Unsupervised Prompt Tuning for Text-Driven Object Detection

Weizhen He$^{1,4}$, Weijie Chen$^{1,2,3,4}$, Binbin Chen$^2$, Shicai Yang$^2$, Di Xie$^{2,3}$, Luojun Lin$^4$, Donglian Qi$^1$, Yueting Zhuang$^1$

$^1$Zhejiang University, $^2$Hikvision Research Institute, $^3$Key Laboratory of Peace-building Big Data of Zhejiang Province, $^4$Fuzhou University

{hewz, chenweijie, qidl, yzhuang}@zju.edu.cn, {chenbinbin8, yangshicai, xiedi}@hikvision.com, linluojun2009@126.com

Abstract

Grounded language-image pre-trained models have shown strong zero-shot generalization to various downstream object detection tasks. Despite their promising performance, the models rely heavily on the laborious prompt engineering. Existing works typically address this problem by tuning text prompts using downstream training data in a few-shot or fully supervised manner. However, a rarely studied problem is to optimize text prompts without using any annotations. In this paper, we delve into this problem and propose an Unsupervised Prompt Tuning framework for text-driven object detection, which is composed of two novel mean teaching mechanisms. In conventional mean teaching, the quality of pseudo boxes is expected to optimize better as the training goes on, but there is still a risk of overfitting noisy pseudo boxes. To mitigate this problem, 1) we propose Nested Mean Teaching, which adopts nested-annotation to supervise teacher-student mutual learning in a bi-level optimization manner; 2) we propose Dual Complementary Teaching, which employs an offline pre-trained teacher and an online mean teacher via data-augmentation-based complementary labeling so as to ensure learning without accumulating confirmation bias. By integrating these two mechanisms, the proposed unsupervised prompt tuning framework achieves significant performance improvement on extensive object detection datasets.

1. Introduction

Object detection, which aims to locate and classify objects in an image, is a very fundamental task in computer vision. Recently, with the development of vision-language foundation models, object detection tends to be an open-vocabulary by learning knowledge from large-scale heterogeneous image-text data. Grounded Language-Image Pre-training (GLIP) [23] is one of the leading models, which detects target objects directly with the pre-defined prompts using category name (such as “[CLS],[CLS],...,[CLS]” in Fig. 1a) without being optimized by task-specific training data. Although those pre-trained models are endowed with a promising zero-shot generalization ability, it is crucial and necessary to transfer the knowledge to various downstream tasks [47, 42, 35].

An emerging trend to solve this problem is Prompt Tuning [51, 50, 21, 39, 25], which freezes the main body of the
network and merely optimizes text prompts (prompt embeddings) using downstream training data in a few-shot or fully supervised manner. However, this paradigm requires training data with annotations, which violates the original intention of zero-shot inference. As shown in Fig. 1, it naturally comes a question: can we conduct prompt tuning with the exposure of downstream data without human labeling?

In this paper, we take GLIP as an example pre-trained model to study this problem. To the best of our knowledge, this is the first attempt in this field to study unsupervised prompt tuning for text-driven object detection. Preliminarily, the baseline framework is developed from the mean teacher based self-training [30], facilitating teacher-student mutual learning. Since only text prompts are allowed to be optimized, both teacher and student share the same frozen network (text encoder and image encoder). Specifically, given the task-specific pre-defined prompts as initialization, the student is the network with learnable prompts, and the teacher is the one with momentum prompts. In this way, the teacher model annotates the unlabeled images to drive student prompt tuning, and then the teacher prompt is updated by the student prompt in an exponential moving average (EMA) manner. However, the pseudo boxes are inevitably noisy, causing the student to suffer from the noisy pseudo boxes, which in turn affects the teacher. To address this issue, we advance the conventional mean teaching process into two simple yet effective variants:

**Nested Mean Teaching (NMT).** The performance of the teacher model is equivalent to the quality of pseudo boxes. From this perspective, learning a good mean teacher can be formulated as a “learning to annotate” problem. Another insight is that mean teaching has been proven effective to optimize pseudo label, a next $k$-step mean teacher can naturally provide high-quality pseudo label, which driving the current timestamp to avoid overfitting on noisy label. Inspired by the above consideration, we aim to learn a delayed-annotator in a nested $k$-step mean teacher optimization manner, which comprises a nested inner loop and an outer loop. As shown in Fig. 2b, the inner loop acts as an annotator optimization process, which nests $k$-step ghosted teacher-student mutual learning to achieve better pseudo boxes. Note that both teacher and student models are discarded after inner-loop training, and only the pseudo labels are propagated to the outer loop. The outer loop optimizes the teacher in an EMA manner by taking student learning as a bridge using the pseudo boxes from the nested inner loop. These two loops are interpretable. Since the quality of pseudo boxes is expected to be better during teacher-student mutual learning, the $k$-step ghosted optimization in the nested inner loop provides better pseudo boxes to drive the optimization of the outer loop mean teacher.

