CapDet: Unifying Dense Captioning and Open-World Detection Pretraining

Yanxin Long[1*] Youpeng Wen[1*] Jianhua Han[2] Hang Xu[2] Pengzhen Ren[1]
Wei Zhang[2] Shen Zhao[1†] Xiaodan Liang[1,3†]
[1]Shenzhen Campus of Sun Yat-sen University [2]Huawei Noah's Ark Lab [3]MBZUAI
longyx9@mail2.sysu.edu.cn, wenyoupeng0@outlook.com, hanjianhua4@huawei.com,
chromexbjxh@gmail.com, renpzh@mail.sysu.edu.cn, wz.zhang@huawei.com,
zs-06@163.com, xdliang328@gmail.com

Abstract

Benefiting from large-scale vision-language pre-training on image-text pairs, open-world detection methods have shown superior generalization ability under the zero-shot or few-shot detection settings. However, a pre-defined category space is still required during the inference stage of existing methods and only the objects belonging to that space will be predicted. To introduce a "real" open-world detector, in this paper, we propose a novel method named CapDet to either predict under a given category list or directly generate the category of predicted bounding boxes. Specifically, we unify the open-world detection and dense caption tasks into a single yet effective framework by introducing an additional dense captioning head to generate the region-grounded captions. Besides, adding the captioning task will in turn benefit the generalization of detection performance since the captioning dataset covers more concepts. Experiment results show that by unifying the dense caption task, our CapDet has obtained significant performance improvements (e.g., +2.1% mAP on LVIS rare classes) over the baseline method on LVIS (1203 classes). Besides, our CapDet also achieves state-of-the-art performance on dense captioning tasks, e.g., 15.44% mAP on VG V1.2 and 13.98% on the VG-COCO dataset.

1. Introduction

Most state-of-the-art object detection methods [37, 38, 55] benefit from a large number of densely annotated detection datasets (*e.g.*, COCO [30], Object365 [40], LVIS [14]). However, this closed-world setting results in the model only being able to predict categories that appear in the training set. Considering the ubiquity of new concepts in real-world scenes, it is very challenging to locate and identify these new visual concepts. This predictive ability of new concepts in open-world scenarios has very important research value in real-world applications such as object search [32, 34], instance registration [50], and human-object interaction modeling [12].

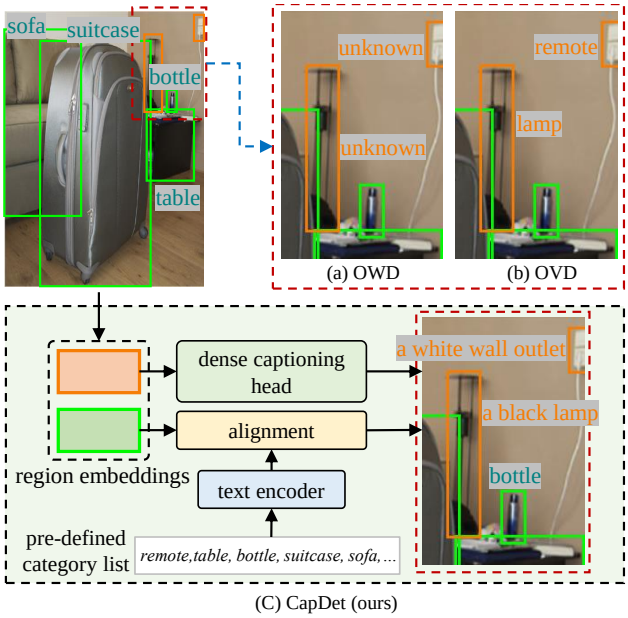


Figure 1. Comparison of the different model predictions under OWD, OVD, and our setting. (a) OWD methods [16, 20, 53] are not able to describe the detailed category of the detected unknown objects and (b) the performance of OVD methods [9, 14, 45] usually depends on the pre-defined category list during the inference. (c) With the unification of two pipelines of dense captioning and open-world detection pre-training, our CapDet can either predict under a given category list or directly generate the description of predicted bounding boxes.

Currently, the open world scenario mainly includes two tasks: *open world object detection* [20] (OWD) and *open-vocabulary object detection* [49] (OVD). Although the paradigms of OWD and OVD tasks are closer to the real world, the former cannot describe the specific concept of the detected unknown objects and requires a pre-defined category list during the inference. Specifically, as shown

*Equal contribution.
†Corresponding authors.

in Figure 1, previous OWD methods [16, 20, 53] would recognize new concepts not in the predefined category space as "unknown". Further, another line of task OVD requires the model to learn a limited base class and generalize to novel classes. Compared to the *zero-shot object detection* (ZSD) proposed by [36], OVD allows the model to use external knowledge, *e.g.*, knowledge distillation from a large-scale vision-language pre-trained model [9, 14], image-caption pairs [49], image classification data [54], grounding data [28, 45, 51]. With the external knowledge, OVD methods show a superior generalization capacity to detect the novel classes within a given category space. However, as shown in Figure 1, when given an incomplete category list, OVD can only predict the concepts that appear in the given category list, otherwise, there will be recognition errors, (*i.e.*, as illustrated in Figure 1 (b), the OVD methods prone to predict the "wall socket" as "remote", since the latter is in the category list but not the former).

Thus, under the OVD setting, we mainly face the following two challenges: (*i*) it is difficult to define a complete list of categories; (*ii*) low response values on rare categories often lead to recognition errors. This is mainly because we cannot exhaustively enumerate new objects in the real world, and secondly, it is difficult to collect enough samples for rare classes. However, the fact that rare objects in the real world, even some new objects that are unknown to humans, such as UFOs, do not prevent people from using natural language to describe it as "a flying vehicle that looks like a Frisbee".

Therefore, based on the above observations, in this paper, we consider a new setting that is closer to the open world and real scenes, *i.e.*, we expect the model to both detect and recognize concepts in a given category list, and to generate corresponding natural language descriptions for new concepts or rare categories of objects. Early dense captioning methods [11, 19] can locate salient regions in images and generate the region-grounded captions with natural language. Inspired by this, to address the challenges faced in the OVD setting, we propose to unify the two pipelines of dense captioning and open-world detection pre-training into one training framework, called **CapDet**. It empowers the model with the ability to both accurately detect and recognize common object categories and generate dense captions for unknown and rare categories by unifying the two training tasks.

