Fine-Grained Visual–Text Prompt-Driven Self-Training for Open-Vocabulary Object Detection

Yanxin Long, Jianhua Han, Runhui Huang, Hang Xu, Yi Zhu, Chunjing Xu, and Xiaodan Liang

Abstract—Inspired by the success of vision–language methods (VLMs) in zero-shot classification, recent works attempt to extend this line of work into object detection by leveraging the localization ability of pretrained VLMs and generating pseudolabels for unseen classes in a self-training manner. However, since the current VLMs are usually pretrained with aligning sentence embedding with global image embedding, the direct use of them lacks fine-grained alignment for object instances, which is the core of detection. In this article, we propose a simple but effective fine-grained visual-text prompt-driven self-training paradigm for open-vocabulary detection (VTP-OVD) that introduces a fine-grained visual–text prompt adapting stage to enhance the current self-training paradigm with a more powerful fine-grained alignment. During the adapting stage, we enable VLM to obtain fine-grained alignment using learnable text prompts to resolve an auxiliary dense pixelwise prediction task. Furthermore, we propose a visual prompt module to provide the prior task information (i.e., the categories need to be predicted) for the vision branch to better adapt the pretrained VLM to the downstream tasks. Experiments show that our method achieves the state-of-the-art performance for open-vocabulary object detection, e.g., 31.5% mAP on unseen classes of COCO.

Index Terms—Open-vocabulary object detection, prompt learning, self-training, vision–language.

I. INTRODUCTION

THE dominant object detection paradigm uses supervised learning to predict limited categories of the object. However, the existing object detection datasets usually contain only a few categories due to the time-consuming labeling procedure, e.g., 20 in PASCAL VOC [1] and 80 in COCO [2]. Some methods [3] attempt to expand the categories to a larger scale by adding more labeled datasets. However, it is time-consuming and difficult to guarantee sufficient instances in each class due to the naturally long-tailed distribution. On the other hand, previous methods [4], [5], [6], [7] follow the setting of zero-shot detection (ZSD) and align the visual embeddings with the text embeddings generated from a pretrained text encoder on base categories, but they still have a significant performance gap compared with the supervised counterpart.

Benefiting from the massive scale of datasets collected from the web [8], [9], recent vision–language pretraining (VLP) models, e.g., CLIP [10], have shown a surprising zero-shot classification capability by aligning the image embeddings with the corresponding caption. However, it is a challenging research direction to transfer this zero-shot classification ability to object detection in the dense prediction framework, because fine-grained pixel-level alignment, which is essential for dense tasks, is missing in the current visual–language pretraining models.

There exist some attempts to train an open-vocabulary detector by leveraging the zero-shot classification ability of a pretrained vision–language method (VLM). The work [11] proposes a basic self-training pipeline, which uses the activation map of the noun tokens in the caption from the pretrained VLM and generates the pseudo bounding-box label. However, the existing OVD methods just take advantage of the global text–image alignment ability of VLMs, thus failing to capture dense text–pixel alignment and severely hindering the self-training performance of OVD tasks. Directly using
them by activation map cannot fully adapt to downstream detection tasks, which require better dense representations. For example, as shown in the upper part of Fig. 1, directly using the pretrained VLM can only obtain a low-quality and incomplete dense score map for the dog in the input image, which is harmful for the next pseudolabeling stage. Recent works [12], [13] attempt to build generic object detection frameworks by scaling to larger label spaces, while they are costly required to acquire large-scale annotations from bigger datasets.

To achieve fine-grained alignment over dense pixels and avoid extra data cost, we propose visual–text prompt-driven self-training paradigm for open-vocabulary detection (VTP-OVD), a VTP-OVD, to improve the robustness and generalization capability. Inspired by the recent advance in the learning-based prompting methods [14], [15] from the natural language processing (NLP) community, we design a novel prompt-driven self-training framework (Fig. 2) for better adaptation of the pretrained VLM to new detection tasks. In detail, VTP-OVD designs a fine-grained visual–text prompt adapting stage to obtain more powerful dense alignments for better pseudolabel generation by introducing an additional dense prediction task. Specifically in the adapting stage, to reduce the domain gap of the upstream and downstream tasks and obtain the semantic-aware visual embedding, we introduce the visual and text prompt modules into the learnable image encoder and text encoder of the original VLM, respectively. The text prompts provide dense alignment task cues to enhance category embedding, and the visual prompt module aligns enhanced categories’ information to each pixel.

Furthermore, with the prompt-enhanced VLM after the adapting stage, a better pseudolabel generation strategy is proposed for novel classes by leveraging the nonbase categories’ names as the label input and aligning each connect region of the score map for each category. As far as we know, our VTP-OVD is the first work to use a learnable prompt-driven adapting stage in the OVD task to capture fine-grained pixelwise alignment and therefore to generate better pseudolabels.