**Dual Complementary Teaching (DCT).** Since unsupervised prompt tuning is an optimization process without using any ground-truth annotations, there exist risks of accumulating confirmation bias, which induces false negatives or false positives in an avalanche during teacher-student mutual learning. To mitigate this problem, we develop Dual Teachers, of which an offline teacher (pre-defined prompts) accounts for providing true-positive cues to ensure learning without forgetting true positives, and an online teacher (momentum prompts) learns to recall false negatives. To further promote the collaboration of Dual Teachers, we propose a data-augmentation-based complementary labeling mechanism. The offline teacher initializes sufficient true-positive boxes by feeding weakly-augmented images, while the online teacher recalls false negatives cautiously by feeding strongly-augmented images which is a strict access condition for the introduction of new pseudo boxes so as to avoid cumulatively introducing false-positive boxes.

These two mean teaching mechanisms are orthogonal to each other. We build the Unsupervised Prompt Tuning framework (UPT) by integrating them, where the Dual Complementary Teaching process is optimized in a nested annotation manner. Extensive ablation studies and experiments are carried out on multiple downstream object detection tasks, i.e., Cityscapes [4], Foggy Cityscapes [37], KITTI [11], Sim10K [19], BDD100K [46], WaterColor [17], MS COCO [28], Pasval VOC [7], Ego-Hands [23], and Pistols [23], which vary in dataset scale, categories, context and style shifts, demonstrating the effectiveness of the proposed framework. The main contributions of this paper can be summarized as follows:

- For the first time, we propose a challenging yet meaningful task, namely unsupervised prompt tuning for text-driven object detection, which fills the blank in object detection and pushes the limits of zero-shot inference.
- We build the unsupervised prompt tuning framework by developing two novel mean teaching methods, i.e., Nested Mean Teaching and Dual Complementary Teaching, which advance the conventional mean teaching process from the perspectives of optimizing annotation and learning without accumulating confirmation bias.
- Extensive experiments on numerous downstream object detection datasets demonstrate that the proposed framework can achieve significant performance improvement.

2. Related Works

2.1. Vision-Language Models

Vision-language models show great potential to learn generic visual representations and allow zero-shot transfer to downstream tasks. CLIP [36, 15], FLIP [45] and ALIGN [18] perform cross-modal contrastive learning on image-text pairs. Thanks to the flexibility of language,
the pretrained model can directly perform open-vocabulary image classification. In object detection, ViLD [13] and HierKD [34] distill the knowledge from CLIP into two-stage and one-stage detectors, respectively. Different from them, MDETR [20], GLIP [23], and VLDet [26] directly perform alignment between text and objects. All of them show promising zero-shot performance in downstream object detection tasks. Instead of grounded vision-language pre-training, this paper focuses on push the limits of zero-shot inference via unsupervised prompt tuning so as to adapt downstream tasks without supervision.

2.2. Prompt Tuning

Although vision-language models show promising zero-shot performances [3, 24, 43], they are heavily conditioned on the language input, known as text prompt. However, designing an appropriate prompt requires senior domain expertise, which is very costly since different downstream tasks need different designs. As an alternative, supervised prompt tuning exploits labeled training data to tune context in prompt [51, 50] or introduce feature adapters [9, 48]. Unlike supervised one, our method tunes the prompt without using any annotations, which meets the original intention of zero-shot inference. Beyond our method, there exist unsupervised prompt tuning works [16] on image classification. To the best of our knowledge, our work is the first attempt on object detection.

2.3. Mean Teaching

Mean Teaching is commonly used in semi-supervised learning [40, 30] and self-supervised learning [10, 12, 41, 27], which contains a teacher for pseudo labeling and a student to improve the teacher model by updating knowledge in an EMA manner. Most previous works focus on the designs of data augmentation [30, 8, 49], pseudo label generator [8, 29], and class-balanced training [30, 44, 1, 14, 22] to improve Mean Teaching. In contrast, we advance Mean Teaching into Nested Mean Teaching and Dual Complementary Teaching to avoid overfitting noisy pseudo boxes.

