**50.7**
OVR [60]	COCO Cap	RN50	RPN	19.6	-
CLIP [37]	CLIP400M	RN50	RPN	25.5	9.2
Ours	CC3M	RN50	RPN	26.8	9.6
Ours	CC3M	RN50x4	RPN	**29.6**	**11.3**

Table 4. Zero-shot inference with ground-truth (GT) boxes or RPN boxes on COCO and LVIS datasets. All models use RoIAlign to extract visual representation of proposed image regions. Our pretrained models beat baselines by a clear margin across datasets.

Region-text Pairs	Image-text Pairs	COCO Zero-shot Inference		COCO Generalized (17+48)		
		All (RPN)	All (GT)	Novel	Base	All
✓		26.7	60.4	21.4	55.5	46.6
✓	✓	28.0	62.8	26.8	54.8	47.5

Table 5. Ablation study on pretraining strategies. All models are pretrained on COCO Cap.

Region Type		COCO Zero-shot Inference		COCO Generalized (17+48)		
Random	RPN	All (RPN)	All (GT)	Novel	Base	All
✓		27.1	60.8	25.2	54.5	46.9
	✓	28.0	62.8	26.8	54.8	47.5

Table 6. Ablation study on the type of regions used during pretraining. All models are pretrained on COCO Cap.

to validate our proposed pretraining method.

Results. In Table 3, the detector initialized by our pretrained visual backbone largely outperforms the baselines initialized by ImageNet and CLIP backbones (*e.g.*, +2.4 mAP on COCO and +2.8 mAP on LVIS). Our pretraining results in faster convergence and better accuracy at 1x schedule in this fully supervised setting. Again, when using RN50x4 as the backbone for both teacher model and student model, the performance is significantly improved (eg, +3.9 mAP on COCO, +3.5 mAP on LVIS).

4.3. Zero-shot Inference for Object Detection

Moving forward, we explore directly using RegionCLIP for zero-shot detection without any object annotations.

Setup. The pretrained vision-language models are directly used to recognize image regions. We use the same evaluation datasets and metrics as the experiments in transfer learning (All AP50 for COCO, mAP for LVIS)[1]. We consider two settings: (1) Ground-truth (GT) bounding boxes are used as region proposals. This oracle setting aims at evaluating the recognition performance by eliminating the localization error; (2) The region proposals come from RPN used in pretraining. The performance is thus impacted by both the quality of localization and accuracy of recognition.

Baselines. We consider two baselines: (1) OVR [60] pretrains a visual backbone on image-text pairs of COCO Cap which has close object concepts as COCO detection dataset. We evaluate the pretrained model provided in their code base. (2) CLIP [37] is pretrained on 400M image-text pairs. Both OVR and CLIP consider image-text pairs for pretraining, same as our RegionCLIP.

Results. Table 4 summarizes the results. With GT boxes, our pretrained model outperforms CLIP baseline by a clear margin across datasets (*e.g.*, 61.4 vs. 58.3 All AP50 on COCO, 44.4 vs. 42.2 mAP on LVIS). When compared with OVR, our model demonstrates a much larger margin (*e.g.*, 61.4 vs. 44.5 All AP50 on COCO), not to mention that OVR is pretrained on the same dataset as evaluation. When using RPN proposals, our model still clearly outperforms CLIP and OVR (*e.g.*, 26.8 vs. 19.6 & 25.5 on COCO, 9.6 vs. 9.2 on LVIS). Note that using GT boxes better characterizes the recognition performance of a pretrained model than using RPN, since RPN injects additional localization errors. These results suggest that our pretraining with region-text alignment improves the recognition of image regions. With RN50x4 architecture as the backbones of teacher and student models, the zero-shot inference performance is further improved across datasets and settings (*e.g.*, +6.3 mAP on LVIS with GT, +2.8 All AP50 on COCO with RPN).

4.4. Ablation Study

Finally, we conduct ablation studies using COCO Cap on zero-shot inference and transfer learning.

[1] The breakdown metrics (*e.g.*, Novel and Base) are omitted in zero-shot inference since no detection annotations are used.

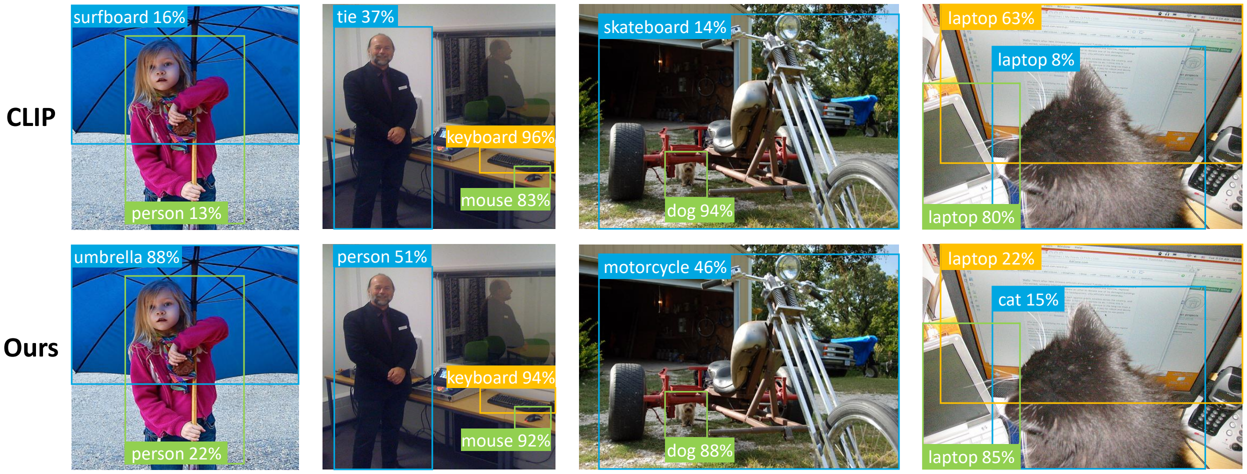

Figure 3. Visualization of zero-shot inference on COCO dataset with *ground-truth boxes*. Without finetuning, the pretrained models (top: CLIP, bottom: Ours) are directly used to recognize image regions into the categories in COCO. (Image IDs: 9448, 9483, 7386, 4795)