We evaluate our VTP-OVD on the popular object detection dataset COCO [2] under the well-known zero-shot setting and further validate its generalization capability on the PASCAL VOC [1], Objects365 [16], and LVIS [3] benchmarks. The experimental results show that our VTP-OVD achieves the state-of-the-art performance on detecting novel classes (without annotations), e.g., 31.5% mAP on the COCO dataset. Besides, VTP-OVD also outperforms other open-vocabulary detection (OVD) methods when directly adapting the model trained on COCO to perform OVD on the three other object detection datasets. To demonstrate the effectiveness of the two learnable prompt modules, further experiments are conducted to show that fine-tuned with these two modules, VTP-OVD can generate a better dense score map (4.3% mIoU higher) compared with directing using the pretrained CLIP [10] model.

II. RELATED WORK

A. Zero-Shot Object Detection

Most ZSD methods [4], [5], [6], [7] align the visual embeddings to the text embeddings of the corresponding base category generated from a pretrained text encoder. Several methods introduce GCN [17], contrastive learning [18], and a novel “polarity loss” [19] to bridge the gap between the visual embeddings and text embeddings. Inspired by the success of VLMs [10], [20], several methods attempt to perform ZSD by training with image captions or directly leveraging a pretrained VLM. Zareian et al. [21] propose to pretrain the CNN backbone and a vision to language (V2L) module via the grounding task on an image caption dataset, and then the whole architecture is fine-tuned with an additional region proposal network (RPN) module. However, it still suffers from a large performance gap with the SOTA, and the domain of the upstream and downstream datasets must remain similar to maintain performance [11]. Based on a pretrained VLM, [22] distills the learned image embeddings of the cropped proposal regions to a student detector, and [23] further introduces an interembedding relationship matching loss to instill interembedding relationships. While each proposal needs to be fed forward into the image encoder of VLM, which requires huge computation costs. Gao et al. [11] propose a basic self-training pipeline based on ALBEF [20], which first uses Grad-CAM [24] to obtain the dense activation region of specific words in the caption and then generate the pseudo bounding-box label by selecting the proposal that has the largest overlap with the activation region. However, most VLMs lack the ability to perform pixelwise classification since it is pretrained with the correspondence of text embedding and global visual token instead of pixel embeddings. Besides, the reliance on caption data limits the application of this generalization to some datasets without caption. Inspired by [25], our method can obtain the pseudoannotations with the pixel-alignment patches.

B. Vision–Language Pretraining

The pretraining tasks of VLP models can be divided into two categories: image–text contrastive learning tasks and language modeling (LM)-based tasks. The first category, e.g.,
CLIP [10] and ALIGN [8], aims to align the visual feature with textual feature in a cross-modal common semantic space. The other category, e.g., VisualBERT [26], UNITER [9], M6 [27], DALL-E [28], and ERNIE-ViLG [29], uses LM-like objects, include both autoregressive LM (e.g., image captioning [30], VQA [31]) and masked LM (e.g., masked language/region modeling). Different from previous works that only focus on global feature alignment, our proposed model can achieve pixel-aware alignments via prompts and perform better in downstream tasks of pixelwise dense prediction.

C. Prompt Tuning

Freezing the pretrained models with only tuning the soft prompts can benefit from efficient serving and matching the performance of full model tuning. Prompt tuning has been verified as an effective method to mitigate the gap between pretraining and fine-tuning. As a rapidly emerging field in NLP [14], prompt tuning is originally designed for probing knowledge in pretrained language models [32] and now applied in various NLP tasks, e.g., language understanding [33], emotion detection [34], and generation [35]. Prompt tuning has now been extended to the vision–language models. Instead of constructing hand-crafted prompts in CLIP [10], CoOp [15] proposes tuning soft prompts with unified context and class-specific context in the downstream classification task. CPT [36] proposes colorful cross-modal prompt tuning to explicitly align natural language to fine-grained image regions. The main differences between our usage of learnable multimodal prompts with the previous prompt tuning works lie in three aspects: 1) unaligned upstream and downstream tasks; 2) multimodal; and 3) learnable prompts for self-training-based open-vocabulary object detection.

III. METHOD

In this section, we first briefly introduce the VTP-OVD framework. Then we describe the details of different stages in VTP-OVD: 1) learnable multimodal prompts in fine-grained alignment stage, which are used to enable VL model to obtain fine-grained pixelwise alignment ability; 2) better pseudolabel generation strategy in the pseudolabeling stage; and 3) the details of the final self-training stage.