3. Method

In this section, we first introduce the preliminary knowledge about text-driven object detection and prompt tuning. We then present the proposed Nested Mean Teaching framework, which adopts nested-annotation to supervise teacher-student mutual learning in a bi-level optimization manner. Afterward, we propose Dual Complementary Teaching to create complementary labels from Dual Teachers to ensure learning without accumulating confirmation bias. The Nested Mean Teaching and Dual Complementary Teaching comprise the final Unsupervised Prompt Tuning framework.

3.1. Preliminaries

Text-Driven Object Detection. We use GLIP [23] as an example in this paper. Unlike typical object detection designed to match between regions and classes, text-driven object detection aligns each region to the corresponding phrase in a text prompt. For the given text prompt and the input image, text-driven object detection feeds both into the visual encoder EncI(·) and the text encoder EncL(·) to extract region features featI and token features featL:

\[ \text{feat}_I = \text{Enc}_I(x), \quad \text{feat}_L = \text{Enc}_L(p), \]  

where \( x \) is the input image and \( p \) is the corresponding text prompt. After that, the region and token features are fed to the fusion module EncF(·, ·) to achieve the results of object detection.
classification $o_{cls}$ and box regression $o_{reg}$:

$$o_{cls}, o_{reg} = Enc_F(feat_I, feat_L),$$  \hspace{1cm} (2)

Finally, the text-driven object detection is trained with the classification and localization losses, respectively:

$$L = L_{cls}(o_{cls}, y_{cls}) + L_{reg}(o_{reg}, y_{reg}),$$  \hspace{1cm} (3)

where $y_{cls}$ and $y_{reg}$ are the classification and localization ground-truth labels. For more details about text-driven object detection, please refer to the original paper [23].

**Prompt Tuning.** Text-driven object detection has shown promising zero-shot generalization to the downstream tasks but heavily relies on laborious prompt engineering. Existing works typically address this problem by tuning text prompts using downstream training data in a supervised manner. In the prompt tuning, all the model parameters are frozen, and only the text prompt is optimized via end-to-end training:

$$p_t+1 = \min_p L(F(x, p_t), y)$$  \hspace{1cm} (4)

where $y = (y_{cls}, y_{reg})$ is the ground-truth label and $t$ denotes the learning iteration. $F(\cdot, \cdot)$ represents the pre-trained text-driven object detection, which contains an image encoder $Enc_I(\cdot)$, a text encoder $Enc_L(\cdot)$, and a fusion module $Enc_F(\cdot, \cdot)$. However, directly updating the text prompt may distort the semantics of the pre-defined prompt, especially in unsupervised training. Instead, we adopt residual prompt tuning (with $\Delta p$) and apply $L2$ regularization to the residual prompt $\Delta p$ to avoid mode diffusion:

$$\Delta p_{t+1} = \min_p L(F(x, p + \Delta p_t), y) + w\|\Delta p_t\|_2$$  \hspace{1cm} (5)

where $w$ is the weight decay. For simplicity, the regularization term is omitted in the following sections.

### 3.2. Nested Mean Teaching

Mean Teaching is an intuitive baseline to solve unsupervised prompt tuning. As shown in Fig. 2a, the student (with a learnable prompt $\Delta S$) is trained on the unlabeled data $x$ with the pseudo boxes $\hat{y}$ predicted by the teacher (with a momentum prompt $\Delta T$) via strong-weak data augmentation. In turn, the teacher is evolved gradually by updating the prompt from the student in an EMA manner:

$$\hat{y} = A_w^{-1}(F(A_w(x), p + \Delta T_t), \tau_I)$$

$$\Delta S_{t+1} = \min_{\Delta S_t} L(F(A_s(x), p + \Delta S_t), A_s(\hat{y}))$$

$$\Delta T_{t+1} = \mu \Delta T_t + (1 - \mu) \Delta S_{t+1}$$  \hspace{1cm} (6)

where $\mu$ is the momentum coefficient. $\tau_I$ is the confidence threshold to filter out the pseudo boxes. $A_s(\cdot)$ and $A_w(\cdot)$ denote the strong-weak data augmentation. $A_w^{-1}(\cdot)$ is the inverse operation of $A_w(\cdot)$ to map the pseudo boxes of $A_w(x)$ to the original images.

