CapDet: Unifying Dense Captioning and Open-World Detection Pretraining

Yanxin Long1* Youpeng Wen1* Jianhua Han2 Hang Xu2 Pengzhen Ren1 Wei Zhang2 Shen Zhao† Xiaodan Liang1,3†
1Shenzhen Campus of Sun Yat-sen University 2Huawei Noah’s Ark Lab 3MBZUAI
longyx9@mail2.sysu.edu.cn, wenyouspeng@outlook.com, hanjianhua4@huawei.com,
chromexbjxh@gmail.com, renpzh@mail.sysu.edu.cn, wz.zhang@huawei.com,
z-s-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 captioning 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 [33, 34, 50] benefit from a large number of densely annotated detection datasets (e.g., COCO [27], Object365 [36], LVIS [12]). 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 [29, 30], instance registration [45], and human-object interaction modeling [10].

Currently, the open world scenario mainly includes two tasks: open world object detection [18] (OWD) and open-vocabulary object detection [44] (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

![Figure 1. Comparison of the different model predictions under OWD, OVD, and our setting. (a) OWD methods [14, 18, 48] are not able to describe the detailed category of the detected unknown objects and (b) the performance of OVD methods [8, 12, 41] 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.](image-url)
in Figure 1, previous OWD methods [14, 18, 48] 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 [32], OVD allows the model to use external knowledge, e.g., knowledge distillation from a large-scale vision-language pre-trained model [8, 12], image-caption pairs [44], image classification data [49], grounding data [25, 41, 46]. 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 [9, 17] 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 pre-training, we treat the detection task as a semantic alignment task and adopt a dual encoder structure as [41] 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 [6, 16, 31] as a scheme in the domains of natural language processing [1, 5] and computer vision [7] obtains continual attention currently. And it exhibits strong performance and generalization ability on various downstream vision and cross-modal tasks. Among them, CLIP [31] and ALIGN [16] 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 [20, 24] 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 [23, 39] 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 [33], ATSS [47]) having a relatively high detection efficiency and two-stage detectors (i.e., Faster R-CNN [34], Mask R-CNN [15]) 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. [48] 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 [25] converts the detection data into grounding format and proposes a fusion module to learn semantic vision information in grounding data. K-Lite [38] reconstructs the input format of the data in GLIP from sequential to parallel and uses noun hierarchy and definition to format text sequence. DETCLIP [41] 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. [17] utilized a fully convolutional localization network to locate regions of interest (RoIs) and then describe them. Afterward, many methods [26, 43] based on Faster-RCNN [34] and LSTM [11] are proposed to do dense captioning. X. Li et al. [26] 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 [37] 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 \( (x, \{b_i\}_{i=1}^N, y_i) \) for each sample from different sources. Specifically, \( x \in \mathbb{R}^{3 \times h \times w} \) is the input image, \( \{b_i\}_{i=1}^N \in \mathbb{R}^{4 \times N} \) denotes the bounding boxes coordinates for each region of the image, and \( y_i \in \mathbb{R}^{N} \) 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 \( y_i \) consists of the category name and the corresponding definition from the concept dictionary [41], 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:

\[
y_i = \text{"person, a human being."}
\]

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

\[
y_i = \text{"an outlet on the wall."}
\]

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
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 [41] to provide negative concepts to enlarge the concept space in a batch. Specifically, for each iteration, we randomly sample a negative concept set (N samples) in a batch to obtain the final concept set \( y_{M}^{M} \), where M represents the sum of the number of positive and negative samples. Finally, we format the triplet to \((x, \{b_i\}_i, y_{M}^{M})\).

CapDet contains a dual vision-language encoder and takes the triplet \((x, \{b_i\}_i, y_{M}^{M})\) as input. The image encoder \( \Phi_v \) 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 \( \Phi_t \) takes the concept set \( y_{M}^{M} \) 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:

\[
O = \Phi_v(x), W = \Phi_t(y_{M}^{M}), S = OW^T \tag{1}
\]

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 \( L_{align} \) is calculated by the predicted alignment scores \( S \) and the ground-truth alignment matrix \( G \). Following [25, 41], we adopt the ATSS [47] detector as an image encoder, and \( L_{align} \) is typically a sigmoid focal loss. As a one-stage detector, the localization loss contains centeredness loss \( L_{cen} \) and bounding box regression loss \( L_{reg} \). The training objective of detection pre-training can be written as:

\[
L = \begin{cases} 
L_{align} + \alpha L_{reg} + \beta L_{center} , & \text{for detection} \\
L_{align} , & \text{for dense captioning} 
\end{cases} \tag{2}
\]

where \( \alpha \) and \( \beta \) denote the weights for the centerness loss \( L_{cen} \) and box regression loss \( L_{reg} \), respectively. The \( L_{cen} \) is the sigmoid loss and the \( L_{reg} \) is the GIOU loss [35].
Table 1. Zero-shot performance on LVIS [13] MiniVal5k datasets. AP_r / AP_c / AP_f indicate the AP values for rare, common, and frequent categories, respectively. “DH” and “F” in GLIP [25] baselines stand for the dynamic head [3] and cross-modal fusion, respectively. Baselines with * are implemented with our code base. GoldG+ denotes the GoldG plus the COCO [27] 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 $\Phi_c$ 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:

$$\mathcal{L}_{\text{cap}} = -\log p(y_{\text{det}}|\Phi_c(y_{\text{det}}), O)),$$