Basic Notations: We construct a combined categories’ set (denoted as $C$) via extracting the categories from several large-scale detection datasets. The base (seen) categories, novel (unseen) categories, and nonbase categories are denoted as $C_B$, $C_N$, and $C_{\overline{B}}$ (i.e., $C = \{C_B, C_{\overline{B}}, C_N\}$). The input image is represented as $X$. We use $\Phi_s$ and $\Phi_t$ to represent the modified image encoder [37] and text encoder of the original VLM specifically.

A. VTP-OVD Framework

Fig. 3 illustrates the overall pipeline of the proposed VTP-OVD. We use different colors to indicate the parameter fixing or requiring training as clarified in the upper right corner of Fig. 3. First, at the fine-grained visual–text prompt adapting stage, using the learnable text prompt module prompt, and visual prompt module prompt, into the text encoder $\Phi_t$ and image encoder $\Phi_s$, we can fine-tune the whole network under dense supervision of the pseudo dense score map with the other parameters fixed. Second, for pseudolabeling stage, we obtain the pseudolabels of nonbase classes by leveraging the dense classification ability of the fine-tuned VL model and the location ability from a pretrained RPN. The complete training procedure can be found in Algorithm 1.

B. Adapting Stage via Multimodal Prompts

The crucial part of the self-training pipeline for OVD lies in the pseudolabeling stage that determines the final detection performance of novel classes. To generate better pseudolabels for the detection task, our method first aligns each pixel with a category and then adopts a pretrained RPN for bounding-box localization. Since most VLMs are trained via the alignment
between the whole image and the corresponding caption, they lack the dense alignment ability between pixels and categories. Thus, in this section, we focus on modifying the pretrained vision–language model to enhance the current self-training paradigms with fine-grained adapting stage via a newly designed dense alignment loss function and learnable text/visual prompts. Note that to obtain the dense pixel-level visual embeddings instead of the global visual embedding, “c” represents the concatenation operation.

1) Dense Alignment Task: As shown in Fig. 3, we introduce a dense alignment task to fine-tune the VLM under the supervision of the pseudo dense score map \( S \in \mathbb{R}^{H \times W \times |C|} \), where \( C \) denotes all the categories, including multidataset categories, and \(|C|\) represents the number of \( C \). \( S \) is generated by calculating the similarity of pixelwise visual embeddings of input image \( X \in \mathbb{R}^{H \times W} \) with each category via the original VLM.

The dense score map \( S \) is calculated by

\[
V = \Phi_v(X), \quad T = \Phi_t(C), \quad S = \left\{ \frac{V}{\|V\|}, \frac{T}{\|T\|} \right\}
\]

where \( \langle \cdot, \cdot \rangle \) denotes the inner product operation, and \( \| \cdot \| \) denotes the normalization operation. \( V \in \mathbb{R}^{H \times W \times D} \) and \( T \in \mathbb{R}^{|C| \times D} \) represent the visual embeddings and text embeddings specifically. To reduce the domain gap of the upstream and downstream tasks and obtain the semantic-aware visual embedding, we further introduce the visual and text prompt modules into the learnable image encoder and text encoder of the original VLM, respectively. Benefiting from the same task form as the upstream tasks, most previous prompt-based tuning works \([14],[15],[38]\) still perform well on the downstream tasks with different data domains. However, our VTP-OVD uses learnable multimodal prompts to obtain the fine-grained feature alignment, which is different from the upstream global alignment in the task domain. The objective of the dense alignment stage is

\[
L_{FT} = L_{CE}(S, \hat{S})
\]

where \( L_{CE} \) represents the cross-entropy loss. Similar to (1), the model prediction \( \hat{S} \in \mathbb{R}^{H \times W \times |C|} \) with the image input \( X \) is calculated as

\[
\hat{S} = \left\{ \frac{V'}{\|V'\|}, \frac{T'}{\|T'\|} \right\}
\]

where \( V' \in \mathbb{R}^{H \times W \times D} \) and \( T' \in \mathbb{R}^{|C| \times D} \) denote the image feature and text feature integrated with prompt embeddings, respectively.

2) Text Prompt: Prompt tuning has been verified as an effective method to probe knowledge in pretrained language models \([32]\) and applied in various NLP tasks \([33],[35]\). For example, the hand-crafted text prompts, e.g., “a photograph of a {},” have been adopted to adapt the pretrained VLM to different downstream tasks \([10],[22]\). Inspired by these works, we further introduce a learnable text prompt encoder to provide dense alignment task cues for enhancing categories’ embeddings since the downstream text–pixel alignment task is quite different from the upstream global text–image alignment task.