Due to the noisy label, the quality of pseudo boxes corresponds to the performance of the teacher model. From this perspective, the mean teaching can be formulated as a “learning to annotate” problem and has been proven effective to optimize pseudo label. Inspired by the success of mean teaching, we improve it from the consideration that the teacher is evolved gradually as the training goes on, a timestamp $t+1$ teacher is more likely performs better than a previous timestamps $t$ teacher. We create the Nested Mean Teaching framework comprising two optimization loops, including an inner and an outer loop. The inner loop uses $k$-step ghosted teacher-student mutual learning to achieve better pseudo boxes. The outer loop uses the pseudo boxes from the inner loop as supervision to train the student prompt and optimize the teacher in an EMA manner. In this design, these two loops can be mutually optimized.

**Nested Loop (Inner Loop).** The inner loop is similar to the $k$-step conventional Mean Teaching process. Firstly, the ghosted teacher $\Delta S_t'$ and student $\Delta T_t'$ prompts in the inner loop are initialized from the counterparts in the outer loop. Then the ghosted teacher creates pseudo boxes to train the student and update its prompt via EMA:

$$\hat{y} = A_w^{-1}(F(A_w(x), p + \Delta T_t'), \tau_I)$$

$$\Delta S_{t+1} = \min_{\Delta S_t} L(F(A_s(x), p + \Delta S_t), A_s(\hat{y}))$$

$$\Delta T_{t+1} = \mu \Delta T_t + (1 - \mu) \Delta S_{t+1}$$  \hspace{1cm} (7)

where $k = (0, ..., K - 1)$. After k-step mutual learning, the ghosted teacher can create better pseudo boxes to drive the optimization in the outer loop.

**Iterative Loop (Outer Loop).** The teacher and student prompts in the outer loop are optimized given the pseudo boxes achieved from the inner loop:

$$\hat{y} = A_w^{-1}(F(A_w(x), p + \Delta T_{t+K}), \tau_I)$$

$$\Delta S_{t+1} = \min_{\Delta S_t} L(F(A_s(x), p + \Delta S_t), A_s(\hat{y}))$$

$$\Delta T_{t+1} = \mu \Delta T_t + (1 - \mu) \Delta S_{t+1}$$  \hspace{1cm} (8)

Note that the ghosted models are discarded after generating pseudo boxes in each step. We claim that Nested Mean Teaching will take effect under the assumption that the teacher is evolved gradually. That is, the pseudo boxes predicted by the teacher are denoised gradually. Given the better pseudo boxes for $\Delta T_t$, the performance of $\Delta T_{t+1}$ in Fig. 2b will be better than $\Delta T_{t+1}$ in Fig. 2a.

### 3.3. Dual Complementary Teaching

**Dual Teachers.** Due to the lack of annotations, there exists a risk of accumulating confirmation bias in unsupervised prompt tuning. To mitigate this problem, we adopt Dual Teacher to retain true-positives while recalling false-negatives. As illustrated in Fig. 3, the proposed Dual Teacher framework consists of three components: an offline teacher (with a frozen pre-defined prompt), an online
teacher (with a momentum prompt), and a student model (with a learnable prompt). The offline teacher uses the pre-defined prompt and a high confidence threshold $\tau_2$ to generate pseudo boxes, which guarantee the knowledge of the basic true-positive objects. This knowledge will not be forgotten as the training goes on. The online teacher explores false-negatives using the momentum prompt with the other confidence threshold $\tau_1$. The pseudo boxes from the online teacher are expected to be complementary to those from the offline teacher. We combine these two parts of pseudo boxes to supervise the training of the student model:

$$\hat{y} = \mathcal{M}(A_w^{-1}(F(A_w(x), p), \tau_2), A_s^{-1}(F(A_s(x), p + \Delta T_1), \tau_1))$$

(9)

where $\mathcal{M}(\cdot)$ denotes Non-Maximum Suppression (NMS).

In each iteration, we update the online teacher prompt and the student prompt as the manner of Eq.6. If we aim to optimize the Dual Teacher using the proposed Nested Mean Teaching mechanism, we can directly use Eq.9 to replace the first equation in Eq.7 (change $\Delta T_i$ in Eq.9 to $\Delta T'_{i+k}$ for Eq.7) and Eq.8 (change $\Delta T_i$ in Eq.9 to $\Delta T'_{i+K}$ for Eq.8).