Specifically, our CapDet constructs a unified data format for the dense captioning data and detection data. With the data unification, CapDet further adopts a unified pre-training paradigm including open-world object detection and dense captioning pre-training. For open-world detection pretraining, we treat the detection task as a semantic alignment task and adopt a dual encoder structure as [45] to locate and predict the given concepts list. The concepts list contains category names in detection data and region-grounded captions in dense captioning data. For dense captioning pretraining, CapDet proposes a dense captioning head to take the predicted proposals as input to generate the region-grounded captions. Due to the rich visual concepts in the dense captioning data , the integration of dense captioning tasks will in turn benefit the generalization of detection performance.

Our experiments show that the integration of few dense captioning data brings in large improvement in the object detection datasets LVIS, *e.g.*, +2.7% mAP on LVIS. The unification of dense captioning and detection pre-training gains an additional 2.3% increment on LVIS and 2.1% increment on LVIS rare classes. Besides, our model also achieves state-of-the-art performance on dense captioning tasks. Note that our method is the first to unify dense captioning and open-world detection pretraining.

To summarize, our contributions are three folds:

- We propose a novel open-vocabulary object detection framework CapDet, which cannot only detect and recognize concepts in a given category list but also generate corresponding natural language descriptions for new concept objects.
- We propose to unify the two pipelines of dense captioning and open-world detection pre-training into one training framework. Both two pre-training tasks are beneficial to each other.
- Experiments show that by unified dense captioning task and detection task, our CapDet gains significant performance improvements on the open-vocabulary object detection task (*e.g.*, +3.3% mAP on LVIS rare classes). Furthermore, our CapDet also achieves state-of-the-art performance on the dense captioning tasks, *e.g.*, 15.44% mAP on Visual Genome (VG) V1.2 and 13.98% mAP on VG-COCO.

2. Related Work

Vision-Language Pre-training. Vision-Language Pre-training [7, 18, 35] as a scheme in the domains of natural language processing [1, 6] and computer vision [8] obtains continual attention currently. And it exhibits strong performance and generalization ability on various downstream vision and cross-modal tasks. Among them, CLIP [35] and ALIGN [18] as dual-stream methods utilize large-scale image-text pairs on the Internet by cross-modal contrastive learning to get excellent zero-shot classification ability. Single-stream methods [22, 27] unify visual and textual embeddings in a single transformer-based model, which can perform text generation tasks such as image caption and VQA. Some mixed architectures [26, 43] combine single-stream and dual-stream to explore a unified way of vision-language understanding and generation. However, these methods take low-resolution images as input and serve the

task of classification and retrieval. Those vision-language pre-training approaches can not be applied to pure computer vision task directly, *i.e.*, object detection task.

Open World Object Detection / Open-Vocabulary Object Detection. Object detection is a core computer vision task, which aims at localizing objects using a bounding box and classifying them. The mature detection approaches which show great performance on supervised data include one-stage detectors (*i.e.*, YOLO [37], ATSS [52]) having a relatively high detection efficiency and two-stage detectors (*i.e.*, Faster R-CNN [38], Mask R-CNN [17]) having good detection accuracy. However, how to generalize these methods to rare classes and novel concepts in the real world is a challenge. Currently, several object detection approaches for such open-world scenes have attracted extensive attention from academia and industry. These methods are divided into two tasks which are called open-world object detection and open-vocabulary object detection respectively depending on whether to detect the class of unknown classes.

For the OWD task, Zhao et al. [53] proposed a proposal advisor to assist in identifying unknown proposals without supervision and a class-specific expelling classifier to filter out confusing predictions. For the OVD task, GLIP [28] converts the detection data into grounding format and proposes a fusion module to learn semantic vision information in grounding data. K-Lite [42] reconstructs the input format of the data in GLIP from sequential to parallel and uses nouns hierarchy and definition to format text sequence. DetCLIP [45] unifies detection, grounding, and image-text pair data in a paralleled formulation and constructs a concept dictionary to augment the text data, which strikes a balance between performance and efficiency. Differing from all these works, our CapDet can generate an open-set caption of each region proposal to cover situations where the semantics of new object instances are not in the given category list.

Dense Captioning. Dense captioning aims at generating detailed descriptions for local regions, which usually needs to locate visual regions with semantic information and generate captions for these regions. J. Johnson et al. [19] utilized a fully convolutional localization network to locate regions of interest (RoIs) and then describe them. Afterward, many methods [29, 47] based on Faster-RCNN [38] and LSTM [13] are proposed to do dense captioning. X. Li et al. [29] arrange RoI features as a sequence and put them into LSTM with the guidance of the region features to form the complementary object context features. This method also needs ground truth bounding boxes auxiliary tests to achieve good results. But limited by the forget gate mechanism of LSTM, the inputted sequence cannot be too long. Then, the transformer-based method TDC [41] is proposed to tackle the long sequence forgotten problem. Instead, our CapDet proposes a transformer-based caption head to generate a caption using a single-stage detector ATSS while simultaneously achieving open-world detection.

3. Method

The overview of our proposed CapDet is shown in Figure 2. To construct a detector to either predict under a given category list or directly generate the concepts of predicted bounding boxes, we incorporate detection data and dense caption data together. In this section, we will present a unified data format for the detection data and dense caption data in Section 3.1, the model architecture and pre-training objectives for open-world object detection pre-training in Section 3.2 and dense captioning in Section 3.3.

3.1. Unified Formulation

We defined a unified triplet-wise data format $(x, \{\mathbf{b}_i\}_{i=1}^{N}, y_{i=1}^{N})$ for each sample from different sources. Specifically, $x \in \mathbb{R}^{3\times h\times w}$ is the input image, $\{\mathbf{b}_i | \mathbf{b}_i \in \mathbb{R}^4\}_{i=1}^{N}$ denotes the bounding boxes coordinates for each region of the image, and the $y_{i=1}^{N}$ represents the concepts of the corresponding boxes. N denotes the number of regions. A concept y_i formatted as a sentence contains the category and textual description of the corresponding region. In detection data, a concept y consists of the category name and the corresponding definition from the concept dictionary [45], while y_i represents the region-grounded caption in dense caption data. For example, for an image x in detection data, y_i can be:

$$y_i = \text{“}person,\ a\ human\ being.\text{”}$$

For an image x from dense captioning data, y_i can be:

$$y_i = \text{“}an\ outlet\ on\ the\ wall.\text{”}$$

With the triplet, we can learn a unified image-text alignment objective on the detection data and the dense captioning data. The unified formulation also ensures the joint training of open-world object detection pre-training and dense captioning.