As shown in Fig. 4(a), the text prompt encoder maps the learnable prompt tokens \( Q \in \mathbb{R}^{M \times D} \) into the prompt embeddings \( H \in \mathbb{R}^{M \times D} \). \( M \) denotes the total number of prompt tokens. We believe that \( h_i \in H \) should depend on each other since the words in a natural sentence always have a strong contextual relationship. Therefore, for the structure of the proposed text prompt encoder module, it contains a bidirectional long short-term memory network (LSTM) to get the contextual information followed by an ReLU activated two-layer multilayer perceptron (MLP) \([39]\):

\[
H = \text{MLP}(\text{LSTM}(Q)).
\]

After that, we concatenate the same text prompt embeddings with each class token \( c \in C \) to obtain the prompt-enhanced text embedding \( t' \in T' \) by

\[
t' = \Phi_t([\{h_i\}_{i=0}^{L}, c, \{h_i\}_{i=L+1}^{M}])
\]

where \( [\cdot] \) denotes the concatenation operation, and \( L \) denotes the length of prompt tokens added before the category token.

3) Visual Prompt: The text prompts use dense alignment task cues to enhance category embedding. Therefore, to align the enhanced categories’ information to each pixel, a visual prompt encoder is built upon the cross-attention mechanism to generate semantic-aware visual embedding for each pixel to improve dense alignment. Similar to ActionCLIP \([40]\), which uses positional embeddings as visual prompt to provide...
Algorithm 1 VTP-OVD Pseudocode

**Input:** Images $X$, Base categories $C_B$, Nonbase classes $C_{\overline{B}}$, Image encoder $\Phi_i$, Text encoder $\Phi_t$, Pretrained RPN model $\Phi_r$, Open-vocabulary detector $\Phi_d$, Visual prompt $prompt_v$, Text prompt $prompt_t$, Fine-tuning iter $Iter_f$, Training iter $Iter_t$, Inner product function $\langle \cdot, \cdot \rangle$

$\Phi_r, \Phi_t \leftarrow$ initialized w/ CLIP pretrained

# 1. Fine-grained visual–text prompt adapt stage

$c \leftarrow$ all classes $C_B + C_{\overline{B}}$

$V, T \leftarrow$ image feature and text feature $\Phi_i(X), \Phi_t(C)$

$S \leftarrow$ pseudo dense score map $\langle \frac{V}{||V||}, \frac{T}{||T||} \rangle$

$V' \leftarrow$ add random-initialed $prompt_v$

$T' \leftarrow$ add random-initialed $prompt_t$

for $t = 1 \rightarrow Iter_f$

$\hat{S} \leftarrow$ model prediction $\langle \frac{V'}{||V'||}, \frac{T'}{||T'||} \rangle$

$\mathcal{L}_F T \leftarrow$ cross-entropy loss $\mathcal{L}_{CE}(S, \hat{S})$

$V', T' \leftarrow$ updated by $\mathcal{L}_F T$

end for

# 2. Pseudolabeling stage

$P \leftarrow$ generated proposals $\Phi_i(X)$

$\hat{S}_{\overline{B}} \leftarrow$ pseudo nonbase dense score map $\langle \frac{V'}{||V'||}, \frac{T_{\overline{B}}}{||T_{\overline{B}}||} \rangle$

$R \leftarrow$ connected regions of nonbase categories $\hat{S}_{\overline{B}}$

pseudo labels $\leftarrow$ IoU($R, P) < \gamma$

# 3. Self-training stage

for $t = 1 \rightarrow Iter_t$

$\hat{Y} \leftarrow$ model prediction $\Phi_d(X)$

$Y \leftarrow$ ground truths (for $C_B$) and pseudolabels (for $C_{\overline{B}}$)

$\mathcal{L}_T \leftarrow \mathcal{L}_{CLS}(Y, \hat{Y}) + \mathcal{L}_{REG}(Y, \hat{Y})$

$\Phi_d \leftarrow$ updated by $\mathcal{L}_T$

end for

additional temporal information, ours is also formed as a post-network prompt module [40] to provide additional semantic cues and we only fine-tune the visual prompt module while fixing the visual encoder. By taking each pixel embedding in visual features $V = \Phi_i(X)\in \mathbb{R}^{HW\times D}$ as the input of the query and the text embeddings of all the categories $T' \in \mathbb{R}^{|C|\times D}$ as the input of key and value, the cross-attention block outputs the semantic-aware visual prompts $\tilde{V} \in \mathbb{R}^{HW\times D}$ for each pixel. Then they are concatenated with the visual features followed by a one-layer MLP to obtain the semantic-aware visual embeddings $V' \in \mathbb{R}^{HW\times D}$ [see Fig. 4(b)]

$$V' = MLP(V \oplus \tilde{V})$$

$$\tilde{V} = \text{softmax} \left( \frac{V \times T'^T}{\sqrt{D}} \right) \cdot T'$$

(6)

where $\oplus$ and $D$ represent the concatenation operation and the feature dimension of text embeddings $T'$, respectively. MLP denotes the one-layer multilayer perceptron to reduce back the feature dimension.