**Complementary Labeling.** Strong-weak data augmentation is a popular technique in Mean Teaching, which feeds the weakly-augmented images to the teacher to generate pseudo boxes and then uses the strongly-augmented images to optimize the student. However, this kind of technique in Dual Teacher has yet to be studied. We explore the data augmentation technique to promote the collaboration of Dual Teachers for complementary labeling. As shown in Eq.9 and Fig. 3, we arm different teachers with different data augmentation strengths. The offline teacher accounts for initializing sufficient true-positives by feeding weakly-augmented images. The role of the online teacher is designed to recall false-negatives while avoiding introducing false-positives. Hence, contrary to the offline teacher, the strongly-augmented images are fed to the online teachers, which is a strict access condition to introduce new pseudo boxes. In this way, we can avoid introducing false-positives cumulatively.

### 3.4. Method Summary

The proposed unsupervised prompt tuning is summarized in Algorithm 1, which exploits Nested Mean Teaching to assist Dual Complementary Teaching. With carefully design, both methods can be mutually refined.

### 4. Experiments

#### 4.1. Datasets

We validate the effectiveness of the proposed approach on six multi-class object detection tasks (Cityscapes [4], Foggy Cityscapes [37], BDD100K [46], WaterColor [17], Pascal VOC 2012 [7], MS COCO 2017 [28]), and four single-class object detection tasks (KITTI [11], Sim10K [19], EgoHands [23] and Pistols [23]). The detail of the datasets is described in Tab. 1. We can see that there are abundant image styles and object categories to fully val-

---

**Algorithm 1** **Unsupervised Prompt Tuning (UPT)**

**Input:** Pre-trained GLIP $\mathcal{F}$, unlabeled training data $D_s$, pre-defined prompt $p \"[CLS],[CLS],...,[CLS]\"$, two confidence thresholds $\tau_1$ and $\tau_2$ for Dual Teachers, strong-weak data augmentation strategies $A_s$ and $A_w$, EMA rate $\mu$, inner loop steps $K$.

**Output:** Momentum prompt $\Delta T$

1. Initialize $\Delta S = 0$ for the student model.
2. Initialize $\Delta T = 0$ for the online teacher.
3. for each batch $x \in D_s$ do
   4. $\Delta S', \Delta T' = copy(\Delta S), copy(\Delta T)$
   5. // inner loop
   6. for $k$ in $1, ..., K$ do
      7. $y_{off} = A_w^{-1}(F(A_w(x), p), \tau_2)$
      8. $y_{on} = A_s^{-1}(F(A_s(x), p + \Delta T'), \tau_1)$
      9. $y = NMS(y_{off}, y_{on})$
     10. $\Delta S' \leftarrow \min_{\Delta S, L}(F(A_s(x), p + \Delta S'), A_s(y))$
     11. $\Delta T' = \mu \Delta T' + (1 - \mu) \Delta S'$
   12. end for
   13. // outer loop
   14. $y_{off} = A_w^{-1}(F(A_w(x), p), \tau_2)$
   15. $y_{on} = A_s^{-1}(F(A_s(x), p + \Delta T'), \tau_1)$
   16. $y = NMS(y_{off}, y_{on})$
   17. $\Delta S \leftarrow \min_{\Delta S, L}(F(A_s(x), p + \Delta S), A_s(y))$
   18. $\Delta T = \mu \Delta T + (1 - \mu) \Delta S$
   19. end for
Table 1: The statistics of the evaluation datasets, including the pre-defined text prompts, and the image number of the datasets. Since the key metric in these kinds of scenarios is to measure the quality of pseudo labels in the training data, like Test-Time Adaptation or Transductive Learning, the train and test datasets are permitted to share. * means the testing data is the same as the training data.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Pre-defined Text Prompt</th>
<th>Train / Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cityscapes</td>
<td>truck. car. rider. person. train. motorcycle. bicycle. bus.</td>
<td>2065/2085*</td>
</tr>
<tr>
<td>Foggy Cityscapes</td>
<td>truck. car. rider. person. train. motorcycle. bicycle. bus.</td>
<td>8055/1407</td>
</tr>
<tr>
<td>BDD100K</td>
<td>truck. car. rider. person. train. motorcycle. bicycle. bus.</td>
<td>36596/5258</td>
</tr>
<tr>
<td>WaterColor</td>
<td>person. bird. car. cat. bicycle. dog.</td>
<td>2000/2000*</td>
</tr>
<tr>
<td>Pascal VOC</td>
<td>aeroplane. bicycle. ... tmirror</td>
<td>5717/5823</td>
</tr>
<tr>
<td>MS-COCO</td>
<td>person. bicycle. ... toothbrush</td>
<td>118287/5000</td>
</tr>
<tr>
<td>KITTI</td>
<td>car.</td>
<td>6084/6684*</td>
</tr>
<tr>
<td>Sim10K</td>
<td>car.</td>
<td>9975/9975*</td>
</tr>
<tr>
<td>EgoHands</td>
<td>hands.</td>
<td>3840/480</td>
</tr>
<tr>
<td>Pistols</td>
<td>pistol.</td>
<td>2071/2071*</td>
</tr>
</tbody>
</table>