3.2. Open-World Object Detection Pre-training

Based on the unified formulation of detection data and dense captioning data, we regard the captions of regions in dense captioning data as a kind of category and utilize two different sources of data for the open-world object detection pre-training. Compared with the limited class list of detection data, dense caption data contains richer concepts and more semantic information than class names of individual regions. On the other hand, localization and recognition are two essential tasks of object detection. Traditional object

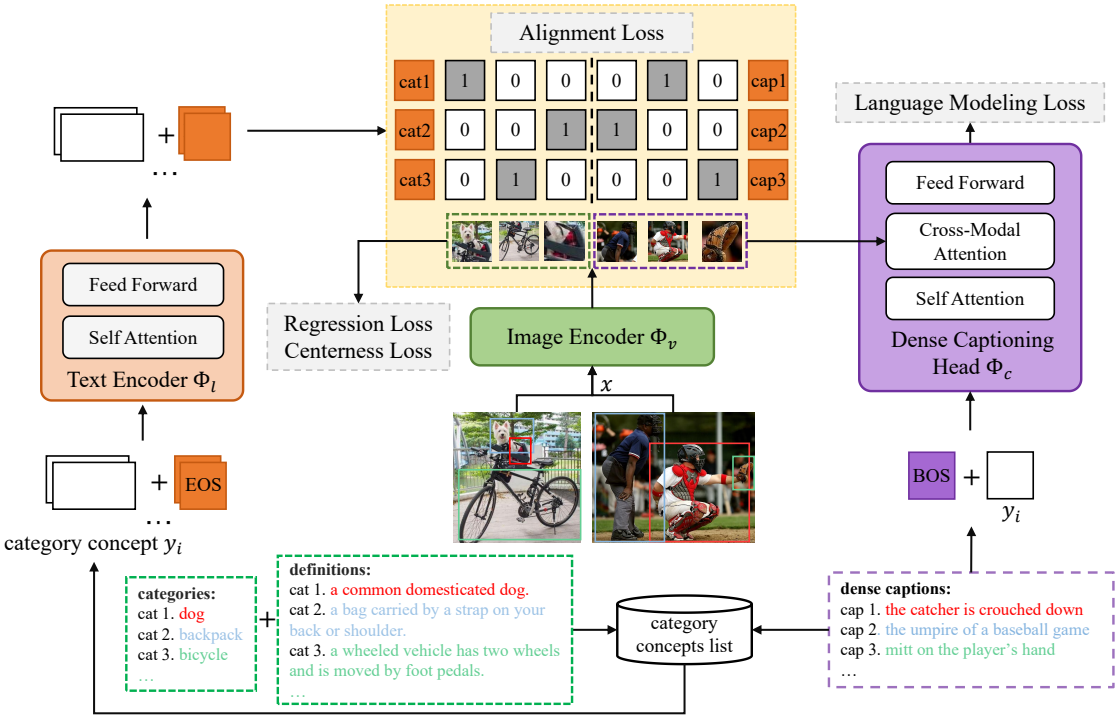


Figure 2. The overall architecture of CapDet. The training paradigm of CapDet contains open-world object detection pre-training and dense captioning. In detection, CapDet contains a dual vision-language encoder. The image encoder generates region embeddings from detection and dense captioning data. The regression loss and centerness loss are introduced to regress the locations. The text encoder takes the category concepts as input to generate the embeddings from the [EOS] token. Then we treat the detection task as a matching task and adopt an alignment loss for the category embeddings and region embeddings. In dense captioning, an additional dense captioning head is proposed to take the region embeddings as input and generate the textual captions for corresponding regions with natural language.

detection always focuses on the salient objects in the image. While the dense captioning data contains lots of annotations which are just parts of an object, *e.g.*, *an ear of an elephant*, it is not suitable to learn those annotations for the localization task. Therefore, we only calculate the localization loss on detection data.

As shown in Figure 2, CapDet predicts the regions and treats the recognition task as a region-category matching task. For efficient learning on the matching task, we adopt the negative sampling proposed by [45] to provide negative concepts to enlarge the concept space in a batch. Specifically, for each iteration, we randomly sample a negative concept set and add to the positive concept set (N samples) in a batch to obtain the final concept set $y_{i=1}^{M}$, where M represents the sum of the number of positive and negative samples. Finally, we format the triplet to $(x, \{\mathbf{b}_i\}_{i=1}^{N}, y_{i=1}^{M})$.

CapDet contains a dual vision-language encoder and takes the triplet $(x, \{\mathbf{b}_i\}_{i=1}^{N}, y_{i=1}^{M})$ as input. The image encoder Φ_v is an object detector that can predict the bounding boxes of regions from the input image x and output the region features $O \in \mathbb{R}^{K\times D}$. The text encoder Φ_l takes the concept set $y_{i=1}^{M}$ as input and obtains the text embeddings $W \in \mathbb{R}^{M\times D}$ from the special token [EOS] concatenated with the text input. K, D denotes the number of predicted regions and region feature dimensions. The alignment score matrix $S \in \mathbb{R}^{K\times M}$ of regions and texts is calculated by:

$$O = \Phi_v(x), W = \Phi_l(y_{i=1}^{M}), S = OW^T \tag{1}$$

where T denotes the transpose operation. A ground-truth alignment matrix $G \in \{0,1\}^{K\times M}$ is constructed to indicate the matching relation of regions and concepts. The alignment loss $\mathcal{L}_{align}$ is calculated by the predicted alignment scores of regions S and the ground-truth alignment matrix G. Following [28, 45], we adopt the ATSS [52] detector as an image encoder, and $\mathcal{L}_{align}$ is typically a sigmoid focal loss. As a one-stage detector, the localization loss contains centeredness loss $\mathcal{L}_{cen}$ and bounding box regression loss $\mathcal{L}_{reg}$. The training objective of detection pre-training can be written as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{align}} + \alpha\mathcal{L}_{\text{reg}} + \beta\mathcal{L}_{\text{center}}, & \text{for detection} \\ \mathcal{L}_{\text{align}}, & \text{for dense captioning} \end{cases} \tag{2}$$

where α and β denote the weights for the centerness loss $\mathcal{L}_{cen}$ and box regression loss $\mathcal{L}_{reg}$, respectively. The $\mathcal{L}_{cen}$ is the sigmoid loss and the $\mathcal{L}_{reg}$ is the GIOU loss [39].