Pretraining strategies. Table 5 studies the effects of different pretraining strategies. While using the region-text pairs already attains plausible performance, adding image-text pairs further improves the results (*e.g.*, +2.4 AP50 with GT boxes on zero-shot inference, +5.4 Novel AP50 on transfer learning). We conjecture that image-text pairs provide contextual information from a global image description that compensates our pseudo region descriptions.

Types of image regions. Table 6 studies the effects of region proposal quality during pretraining. We replace the RPN proposals by randomly sampling the same number of image regions. Using random boxes hurts zero-shot inference (-2.0 AP50 with GT boxes), yet achieves comparable performance in transfer learning (46.9 vs. 47.5 All AP50). These results indicate that our pretraining is robust to the quality of region proposals. Zero-shot inference benefits from higher quality of proposals, yet the gap diminishes when human supervision is available to finetune the model.

Pretraining losses. Table 7 studies the effects of different losses. Combining contrastive and distillation loss has similar results as only using distillation loss when evaluated on zero-shot inference (62.8 vs. 63.1 AP50 with GT boxes), yet achieves the better results on transfer learning (26.8 vs. 24.1 Novel AP50). We hypothesis that the two losses are complementary. Distillation loss helps to inherit knowledge from the teacher model, while contrastive loss enforces discriminative representations for transfer learning.

Visualization. Fig. 3 visualizes the results of zero-shot inference with GT boxes on COCO dataset. Our model predicts more reasonable object categories than CLIP (*e.g.*, the blue regions in 1st and 2nd columns are correctly predicted as “umbrella” and “person” by our model). More visualizations can be found in our supplement.

Pretraining Loss		COCO Zero-shot Inference		COCO Generalized (17+48)		
Contrastive	Distillation	All (RPN)	All (GT)	Novel	Base	All
✓		25.2	58.2	21.8	54.2	45.8
	✓	27.8	63.1	24.1	54.6	46.7
✓	✓	28.0	62.8	26.8	54.8	47.5

Table 7. Ablation study on losses during pretraining. All models use image-level contrastive loss pretrained on COCO Cap.

5. Conclusion

In this paper, we proposed RegionCLIP — a novel region-based vision-language pretraining method that learns to match image regions and their descriptions. Our key innovation is a scalable approach to associate region-text pairs without using human annotation. By learning from such region-level alignment, our pretrained model established new state of the art when transferred to open-vocabulary object detection on COCO and LVIS datasets. Moreover, our pretrained model demonstrated promising results on fully supervised and zero-shot inference for object detection. We believe our work provides a solid step towards region representation learning, and we hope that our work can shed light on vision-language pretraining.

Limitations and Societal Impacts. Our work has several limitations that can be further investigated. (1) Our model does not consider object attributes and relationships, which are beneficial to many vision tasks (*e.g.*, visual grounding) and thus can be an interesting future direction. (2) Our method relies on CLIP’s visual-semantic space and does not update the language encoder. Unfreezing the language encoder may bring additional gains in the pretraining. Moreover, our model is pretrained on image captioning datasets (*e.g.*, CC3M) using CLIP prompts, and thus might inherit undesired biases from the datasets and the prompts.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086. IEEE, 2018.

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, pages 384–400, 2018.

[3] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of machine learning research*, 3:1107–1135, mar 2003.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

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

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[8] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1431–1439, 2015.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009.

[11] Karan Desai and Justin Johnson. VirTex: Learning visual representations from textual annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3270–3277, 2014.

[13] 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.

[14] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[15] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928, 2021.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[19] 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.

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

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5994–6002, 2017.

[24] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, October 2021.

[25] Yasuhide Mori Hironobu, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *in Boltzmann machines", Neural Networks*, page 405409, 1999.

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representa-

tion learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[27] Mao Jiayuan and Kasai Seito. Scene graph parser. *https://github.com/vacancy/SceneGraphParser*, 2018.

[28] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, pages 67–84. Springer, 2016.

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[31] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer, 2020.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019.

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[35] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[36] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[38] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[39] Shafin Rahman, Salman H Khan, and Fatih Porikli. Zero-shot object detection: Joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128(12):2979–2999, 2020.

[40] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Predet: Large-scale weakly supervised pre-training for detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2865–2875, October 2021.

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

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015.

[43] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020.

[44] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics (ACL).

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[47] Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1081–1090, 2018.

[48] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

[49] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence (AAAI)*, 2017.

[50] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

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

[53] Josiah Wang, Katja Markert, and Mark Everingham. Learning models for object recognition from natural language descriptions. In *The British Machine Vision Conference (BMVC)*, volume 1, page 2, 2009.

[54] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[55] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[56] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3060–3069, October 2021.

[57] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[58] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[59] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[61] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, June 2021.

[62] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834, 2021.

[63] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[64] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021.

[65] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[66] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 2020.