4.2. Implementation Details

Network architecture. GLIP-T [23] is used as the basic vision-language object detection architecture for unsupervised prompt tuning, where the image encoder is based on Dynamic Head [5] with Swin-Tiny [32] as the backbone and BERT [6] as the text encoder. GLIP-T is pre-trained on 1) Objects365 [38], which contains 0.66M images with 365 categories, 2) GoldG, which contains 0.8M human-annotated gold grounding data [20] without COCO images.

Data Augmentation Strategy. We use the same data augmentation strategy in the baseline and the proposed method for a fair comparison. We adopt the practice in [2] to set the strong-weak data augmentation strategies.

Optimization. We perform unsupervised prompt tuning for 10K training iterations with a batch size of 4 and a fixed learning rate of 0.0001 on two GPUs. The residual prompt is trained by an AdamW optimizer [33] with the weight decay $\omega$ of 0.25. The EMA rate $\mu$ is set 0.99 by default. In this paper, we choose the checkpoint of the mean teacher on the 10K-th training iteration to report the performance without cherry-picking. Moreover, all downstream tasks share the same hyper-parameters without specific tuning.

Pseudo Labels Generation. Dual Teaching employs two complementary teachers with different labeling thresholds. We set $\tau_1$ 0.5 for the online teacher, and set $\tau_2$ 0.7, a higher confidence threshold, for the offline teacher. Moreover, we transfer the label of “person” to “rider” if the IoU between pseudo boxes of “person” and “bicycle” exceeds 0.3.

Comparison Baselines. To the best of our knowledge, this work is the first attempt to study unsupervised prompt tuning for text-driven object detection on various downstream tasks. We aim to adapt the pre-trained model to downstream tasks, where the source data is unavailable and the source model cannot be reformulated. Some of the existing unsupervised domain adaptation or semi-supervised learning works [2, 31] are heavily relied on the network architecture, which cannot be directly applied to UPT. Therefore, we reproduce a generalized semi-supervised object de-
Table 7: The effect of Dual Teachers on Cityscapes dataset.
Here “ZS”, “Off”, “On” and “DCT” denote GLIP-T zero-shot inference baseline, offline teacher only, online teacher only, and the proposed Dual Complementary Teaching.

<table>
<thead>
<tr>
<th></th>
<th>tru.</th>
<th>car</th>
<th>rid.</th>
<th>per.</th>
<th>tru.</th>
<th>mot.</th>
<th>bic.</th>
<th>bus</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZS</td>
<td>15.7</td>
<td>55.6</td>
<td>10.6</td>
<td>39.5</td>
<td>19.7</td>
<td>44.4</td>
<td>41.6</td>
<td>43.2</td>
<td>33.8</td>
</tr>
<tr>
<td>Off</td>
<td>19.2</td>
<td>46.2</td>
<td>30.2</td>
<td>38.4</td>
<td>33.4</td>
<td>42.8</td>
<td>39.3</td>
<td>44.6</td>
<td>36.8</td>
</tr>
<tr>
<td>On</td>
<td>15.7</td>
<td>66.9</td>
<td>35.5</td>
<td>42.8</td>
<td>24.0</td>
<td>46.8</td>
<td>44.2</td>
<td>43.7</td>
<td>40.0</td>
</tr>
<tr>
<td>DCT</td>
<td>19.7</td>
<td>70.8</td>
<td>36.4</td>
<td>46.3</td>
<td>26.7</td>
<td>47.9</td>
<td>43.8</td>
<td>47.2</td>
<td>42.4</td>
</tr>
</tbody>
</table>

Table 8: Ablation study on data augmentation in Dual Complementary Teaching on Cityscapes dataset.