Model	Backbone	Pre-Train Data	Images Number	LVIS AP	LVIS AP_r / AP_c / AP_f
Mask-RCNN [17]	Swin-T	LVIS	0.1M	34.1	19.1 / 34.0 / 37.0
ATSS [52]	Swin-T	LVIS	0.1M	33.6	19.7 / 32.4 / 37.2
ATSS [52]	Swin-L	LVIS	0.1M	43.9	30.6 / 43.7 / 46.3
MDETR [21]	RN101	GoldG+	0.77M	24.2	20.9 / 24.3 / 24.2
GLIP-T(A) [28]	Swin-T+DH+F	O365	0.66M	18.5	14.2 / 13.9 / 23.4
GLIP-T(C) [28]	Swin-T+DH+F	O365,GoldG	1.43M	24.9	17.7 / 19.5 / 31.0
GLIP-T [28]	Swin-T+DH+F	O365,GoldG,Cap4M	5.43M	26.0	20.8 / 21.4 / 31.0
K-Lite [42]	Swin-T	O365	0.66M	21.3	14.8 / 18.6 / 24.8
K-Lite [42]	Swin-T	O365,GoldG	1.43M	26.1	17.2 / 24.6 / 29.0
GLIPv2-T [51]	Swin-T+DH+F	O365,GoldG,Cap4M	5.43M	29.0	- / - / -
DetCLIP-T(A) [45]	Swin-T	O365	0.66M	28.8	26.0 / 28.0 / 30.0
DetCLIP-T(B) [45]	Swin-T	O365, GoldG	1.43M	34.4	26.9 / 33.9 / 36.3
DetCLIP-T(C)* [45]	Swin-T	O365, VG	0.73M	31.5	27.5 / 30.6 / 33.0
CapDet (Ours)	Swin-T	O365, VG	**0.73M**	**33.8**	**29.6 / 32.8 / 35.5**

Table 1. Zero-shot performance on LVIS [15] MiniVal5k datasets. AP_r / AP_c / AP_f indicate the AP values for rare, common, and frequent categories, respectively. "DH" and "F" in GLIP [28] baselines stand for the dynamic head [4] and cross-modal fusion, respectively. Baselines with * are implemented with our code base. GoldG+ denotes the GoldG plus the COCO [30] caption dataset.

3.3. Dense Captioning

The open-world object detection pre-training ensures CapDet gains the capacity to detect under given an arbitrary category list. However, when the given category list is not complete enough to cover the potential classes on a new domain data, the detector will perform worse on the categories which are not in the given list. Considering such limitation, we propose a dense captioning head Φ_C to generate semantically rich concepts with natural language for the predicted proposals. In the dense captioning task, the model receives an image and produces a set of regions and the corresponding captions. The dense captioning head is a cross-modal decoder that takes the c predicted regions features O generated by the image encoder as input. The captioning (*i.e.*, language modeling) loss is calculated by:

$$\mathcal{L}_{cap} = -\log p(y_{it}|\Phi_c(y_{i(\tau<t)}, O_i)), \tag{3}$$

where y_{it} means the t token in caption y_i corresponding to region feature O_i, and $y_{i(\tau<t)}$ means tokens before t in caption y_i. The overall pre-training loss can be written as:

$$\mathcal{L} = w_d\mathcal{L}_{det} + w_c\mathcal{L}_{cap}, \tag{4}$$

where w_d, w_c denote the weighting factor of $\mathcal{L}_{det}$ and $\mathcal{L}_{cap}$.

To minimize the gap in the type of bounding boxes between the detection data and dense captioning data, we propose a simple way to transform our detector as a class-agnostic detector and only select the top k regions based on the centeredness scores to adapt to the dense captioning task. We can fine-tune our CapDet on the dense captioning data to achieve better performance. Specifically, we propose "object" as the foreground concept and "background" as the background concept. The text encoder Φ_l outputs the concept embeddings $W' \in \mathbb{R}^{2\times D}$. Then the alignment scores $S' \in \mathbb{R}^{K\times 2}$ is calculated by Eqn. 1. The captioning head takes the top k most confident proposal embeddings based on centeredness scores as input to predict the region-grounded captions.

4. Experiment

Implementation Details. For the image encoder, we adopt the Swin-T backbone proposed in Swin-Transformer [31] which is pre-trained on ImageNet-1K [5]. We use 12 layers 8 heads transformer as our text encoder and load a base model checkpoint released by FILIP [46], in order to make a fair comparison with DetCLIP [45]. The structure of the dense captioning head is consistent with that in the text encoder but trained from scratch for a fair comparison. We employ AdamW [23] optimizer and set the batch size to 32. The learning rate is set to 1.4×10^{-4} for the parameters of the image encoder and detection head, and 1.4×10^{-5} for the text encoder and dense captioning head. When fine-tuning the VG dataset to do the dense captioning task, we set the learning rate to 1.4×10^{-4}. Without otherwise specified, all models are trained with 12 epochs and the learning rate is decayed with a factor of 0.1 at the 8-th and the 11-th epoch. The context token

Figure 3. Qualitative visualizations between GLIP-T, DetCLIP-T(C) and CapDet. From top to down, the three rows of images show the LVIS zero-shot detection results of GLIP-T, DetCLIP-T(C), and CapDet respectively. All models are pre-trained on O365 and VG.

length for input text is set to 20. We set the number of input captions to 150, and the number of the region features N is determined by the feature map. The loss weight factor w_c and w_d are both set to 1.0. We build our model on MMDetection [2] code base.

Dataset. Our CapDet is trained with two types of data, including detection data and caption data. Following DetCLIP [45], we use Object365 [40] (it will be abbreviated as O365 in the following paper) as detection data, and sample 0.66M data from O365 v2 for training. Following GLIP [28] and DetCLIP [45], LVIS [15] MiniVal5k (defined in [21]) which has 5000 images is used for detection evaluation. Moreover, we remove the training samples contained in the LVIS dataset for fair zero-transfer evaluation. For dense captioning data, we mainly conduct our experiments on VG [24] V1.2 and VG-COCO (defined in [41]). Following [41], we allocate 77398 images for training and 5000 images for validation and testing on VG. As demonstrated in [24], the ground-truth bounding boxes of VG are much denser than the other object detection datasets, *i.e.*, the average number of per sample in MS COCO [30] is only 7.1 *vs.* 35.4 in VG. Then an intersection of VG V1.2 and MS COCO is proposed by [41] and is denoted as VG-COCO, which has 38080 images for training, 2489 for validation, and 2476 for testing.