C. Pseudolabel Generation

Based on the more precise dense score map generated by the fine-tuned VLM with learnable text and visual prompts, we can obtain the nonbase classes’ pseudolabels by additionally leveraging the location ability of a pretrained RPN $\Phi_r$ [shown in Fig. 3(b)]. Note that $\Phi_r$ is only trained on base classes, and the experiments in [22] already demonstrate that training only on the base categories can achieve comparable recall to average recall of the overall categories. Previous pseudolabeling generation strategy [11] used the objects of interest in the caption, which not only harms the transfer ability to the detection datasets without the caption but also is limited to the uncompleted description of caption data. For example, such a strategy cannot generate the pseudolabels of classes that are not included in the caption.

To address these issues, we directly use nonbase classes’ names as the input of fine-tuned text encoder $\Phi_t$. Besides, to better distinguish the background classes, instead of directly using the word “background,” we treat all the base classes as background since we already have the ground-truth annotation of these classes. Then we obtain the pseudo dense score map of nonbase classes $\hat{S}_{\overline{B}} \in \mathbb{R}^{HW\times|C_{\overline{B}}|}$ from the fine-tuned VLM with the procedure same as (3). For each image $x_i$, we first compute the connected regions $r_j \in \mathbb{R}^{HW}$ for the $j$th nonbase category on $\hat{S}_{\overline{B}}$ by setting a similarity threshold $\delta$, and then we adopt the intersection of union (IoU) of the $k$th proposal $p_k^i$ from RPN and $r_j$ as confidence score

$$\text{score}_{i,j,k} = \text{IoU}(p_k^i, r_j)$$

(7)

where $p_k = \Phi_i(x_i)$, and $\text{score}_{i,j,k}$ denotes the confidence score that the proposal $p_k$ belongs to the $j$th nonbase category for image $x_i$. Finally, a hard score threshold $\gamma$ is adopted on the confidence score to filter out pseudoboxes with less confidence.

**Self-Training Stage:** After obtaining the pseudo bounding-box labels for nonbase classes, together with the ground-truth annotations of base classes, we are able to train a final open-vocabulary detector $\Phi_d$ to fulfill the self-training pipeline [shown in Fig. 3(c)] via Faster-RCNN [41]. We build $\Phi_d$ by replacing the last classification layer with the text embeddings generated by the text encoder on all the categories $V'$. The objective of the self-training stage $\mathcal{L}_T$ is calculated as

$$\mathcal{L}_T = \mathcal{L}_{CLS}(Y, \hat{Y}) + \mathcal{L}_{REG}(Y, \hat{Y})$$

(8)

where $\mathcal{L}_{CLS}$ and $\mathcal{L}_{REG}$ denote the cross-entropy loss for the classification head and L1-loss for regression head, respectively, and $\hat{Y}$ and $Y$ represent the model prediction $\Phi_d(X)$ and annotations combined with pseudolabels of nonbase classes and the ground truth of base classes. The trained detector $\Phi_d$ can then transfer to other detection datasets by providing the class names.

IV. EXPERIMENTS

A. Open-Vocabulary Object Detection Setups

1) Benchmark Setting: We benchmark on the widely used object detection dataset COCO 2017 [2]. Following the well-known settings [4] adopted by many OVD methods, we divide the categories used in COCO into 48 base (seen) classes and 17 novel (unseen) classes. Each method can only be trained on annotations of the base categories and then it predicts the novel categories without seeing any annotations of these categories. We report the detection performance of both the base and
novel classes during inference, as the generalized settings used in [21]. We also evaluate the generalization ability of the trained detector by directly transferring it to three other object detection datasets, including PASCAL VOC [1], LVIS [3], and Objects365 [16].

2) Evaluation Metric: We use mean average precision (mAP) with IoU threshold 0.5 as the evaluation metric. We pay more attention to the novel class performance since we aim to build an open-vocabulary detector, and the annotations of base classes are already provided.

B. Implementation Details

We use a pretrained CLIP (RN50 × 16) [10] model as the VLM. Note that our model is compatible with the conventional VL models. For simplicity, we take CLIP as an example. All the detection models are implemented on the mmdet [42] codebase and follow the default setting (i.e., 1x schedule unless otherwise mentioned. At the adapting stage, the prompt-enhanced CLIP is trained for five epochs, with the text prompt learning rate set to 1e−1 and the visual prompt set to 1e−5 separately. For the pseudolabeling stage, the RPN is trained only on the base classes in the 2x schedule. The objectness score threshold for RPN is set to 0.98. The similarity threshold δ and score threshold γ are set to 0.6 and 0.4, respectively. For the self-training stage, following the default setting, we adopt a Mask-RCNN (ResNet50) [46] with the last classification layer replaced by class embeddings output by the text encoder. We train the detector for 12 epochs, with the learning rate decreased by a factor of 0.1 at 8 and 11 epochs with 8 V100 GPUs. The initial learning rate is set to 0.04 with batch size 32, and the weight decay is set to 1e−4. We extract the categories of large-scale object detection dataset (i.e., LVIS [3] and Objects365 [16]) as the combined categories set C, which contains about 1k categories. We generate non-base pseudolabels for training, while evaluate only on novel categories and base categories following [21].