<table>
<thead>
<tr>
<th>Strong Augmentation</th>
<th>Online</th>
<th>Offline</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>×</td>
<td>40.5</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>44.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Ablation study on the effect of Dual Complementary Teaching and Nested Mean Teaching.

<table>
<thead>
<tr>
<th></th>
<th>Zero-shot</th>
<th>DCT NMT</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased Teacher</td>
<td>- -</td>
<td>- -</td>
<td>33.8</td>
</tr>
<tr>
<td>UPT (Ours)</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>43.4</td>
</tr>
</tbody>
</table>

4.3. Main Results

We present the results of ten datasets and conduct performance comparisons among GLIP-T zero-shot inference, ensemble zero-shot inference, Unbiased Teacher, and the proposed UPT framework. As shown in Tab. 2-Tab. 6, UPT achieves significant improvements against the zero-shot baseline and outperforms by about 0.6-10.5 mAP on each dataset without using any annotations. Compared with ensemble zero-shot baseline and the reproduced Unbiased Teacher, the proposed UPT is superior to them significantly. There still exists performance gap between UPT and Oracle, the latter of which is supervised by ground-truth labels and is the performance upper bound of UPT, leaving an improvement space for future works. Note that the performances on MS-COCO and Pascal VOC datasets are not as impressive as on other datasets because the performance gap between the zero-shot baseline and the oracle result is limited, which leaves a limited performance improvement space for unsupervised learning. Also, TTA is an inference trick by assembling multi-view results, which is orthogonal to UPT. Using TTA, UPT can further improve the performances on MS-COCO and Pascal VOC to 65.7 and 85.8 mAP, which surpass the zero-shot baseline with TTA.

4.4. Ablation Study

4.4.1 Dual Complementary Teaching (DCT)

The effect of Dual Teachers. To validate the effectiveness of Dual Teachers, we design the experiments using online or offline teachers only. As shown in Tab. 7, compared with online or offline teacher only, the proposed Dual Complementary Teaching can boost the performance of zero-shot performance of GLIP-T by 8.6 mAP without using any annotation. The complementary mechanism takes advantage of distinguishing fine grained categories such as

Figure 4: The quality of pseudo boxes during training. We report the results of recall rate and accuracy of “Dual Teachers” and “Online Teacher only” on Cityscapes dataset.

Figure 5: Ablation study on the confidence threshold $\tau_1$ in Online Teacher and $\tau_2$ in Offline Teacher (Cityscapes). Left: Fix $\tau_1$ as 0.5 and tune $\tau_2$ from 0.6 to 0.8. Right: Fix $\tau_2$ as 0.7 and tune $\tau_1$ from 0.4 to 0.6.

Figure 6: Visualization of pseudo boxes before (Left) and after (Right) nested annotation. Green, red and gold boxes denote true-positives, false-positives and false-negatives.
“truck”, “car” and “bus”, which is challenging for single-teacher to distinguish. Fig. 4 evaluates the quality of pseudo labels in terms of recall rate and accuracy. Compared with online teacher only, Dual Teachers provide more and more reliable pseudo boxes as the training goes on, which is more susceptible to noise labels without forgetting the original confident true-positive boxes during training.

The effect of data-augmentation-based complementary labeling. To demonstrate the effectiveness of the proposed data augmentation strategy in DCT framework, Tab. 8 ablates the data augmentation used by Dual Teachers. We observe that the online teacher with strong augmentation boosts the dual teachers with weak augmentation baseline by 1.9 mAP, indicating that feeding strongly-augmented images to the online teacher can prevent introducing false-positives. In contrast, changing the weak augmentation to strong augmentation of offline teacher decreases the weakly-augmented dual teachers and only strong augmentation online teacher by 0.2 mAP and 0.7 mAP, which claims that feeding weakly-augmented images to the offline teacher can prevent forgetting confident true-positives. In the proposed data-augmentation-based labeling mechanism, Dual Teachers are expected to act better complementary roles for pseudo labeling.

Analysis of hyper-parameters. The confidence threshold to filter out pseudo boxes in Mean Teaching for object detection is an important hyper-parameter. Here we study the effect of $\tau_1$ for online teacher and $\tau_2$ for offline teacher. As shown in Fig. 5, we can achieve the best result when $(\tau_1, \tau_2)$ are set (0.5, 0.7). Here $\tau_2$ is set 0.7 so as to achieve low-noisy true-positives with high confidence. If $\tau_2$ is set to a lower score, there exists a risk that the offline teacher may introduce abundant false-positives, which may harm the optimization of the online teacher.