Benchmark Settings. We mainly evaluate our method on open-vocabulary object detection and dense captioning task. For open-vocabulary object detection, we evaluate the direct domain transfer on LVIS [15] which contains 1203 categories. Following [28, 45], we metric the zero-shot detection performance by the Average Precision (AP) on a 5k subset. The annotations of LVIS data are split into three folds, *i.e.*, rare, common, and frequency, based on the number of categories. Since there is almost no overlap between the rare classes and the classes of training dataset Objects365 [40], the AP of the rare classes shows a valuable zero-shot detection performance. For the dense captioning task, we follow the setting of [19] to evaluate the VG and VG-COCO. The evaluation metric we adopt is the mean Average Precision proposed by [19] which is calculated across a range of thresholds for both localization and language accuracy, *i.e.*, the intersection over union (IOU) thresholds .3, .4, .5, .6, .7 are used for localization and the METEOR score' thresholds 0, .05, .1, .15, .2, .25 is adopted for evaluating the language generation.

4.1. Open-world Detection Results

Table 1 shows our zero-shot object detection performance on LVIS. We mainly train our CapDet with the backbone Swin-T [31] on the detection data Objects365 [40] and dense captioning data (VG [24]). Since DetCLIP does not report the performance on O365 and VG, we train DetCLIP on the two datasets under the same settings and denote it as DetCLIP-T(C) for a fair comparison. Comparing the 11th row and 12th row, our CapDet outperforms DetCLIP-T(C) on the same data scale and backbone with an extra simple caption head. Moreover, our model's zero-shot performance even surpasses the fully-supervised model with the same backbone by a large margin on rare classes, *i.e.*, CapDet outperforms ATSS by 9.9%.

Qualitative Visualizations Figure. 3 illustrates the detection results on LVIS [24] dataset from GLIP-T, DetCLIP-T(C), and CapDet. All three models are trained on O365

Figure 4. Qualitative visualizations between JIVC and CapDet. "w/o ft" means do caption without finetune, while "w/ ft" means with finetune.

and VG, and details are given in Section 4.3. Given a category list, the rare classes are detected more precisely by our CapDet, *e.g.*, "kitchen table" in the first column, "horse buggy" in the third column, and "fishbowl" in the sixth column that our model CapDet detects correctly but the other two not.

4.2. Dense Captioning Results

Due to the target bounding boxes in dense captioning data containing lots of local structures of objects and being much denser than the bounding boxes in object detection data, we do not regress the bounding box in the pre-training stage. The previous works directly train on the dense captioning data and generate captions on the top k proposals ranking by a confidence score. When fine-tuning our model on the VG dataset for the dense captioning tasks, we transform our CapDet into a class-agnostic detector. Specifically, we propose "object" as the foreground concept and "background" as the background concept for computing alignment scores. The scores are used as proposal confidences to predict the region-grounded captions.

Table 2 and Table 3 show CapDet significantly outperforms the latest work TDC [41] by 2.5% on mAP on VG and TDC+ROCSU [41] by 2.08%, respectively. It is worth noticing that, even against given the ground-truth bounding boxes with the previous method COCG [29] denoted as COCOG&GT, our CapDet still gains a 43.80% mAP increase and achieves state-of-the-art. One important reason is that the excellent detection performance of our model assists the localization ability of dense captioning tasks.

Method	mAP(%)
FCLN [19]	5.16
JIVC [44]	9.96
ImgG [29]	9.68
COCD [29]	9.75
COCG [29]	10.39
CAG-Net [47]	10.51
TDC [41]	11.90
CapDet (Ours)	**15.44**

Table 2. Comparison of mAP (%) performance on dense captioning benchmark on the VG V1.2 dataset.

Qualitative Visualizations. Figure 4 shows a qualitative visualization comparison between JIVC [44] and our CapDet. The three image rows from top to bottom are the visualization of JIVC, CapDet without fine-tuning, and CapDet with finetuning. In the second row, CapDet can locate more objects than JIVC, owing to our model's superior localization performance. After finetuning, CapDet can further describe a region rather than a single object such as

Method	mAP(%)
FCLN [19]	4.23
JIVC [44]	7.85
Max Pooling [29]	7.86
ImgG [29]	7.81
COCD [29]	7.92
COCG [29]	8.90
COCG-LocSiz [29]	8.76
COCG&GT [29]	9.79
TDC+ROCSU [41]	11.9
CapDet (Ours)	**13.98**

Table 3. Comparison of mAP (%) performance on the dense captioning benchmark on the VG-COCO Dataset.

MODEL	DC HEAD	LVIS	
		AP	AP_r / AP_c / AP_f
GLIP-T	✗	30.4	22.5 / 29.0 / 33.0
GLIP-T	✓	33.1	27.0 / 32.1 / 35.0
DETCLIP-T	✗	31.5	27.5 / 30.6 / 33.0
DETCLIP-T	✓	33.8	29.6 / 32.8 / 35.5

Table 4. Ablations on integrating our dense captioning head into different baselines.

Pre-training Data	Fine-tune	DCap mAP(%)	Box mAP(%)
VG	✗	12.86	27.65
O365,VG	✗	4.72	9.65
VG	✓	13.83	28.58
O365,VG	✓	15.44	30.61

Table 5. Ablations on incorporating data from different sources. "DCap" stands for the dense caption mAP.

"two women in a kitchen" in the 5-th column.

4.3. Ablation Studies

4.3.1 Ablations for Unified Pre-training

Effect on different baselines. Table 4 investigates the advantages of dense captioning heads on different baselines. We integrate our dense captioning head with GLIP-T or DetCLIP-T. The GLIP-T is implemented with parallel text encoding without external knowledge following the setting as ablations in [42] on our code base. All the results are pre-trained on Objects365 and VG. The results show that our dense captioning head is able to boost the generalization and model-agnostic.