C. Main Results

1) COCO Dataset: We compare our VTP-OVD with the existing open-vocabulary methods on the COCO dataset [2]. As shown in Table I, VTP-OVD achieves the state-of-the-art performance (i.e., 31.5% mAP) on the novel categories of the COCO dataset and 4.5% mAP improvement on overall categories compared with another self-training based method OVD-ALBEF [11]. Besides, by comparing with the VTP-OVD (w/o FT), e.g., 31.5% mAP versus 29.8% mAP, we further demonstrate the necessity of the fine-grained adapting stage with learnable visual and text prompts for dense alignment tasks. Note that we use the basic vision–language model CLIP [10] and do not use the caption data, guaranteeing the transfer ability to other VLMs or pretraining datasets without caption information. Clarification should be made that the self-training-based methods (OVD-ALBEF [11] and ours) usually achieve slightly worse on base classes compared with the knowledge-distillation (KD) method (ViLD [22]). We attribute this to the reason that the self-training methods make the model optimize more toward novel classes via generating massive pseudolabels on them, while the KD-based

| METHOD | VL MODEL | USING COCO CAPTION | NOVEL | BASE | OVERALL |
|--------|-----------|---------------------|-------|------|---------|
| SB [4] | x         | x                   | 0.31  | 29.2 | 24.9    |
| LAB [4] | x         | x                   | 0.22  | 20.8 | 18.0    |
| DSES [4] | x      | x                   | 0.27  | 26.7 | 22.1    |
| DELO [43] | x    | x                   | 3.41  | 13.8 | 13.0    |
| PL [44] | x         | x                   | 4.12  | 35.9 | 27.9    |
| OVR-CNN [21] | -    | ✓                   | 22.8  | 46.0 | 39.9    |
| ViLD* [22] | CLIP | x                   | 27.6  | 59.5 | 51.3    |

Self-training based open-vocabulary detection methods

| METHOD | VL MODEL | USING COCO CAPTION | NOVEL | BASE | OVERALL |
|--------|-----------|---------------------|-------|------|---------|
| OVD-ALBEF [11] | ALBEF | ✓                   | 30.8  | 46.1 | 42.1    |
| DETIC [12] | -       | x                   | 27.8  | 47.1 | 45.0    |
| VTP-OVD (w/o FT) | CLIP | x                   | 29.8  | 51.8 | 46.1    |
| VTP-OVD (OURS) | CLIP | x                   | 31.5  | 51.9 | 46.6    |

Table I: OVD performance comparison on the COCO datasets. VTP-OVD (w/o FT) denotes the VTP-OVD without fine-grained adapting stage. We can observe that VTP-OVD achieves the state-of-the-art detection performance on the novel classes without the fine-tuning stage, the performance on the novel classes suffers a large drop. Note that OVR-CNN pretrains a self-designed vision–language model with a V2L module by itself. ViLD* trains the detector with data augmentations of large-scale jittering (LSJ) [42] and longer training schedule (16×). mAP (%) is reported.
methods try to keep the performance of the base class by only distilling the classification ability.

2) Generalization Abilities: To further demonstrate the open-vocabulary ability of the detector trained through the VTP-OVD pipeline, we directly transfer the final detection model trained on the COCO datasets, to other detection datasets, including PASCAL VOC [1], Object365 [16], and LVIS [3]. Benefit from the training on nonbase pseudolabels, the experimental results in Table II show that our VTP-OVD achieves the best generalization ability even adopted to the datasets with much more categories than the pretrained dataset, i.e., 81 classes in COCO versus 1203 classes in LVIS. Note that we do not compare with ViLD [22] since it does not provide either this experimental result or the code.

D. Ablation Study

1) What Is the Effect of the Learnable Prompt Modules for Pseudolabel Generation?: We introduce two learnable prompt modules to obtain the dense alignment ability,
which is important for pseudo bounding-box label generation. To analyze the effect of the proposed visual and text prompts, we conduct the quantitative experiments with different combinations of prompt modules to show the quality of pseudolabels (with mAP) and dense score map (with mIoU on novel classes). As shown in Table III, adopting hand-crafted prompts improves the baseline (without prompt) by 1.2% mAP increment on pseudolabels’ quality, while adding text and visual prompts separately gains additional 1.7% and 1.4% increment, respectively. Combining visual and text prompts makes better performance (+3.5%) and the best dense score map (35.4% mIoU). After adopting the self-training stage, it reaches the state-of-the-art detection performance (31.5% mAP) on COCO novel classes.