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

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

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

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 $\Phi_t$ outputs the concept embeddings $W^t \in \mathbb{R}^{2 \times D}$. Then the alignment scores $S_t \in \mathbb{R}^{k \times 2}$ 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 [28] which is pre-trained on ImageNet-1K [4]. We use 12 layers 8 heads transformer as our text encoder and load a base model checkpoint released by FILIP [42], in order to make a fair comparison with DetCLIP [41]. 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 [21] optimizer and set the batch size to 32. The learning rate is set to $1.4 \times 10^{-4}$ for the parameters of the image encoder and detection head, and $1.4 \times 10^{-5}$ 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 $1.4 \times 10^{-4}$. 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 Det-CLIP [41], we use Object365 [36] (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 [25] and DetCLIP [41], LVIS [13] MiniVal5k (defined in [19]) 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 [22] V1.2 and VG-COCO (defined in [37]). Following [37], we allocate 77398 images for training and 5000 images for validation and testing on VG. As demonstrated in [22], 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 [27] is only 7.1 vs. 35.4 in VG. Then an intersection of VG V1.2 and MS COCO is proposed by [37] 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 [13] which contains 1203 categories. Following [25, 41], 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 [36], the AP of the rare classes shows a valuable zero-shot detection performance. For the dense captioning task, we follow the setting of [17] to evaluate the VG and VG-COCO. The evaluation metric we adopt is the mean Average Precision proposed by [17] 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 [28] on the detection data Objects365 [36] and dense captioning data (VG [22]). 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 [22] dataset from GLIP-T, DetCLIP-T(C), and CapDet. All three models are trained on O365.
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 [37] by 2.5% on mAP on VG and TDC+ROCSU [37] by 2.08%, respectively. It is worth noticing that, even against given the ground-truth bounding boxes with the previous method COCG [26] 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.

Qualitative Visualizations. Figure 4 shows a qualitative visualization comparison between JIVC [40] 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

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCLN [17]</td>
<td>5.16</td>
</tr>
<tr>
<td>JIVC [40]</td>
<td>9.96</td>
</tr>
<tr>
<td>ImgG [26]</td>
<td>9.68</td>
</tr>
<tr>
<td>COCD [26]</td>
<td>9.75</td>
</tr>
<tr>
<td>COCG [26]</td>
<td>10.39</td>
</tr>
<tr>
<td>CAG-Net [43]</td>
<td>10.51</td>
</tr>
<tr>
<td>TDC [37]</td>
<td>11.90</td>
</tr>
<tr>
<td>CapDet (Ours)</td>
<td>15.44</td>
</tr>
</tbody>
</table>

Table 2. Comparison of mAP (%) performance on dense captioning benchmark on the VG V1.2 dataset.
<table>
<thead>
<tr>
<th>Method</th>
<th>mAP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCLN [17]</td>
<td>4.23</td>
</tr>
<tr>
<td>JIVC [40]</td>
<td>7.85</td>
</tr>
<tr>
<td>Max Pooling [26]</td>
<td>7.86</td>
</tr>
<tr>
<td>ImgG [26]</td>
<td>7.81</td>
</tr>
<tr>
<td>COCD [26]</td>
<td>7.92</td>
</tr>
<tr>
<td>COCG [26]</td>
<td>8.90</td>
</tr>
<tr>
<td>COCG-LocSiz [26]</td>
<td>8.76</td>
</tr>
<tr>
<td>COCG&amp;GT [26]</td>
<td>9.79</td>
</tr>
<tr>
<td>TDC+ROCSU [37]</td>
<td>11.9</td>
</tr>
<tr>
<td>CapDet (Ours)</td>
<td>13.98</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>MODEL</th>
<th>DC HEAD</th>
<th>AP</th>
<th>LVIS AP</th>
<th>AP_r / AP_c / AP_f</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLIP-T</td>
<td>✗</td>
<td>30.4</td>
<td>22.5 / 29.0 / 33.0</td>
<td></td>
</tr>
<tr>
<td>GLIP-T</td>
<td>✓</td>
<td>33.1</td>
<td>27.0 / 32.1 / 35.0</td>
<td></td>
</tr>
<tr>
<td>DETCLIP-T</td>
<td>✗</td>
<td>31.5</td>
<td>27.5 / 30.6 / 33.0</td>
<td></td>
</tr>
<tr>
<td>DETCLIP-T</td>
<td>✓</td>
<td>33.8</td>
<td>29.6 / 32.8 / 35.5</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Pre-training Data</th>
<th>Fine-tune</th>
<th>DCap mAP(%)</th>
<th>Box mAP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VG</td>
<td>✗</td>
<td>12.86</td>
<td>27.65</td>
</tr>
<tr>
<td>O365,VG</td>
<td>✓</td>
<td>4.72</td>
<td>9.65</td>
</tr>
<tr>
<td>VG</td>
<td>✓</td>
<td>13.83</td>
<td>28.58</td>
</tr>
<tr>
<td>O365,VG</td>
<td>✓</td>
<td>15.44</td>
<td>30.61</td>
</tr>
</tbody>
</table>

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

There 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/
References