Effect of dense captioning data. Table 1 shows the efficiency of incorporating dense captioning data. Specifically, only 0.07M data added, the DetCLIP-T(C) gains +2.7% overall AP and +1.5% AP_r on LVIS compared to DetCLIP-T(A). The performance of DetCLP-T(A) on rare categories also outperforms DetCLIP-T(C) train on Objects365 and GOLDG, while the data size is 1.43M *vs.* 0.73M.

4.3.2 Ablations for dense captioning

We investigate the impact of training policy and data from different sources on the dense captioning task. As shown in row1 in Table 5, our CapDet still achieves a significant performance which is directly trained on VG outperforms the previous task (*i.e.*, TDC [41] in Table 2). Row2 is our CapDet and is pre-trained on Objects365 and VG, while only the bounding box in the Objects365 is regressed, and then transformed on a dense captioning task. Since the type of bounding boxes in dense captioning is different from the detection data, the result of the direct transforming to dense captioning is worse. However, we've proved that our model still keeps the dense captioning capacity on the salient objects in Figure 4. The results in row3 and row4 indicate that pre-training on the detection data Objects365 is also beneficial to the dense captioning task.

5. Limitations

These are a few issues that we need to improve in the future: (1) Although our unification training paradigm works well on open-vocabulary object detection and dense captioning task, the training of dense captioning generation costs lots of time. (2) In addition, existing dense captioning data is high-cost to collect. We will research how to collect large-scale dense captioning data by auto annotation and get better performance with the scaled-up data.

6. Conclusion

In this paper, we propose a novel open-world object detection method named CapDet. Our CapDet is more practical in the open world and real scenes. Specifically, CapDet introduces a unification training framework including open-world object detection pre-training and dense captioning. The unification enables our CapDet to localize and recognize concepts in an arbitrary given category list or directly generate textual captions for predicted new concept objects. Experiments show that the design of unification is both beneficial to open-world object detection tasks and dense captioning tasks. In the future, our CapDet can be easily injected into any open world and real scenes tasks. The unification framework can also be integrated into any other OWD/OVD methods to generate semantic-rich concepts for unknown/novel objects.

Acknowledgements We gratefully acknowledge the support of MindSpore[1], CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

[1]https://www.mindspore.cn/

Appendix for CapDet: Unifying Dense Captioning and Open-World Detection Pretraining

Appendix A. Detailed Experimental Settings

The detailed architecture parameters for different modules of CapDet are shown in Table 6. For the learning rate scheduler, we assign a base learning rate and then linearly warm it up to the peak learning rate according to the effective total batch size by a square root strategy, $lr_{peak} = lr_{base} \times \sqrt{\text{batchsize}/16}$, *e.g.*, we set image encoder base learning rate to 1×10^{-4} and it automatically scales to 1.4×10^{-4}. The training hyperparameters used for CapDet are shown in Table 7.

Image Encoder	Value
backbone	swin-t
neck	fpn
input resolution	1333×800
Text Encoder	**Value**
width	512
heads	8
layers	12
Cross-Modal Decoder	**Value**
width	512
heads	12
layers	12

Table 6. Detailed architecture parameters for different module.

Hyperparameter	Value(%)
Image encoder lr	1.4×10^{-4}
Text encoder lr	1.4×10^{-5}
Crossmodal decoder lr	1.4×10^{-5}
Learning policy	CosineAnnealing
warmup ratio	0.0001
warmup iters	1000
batchsize	32
weight decay	0.05
w_c	1
w_d	1

Table 7. The training hyperparameters used for CapDet.

Appendix B. Fine-tuning Results on LVIS

We provide the fine-tuning results on LVIS in Table 8 below. We observe that CapDet outperforms the baseline DetCLIP with 1.2% AP on average and 6.5% AP on rare classes. Besides, though pre-trained with fewer data and tasks, CapDet shows a competitive performance compared with the GLIPv2.

Model	Backbone	Pre-Train Data	Images Number	LVIS AP	LVIS AP_r / AP_c / AP_f
DetCLIP-T(C)* [45]	Swin-T	O365, VG	0.73M	45.6	33.6 / 45.8 / 47.5
GLIPv2-T [28]	Swin-T+DH+F	O365, GoldG, Cap4M	5.43M	50.6	- / - / -
CapDet (Ours)	Swin-T	**O365, VG**	**0.73M**	**47.2**	**40.1 / 46.9 / 48.7**

Table 8. Fine-tuning performance on LVIS [15] MiniVal5k datasets. AP_r/AP_c/AP_f indicate the AP values for rare, common, and frequent categories. 'DH' and 'F' in GLIP [28] baselines stand for the dynamic head [4] and cross-modal fusion.

Appendix C. Open-World Detection Results on LVIS Full Validation Set

Table 9 reports our zero-shot object detection performance on LVIS [15] full validation set. Following [28, 45], we take the class names with additional manually designed prompts as input of text encoder. Comparing the 5th row and 6th row, our CapDet still outperforms DetCLIP-T(C) on the same data scale and backbone with an extra simple caption head. The zero-shot performance surpasses the previous methods with the same backbone by a large margin on rare classes, *e.g.*, CapDet trained on fewer data outperforms GLIP-T [28] by 10.8% on AP_r.

Appendix D. Analysis of the Improvements on OVD

We attribute the improvements on OVD to the reason that the incorporation of captioning head brings more generalizability for the region features, which in turn helps the learning of OVD task. Specifically, the dense captioning task is essentially a sequential classification task with a large enough class space (*i.e.*, word tokens), while alignment task is a single-step

Model	Backbone	Pre-Train Data	Images Number	LVIS Val Full	
				AP	AP_r / AP_c / AP_f
GLIP-T(A) [28]	Swin-T+DH+F	O365	0.66M	12.3	6.00 / 8.00 / 19.4
GLIP-T [28]	Swin-T+DH+F	O365,GoldG,Cap4M	5.43M	17.2	10.1 / 12.5 / 25.2
DetCLIP-T(A) [45]	Swin-T	O365	0.66M	22.1	18.4 / 20.1 / 26.0
DetCLIP-T(C) [45]	Swin-T	O365, VG	0.73M	23.5	18.4 / 21.6 / 27.9
CapDet (Ours)	Swin-T	O365, VG	0.73M	**26.1**	**20.9 / 24.4 / 30.2**

Table 9. Zero-shot transfer performance on LVIS [15] full validation dataset. AP_r/AP_c/AP_f indicates the AP values for rare, common, and frequent categories. 'DH' and 'F' in GLIP [28] baselines stand for the dynamic head [4] and cross-modal fusion.

classification task with a limited class space. Therefore, training with dense captioning tasks will bring the region feature into a more proper location in feature space rather than simply pulling them together via only detection task. As shown in Table 10, we further conduct the experiments to demonstrate the effectiveness of pre-training under captioning. By comparing the row 2 and 5, we observe that even with only dense captioning data (VG data), pre-training with the dense captioning paradigm also brings a significant improvement.