2) What Is the Effect of Different Text Prompt Structures?:
As shown in Table IV, we evaluate the effect of different structures of only the text prompts. Based on the input learned embeddings, separately using LSTM and MLP obtains a relatively small improvement due to the lack of the contextual information of prompt tokens. Adopting both LSTM and MLP for the association of prompt tokens achieves the best performance, i.e., 25.4% mAP for generated pseudolabels on novel classes of COCO. The ablations of the category token position and prompt token number which are decided by $L$ and $M$ in (4) are shown in Table V.

3) How the Pseudo labeling Robust to Hyperparameters?:
As shown in Fig. 5, we conduct the ablation studies on the two most important hyperparameters, including the similarity
threshold $\delta$ for the dense score map to compute connected regions and objectness threshold for RPN to filter the low-confident proposals, in the fine-tuning stage. Note that when we ablate one parameter, we keep the other fixed as the default value. Observation can be made that our pseudolabeling strategy is robust to these two parameters and achieves the highest when similarity threshold and objectness threshold are set to 0.6 and 0.98, respectively. Low similarity and objectness threshold bring much false positives, while high similarity and objectness threshold decrease the overall recall.

E. Qualitative Results

1) Qualitative Examples: We visualize the final detection performance of VTP-OVD on the novel classes of the COCO dataset and the transfer performance on Objects365, LVIS, and Pascal VOC in Fig. 6. We can observe that our VTP-OVD can obtain high-quality bounding boxes for novel classes on COCO even without any annotation of these categories. Besides, as shown in the third row of Fig. 6, our VTP-OVD can detect rare classes such as fireplug and polar bear, demonstrating its OVD ability.

However, we also find our VTP-OVD fails in detecting several novel categories, including skateboard, snowboard, etc. during the pseudolabeling stage and self-training stage. By observing the dense score map and the pseudolabels of skateboard class in Fig. 7, we attribute this phenomenon to the generation of wrong pseudolabels.

To explain the reason, we think that some objects are often related to the environment, such as skateboards, skateboarders, and skateboard playground often appear together, thus trained with the alignment based on global image information and text embeddings, and the current vision–language model usually cannot distinguish these objects from other objects in same scene.

2) Effect of Prompt Modules: In Fig. 8, we visualize the effect of proposed prompt modules on the generation of pseudolabels of novel classes. By comparing the dense score maps in the third and sixth columns, observation can be made that through learnable visual–text prompts, the VLM can achieve a better dense alignment and improve the quality of pseudolabels. Specifically, adopting the visual prompt tends to fill the dense alignment of the object, while adopting the text prompt usually explores the new classes or objects. Combining these two prompts makes the better performance to generate more precise pseudo bounding boxes.

We also visualize the distribution of the pixel embeddings generated by the vision–language model CLIP (with or without learnable prompt modules) via t-SNE [47] on novel classes of the COCO dataset in Fig. 9. Observation can be made that adopting learnable prompt modules helps cluster the pixel embeddings in the same class and separate pixel embeddings from different categories, demonstrating their effects on dense alignments.

V. Conclusion

In this article, we propose a novel open-vocabulary pipeline named VTP-OVD. VTP-OVD introduces a new fine-tuning stage to enhance the self-training paradigm with dense alignments by adopting two learnable visual and text prompt modules. The experimental results show that VTP-OVD achieves the state-of-the-art performance on the novel classes of the COCO datasets and the best transfer performance when directly adapting the model trained on COCO to the PASCAL VOC, Object365, and LVIS datasets. Additional experiments also show that after fine-tuning with two learnable prompt modules, VTP-OVD obtains a more precise dense score map (4.3% higher on mIoU) on novel classes. Nevertheless, VTP-OVD is a general pipeline of adopting the pretrained vision–image encoder to dense prediction tasks, which can be easily extended to other tasks, e.g., open-vocabulary instance segmentation. We hope VTP-OVD can serve as a strong baseline for future research on different open-vocabulary tasks.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.
A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, “Open-vocabulary object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5356–5364.

B. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, “Zero-shot object detection,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 384–400.

B. Demirel, R. G. Cinbis, and N. Izkizer-Cinbis, “Zero-shot object detection by hybrid region embedding,” 2018, arXiv:1805.06157.

S. Rahman, S. Khan, and N. Barnes, “Transductive learning for zero-shot object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6082–6091.