Model	Pre-Train Data	LVIS	
		AP	AP_r / AP_c / AP_f
DetCLIP-T [45]	O365	28.8	26.0 / 28.0 / 30.0
	VG	10.3	8.6 / 10.1 / 10.8
	O365, VG	31.5	27.5 / 30.6 / 33.0
CapDet	O365	28.5	25.2 / 27.5 / 29.9
	VG	11.4	10.2 / 11.1 / 11.8
	O365, VG	**33.8**	**29.6 / 32.8 / 35.5**

Table 10. Zero-shot performance on LVIS [15] MiniVal5k datasets. AP_r / AP_c / AP_f indicate the AP values for rare, common, and frequent categories, respectively. "DH" and "F" in GLIP [28] baselines stand for the dynamic head [4] and cross-modal fusion, respectively.

Appendix E. 'Real' Open-world Object Detection Deployment Strategy

In this paper, the detection and dense captioning task are illustrated separately for better understanding and comparison with other methods, since no benchmark has considered combining these two tasks. For the practical deployment, we propose a simple two-stage ensemble way to stay true to the motivation. Specifically, in the first stage, we execute detection on images among the pre-defined categories list and treat the proposals with maximum alignment scores among all classes less than a threshold as 'unknown' objects. Then in the second stage, we generate the captions for the 'unknown' objects. To demonstrate the effectiveness of the proposed strategies, We conduct detection on the images with 80 categories of COCO and regenerate captions for the 'unknown' objects. As shown in the Figure 5 , our proposed strategy expands the semantic space of the limited categories list and shows reasonable results.

Appendix F. More Ablation Studies

Ablations on Pre-trained Language Model Table 11 reports the effect of different tokenizers and pre-trained language models loaded for text encoder. We ablate two kinds of pre-trained language models and corresponding tokenizers for our text encoder. For dense captioning head, we construct the same decoder as BLIP [26] decoder and keep the tokenizer the same as the text encoder. The results indicate the FILIP [46] encoder with byte pair encoding performs a better generalization, since it is pre-trained on a larger scale of data, *i.e.*, 300M in FILIP [46] *vs.* 128M in BLIP [26].

Ablations on the Weighting Factor of Dense Captioning Loss We study the effect of weights of detection loss and dense captioning loss during pre-training. We set the weighting factor of detection loss w_d to 1.0. Table 12 provides the ablations of the weighting factor of dense captioning loss w_c. We choose $w_c = 1$ for CapDet, since the result of overall AP is the best.

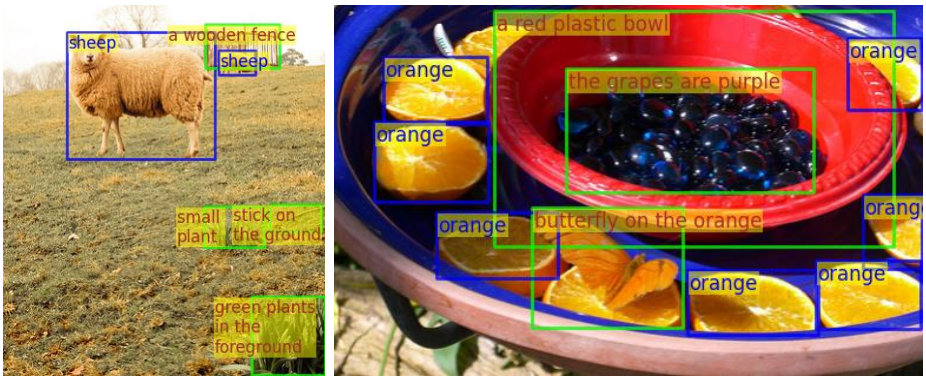


Figure 5. Deployment results.

Pre-trained Model	Tokenizer	Vocab Size	DC Head	LVIS AP	LVIS AP_r / AP_c / AP_f
BLIP [26]	WordPiece	30524	✗	30.4	26.7 / 29.4 / 32.0
			✓	32.4	27.4 / 31.8 / 33.9
FILIP [46]	BPE	49408	✗	31.5	27.5 / 30.6 / 33.0
			✓	33.8	29.6 / 32.8 / 35.5

Table 11. Effect of different tokenizers and language models. 'DC Head' and 'BPE' stand for the integration of Dense Captioning Head and Byte Pair Encoding.

w_c	LVIS AP	LVIS AP_r / AP_c / AP_f
0.5	33.6	31.0 / 32.8 / 34.9
1.0	**33.8**	29.6 / 32.8 / 35.5
1.5	33.5	32.0 / 32.1 / 35.0

Table 12. Effect of weighting factor of dense captioning loss.