H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. S. Khan, “Synthesizing the unseen for zero-shot object detection,” in Proc. Asian Conf. Comput. Vis., 2020, pp. 1–16.

C. Jia et al., “Scaling up vision and language-representation learning with noisy text supervision,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 4904–4916.

Y. Chen et al., “UNITER: Universal image-text representation learning,” in Proc. 16th Eur. Conf. Comput. Vis. (ECCV), in Lecture Notes in Computer Science, Glasgow, U.K., vol. 12375, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Springer, Aug. 2020, pp. 104–120.

A. Radford et al., “Learning transferable visual models from natural language supervision,” 2021, arXiv:2103.00020.

M. Gao et al., “Towards open vocabulary object detection without human-provided bounding boxes,” 2021, arXiv:2111.09452.

Z. Yu, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, “Detecting twenty-thousand classes using image-level supervision,” 2022, arXiv:2201.02605.

C. Feng et al., “PromptDet: Towards open-vocabulary detection using uncurated images,” 2022, arXiv:2203.16513.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” 2021, arXiv:2107.13586.

K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” 2021, arXiv:2109.01134.

S. Shao et al., “Object365: A large-scale, high-quality dataset for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8430–8439.

C. Yan, Q. Zheng, X. Chang, M. Luo, C.-H. Yeh, and A. G. Hauptmann, “Semantics-preserving graph propagation for zero-shot object detection,” IEEE Trans. Image Process., vol. 29, pp. 8163–8176, 2020.

C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, “Semantics-guided contrastive network for zero-shot object detection,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jan. 4, 2022, doi: 10.1109/TPAMI.2021.3140070.

S. Rahman, S. Khan, and N. Barnes, “Polarity loss: Improving visual-semantic alignment for zero-shot detection,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 30, 2020, doi: 10.1109/TNNLS.2022.3184821.

J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 9694–9705.

A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, “Open-vocabulary object detection using captions,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14393–14402.

X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” 2021, arXiv:2104.13921.

H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. S. Khan, “Bridging the gap between object and image-level representations for open-vocabulary detection,” in Proc. Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 33781–33794.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.

Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, “From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation,” IEEE Trans. Neural Netw. Learn. Syst., early access, Feb. 10, 2022, doi: 10.1109/TNNLS.2022.3145602.
Jianhua Han received the bachelor’s and master’s degrees from Shanghai Jiao Tong University, Shanghai, China, in 2016 and 2019, respectively. He is currently a Researcher with the Huawei Noah’s Ark Lab, Huawei Technologies Company Ltd., Shenzhen, China. His research interests lie primarily in deep learning and computer vision.

Runhui Huang received the bachelor’s degree from the School of Intelligent Systems Engineering, Sun Yat-sen University at Shenzhen, Shenzhen, China, in 2021, where he is currently pursuing the master’s degree. He works with the Human Cyber Physical Intelligence Integration Lab, China, under the supervision of Prof. Xiaodan Liang. His research interests include vision-and-language pretraining.

Hang Xu received the B.Sc. degree from Fudan University, Shanghai, China, in 2012, and the Ph.D. degree in statistics from The Hong Kong University, Hong Kong, in 2018. He is currently a Senior Researcher with Huawei Noah’s Ark Lab, Shanghai. He has published more than 50 papers in top-tier AI conferences, such as NeurIPS, CVPR, ICCV, and ICLR and some statistics journals, e.g., *Computational Statistics and Data Analysis* and *Statistical Computing*. His research interests include machine learning, vision-language models, object detection, and AutoML.

Yi Zhu received the B.S. degree in software engineering from Sun Yat-sen University at Guangzhou, Guangzhou, China, in 2013. She is currently pursuing the Ph.D. degree in computer science with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China. Her current research interests include object recognition, scene understanding, weakly supervised learning, and visual reasoning.

Chunjing Xu received the bachelor’s degree in math from Wuhan University, Wuhan, China, in 1999, the master’s degree in math from Peking University, Beijing, China, in 2002, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2009. He was an Assistant Professor and an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China. He became the Director of the computer vision laboratory in Huawei Noah’s Ark Lab, Central Research Institute, Shanghai, China, in 2017. His main research interests include machine learning and computer vision.

Xiaodan Liang received the Ph.D. degree from Sun Yat-sen University at Shenzhen, Shenzhen, China, in 2016, advised by Liang Lin. She was a Post-Doctoral Researcher with the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA, from 2016 to 2018, working with Prof. Eric Xing. She is currently an Associate Professor with Sun Yat-sen University. She has published several cutting-edge projects on human-related analysis, including human parsing, pedestrian detection and instance segmentation, 2-D/3-D human pose estimation, and activity recognition.