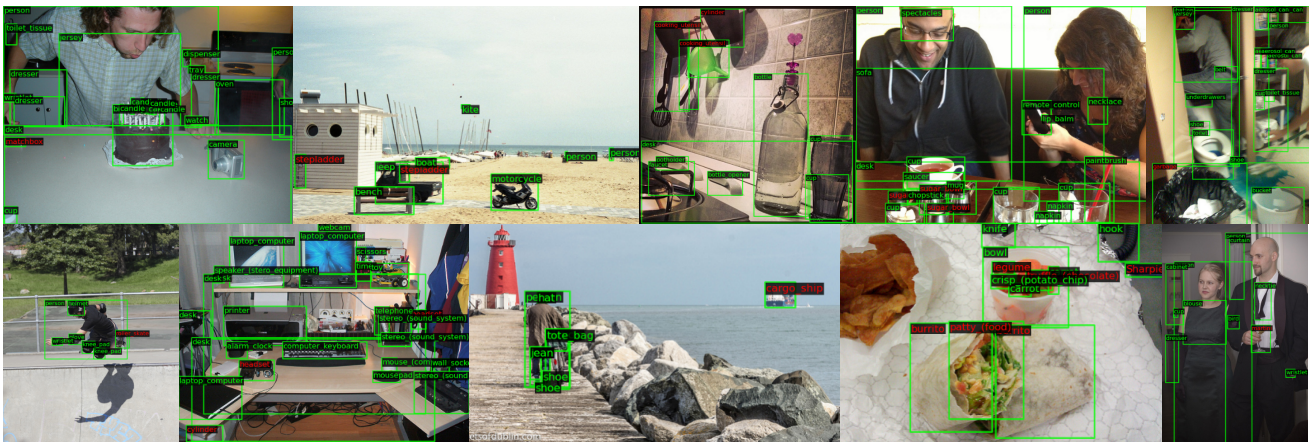


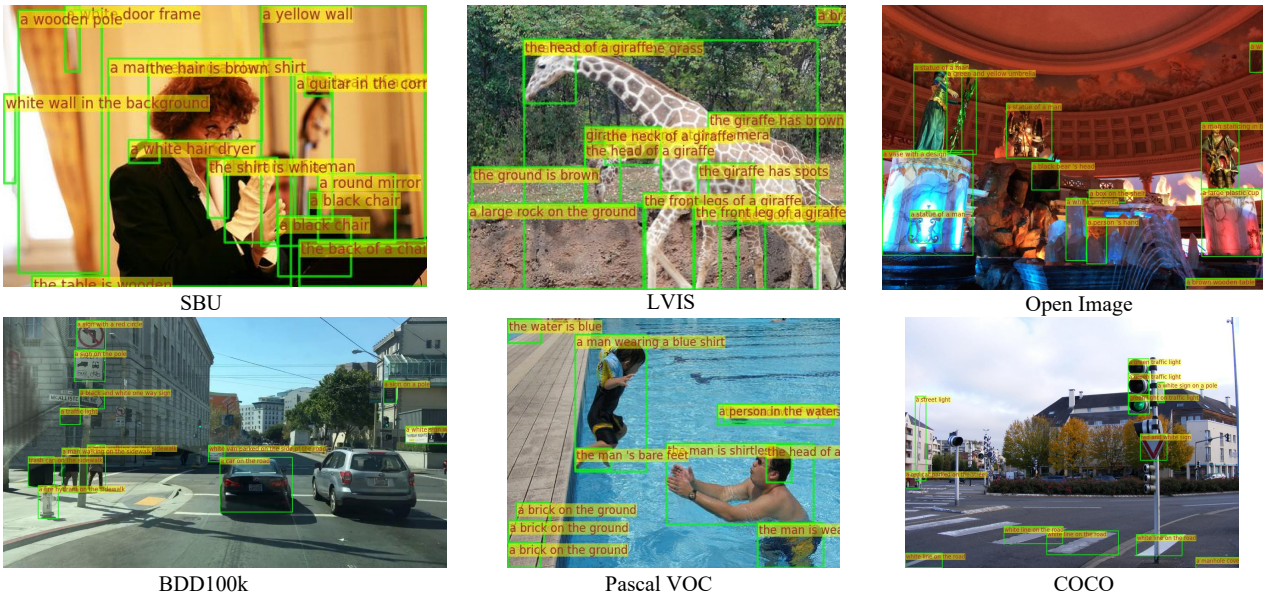
Figure 6. Qualitative visualizations on LVIS.

Appendix G. More Qualitative Results

Open-World Detection Results Figure 6 illustrates more detection results on LVIS [15] dataset from our CapDet. We highlight the detected rare classes's text in red.

Figure 7. Qualitative visualizations on VG.

Dense Captioning Results Figure 7 shows more captioning results on VisualGenome [24] dataset. Our model CapDet locates not only "object" such as "bicycle" but also "region" such as "a shadow on the ground". We also explored the zero-shot generalization ability of CapDet. We directly use our model to do the zero-shot dense captioning task without finetuning on serveral datasets, which include SBU [33], LVIS [15], Open Image [25], BDD100K [48], Pascal VOC [10] and COCO [3]. As shown in Figure 8, CapDet can accurately locate objects and generate corresponding region-grounded captions.



SBU LVIS Open Image

BDD100k Pascal VOC COCO

Figure 8. Qualitative visualizations on several datasets.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. 2

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *ArXiv*, abs/1906.07155, 2019. 6

[3] 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. 12

[4] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 5, 9, 10

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 2

[7] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[9] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 2

[10] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 12

[11] Yiqi Gao, Xinglin Hou, Yuanmeng Zhang, Tiezheng Ge, Yuning Jiang, and Peng Wang. Caponimage: Context-driven dense-captioning on image. *arXiv preprint arXiv:2204.12974*, 2022. 2

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 1

[13] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012. 3

[14] 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. 1, 2

[15] 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. 5, 6, 9, 10, 11, 12

[16] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9235–9244, 2022. 1, 2

[17] 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. 3, 5

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[19] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 2, 3, 6, 7, 8

[20] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1, 2

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 5, 6

[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6, 12

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 12

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2, 10, 11

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 3, 4, 5, 6, 9, 10

[29] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8650–8657, 2019. 3, 7, 8

[30] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5, 6

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 5, 6

[32] Jingjing Meng, Junsong Yuan, Jiong Yang, Gang Wang, and Yap-Peng Tan. Object instance search in videos via spatio-temporal trajectory discovery. *IEEE Transactions on Multimedia*, 18(1):116–127, 2015. 1

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 12

[34] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[36] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, pages 547–563. Springer, 2018. 2

[37] 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. 1, 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 3

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1, 6

[41] Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3, 6, 7, 8

[42] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022. 3, 5, 8

[43] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 2

[44] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017. 7, 8

[45] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 1, 2, 3, 4, 5, 6, 9, 10

[46] Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ArXiv*, abs/2111.07783, 2022. 5, 10, 11

[47] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6241–6250, 2019. 3, 7

[48] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 12

[49] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2

[50] Haoyang Zhang and Xuming He. Deep free-form deformation network for object-mask registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4251–4259, 2017. 1

[51] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 2, 5

[52] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 3, 4, 5

[53] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yuqing Ma, Yixuan Qiao, and Duorui Wang. Revisiting open world object detection. *arXiv preprint arXiv:2201.00471*, 2022. 1, 2, 3

[54] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1