4.4.2 Nested Mean Teaching (NMT)

The effect of nested annotation. Mean Teaching can be regarded as a process of label denoising. Under the assumption that the pseudo boxes tend to be better as the training goes on, we use inner loop to fetch the better pseudo boxes to supervise the outer loop. From the visualization of pseudo boxes in Fig. 6, Nested Mean Teaching successfully filters out the false positive box in Fig. 6a and excavates more true positive samples from backgrounds in Fig. 6b. As shown in Tab. 9, Nested Mean Teaching can further boost Dual Complementary Teaching to a more competing level.

The effect of $k$-step. To study the effect of $k$-step in the inner loop for Nested Mean Teaching, we ablate $K$ from 1 to 4 and "K=0" represents the normal Dual Complementary Teaching without $k$-step adaptation in the inner loop, and test the performance on Cityscapes dataset. As presented in Tab. 10, $K = 3$ shows a more stable performance improvement. Therefore, we set $K = 3$ as a default setting to verify the method in this paper.

4.5. Prompt Tuning Rather Than Model Tuning

We follow the Prompt Tuning experiments setting except the learning rate (1e-6 instead) and AdamW optimizer weight decay (5e-2 instead) to conduct experiments on model tuning (MT, tuning the entire model parameters) and visual model tuning (MT-V, only tuning the visual encoder). As shown in Tab. 11, both MT and MT-V are easily trapped into negative transfer due to the absence of labeled data. Prompt tuning is more stable and shows better performance in the unsupervised setting (see Tab. 2). Also, we can reuse the model parameters by tuning prompts for different tasks. Both are the important reasons to study prompt tuning over model parameter tuning.

4.6. Comparison with Few-Shot Prompt Tuning

We vary the amount of task-specific annotated data, from zero-shot (inference with the pre-trained model), to $X$-shot (we randomly sample the dataset such that there are at least $X$ examples per category) and using all data in the training dataset. We compare our Unsupervised Prompt Tuning (UPT) method with the zero-shot and few-shot training mAP curve in Fig. 8. UPT outperforms the 5-shot supervised GLIP-T on Cityscapes dataset. Tab. 12 shows the specific result on each category.

4.7. Qualitative Analysis

As shown in Fig. 7, we present the qualitative results of unsupervised prompt tuning on the downstream tasks.
Figure 7: Qualitative results on the downstream tasks. From left to right: WaterColor, Pistols, Cityscapes, Foggy Cityscapes. From top to bottom: zero-shot inference of GLIP-T, Unbiased Teacher, UPT (Nested Mean Teaching + Dual Complementary Teaching). Green, red and gold boxes denote true-positives, false-positives and false-negatives, respectively.

of WaterColor, Pistols, Cityscapes, and Foggy Cityscapes datasets. The visualization shows that the proposed UPT can significantly boost the performance to adapt the downstream tasks, which reduces false-positives and recall false-negatives.

5. Conclusion

In this paper, for the first time, we study a challenging yet meaningful task, unsupervised prompt tuning for text-driven object detection, which can extend the promising out-of-distribution zero-shot inference capacity for the vision-language object detection models. To solve this task, we propose a novel framework composed of Nested Mean Teaching and Dual Complementary Teaching mechanisms, which we hope can inspire future works in this field.

6. Limitations

Nested Mean Teaching will take effect under the assumption that the online teacher is evolved during training. If the online teacher degenerates during training, Nested Mean Teaching may aggravate the degeneration process. Dual Complementary Teaching (so as the conventional Mean Teaching) works when provided not bad zero-shot performance as the initialization for pseudo labeling. Extremely speaking, when the zero-shot accuracy approaches zero, unsupervised prompt tuning is unsolvable.

Figure 8: Data efficiency of few-shot prompt tuning on Cityscapes dataset, X-axis is the amount of task specific data (providing at least X examples per category), Y-axis is the average AP across 8 categories. We also mark the UPT performance in the figure for a clear comparison with the few-shot mAP curve.

Acknowledgements

This work was sponsored by National Key R&D Program of China (2023YFE0204200), National Natural Science Foundation of China (No.62127803), Key R&D Project of Zhejiang Province (No.2022C01056) and the National Natural Science Foundation of Zhejiang Province (NO.LQ21F030017).
References


[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML. 2021.


[27] Luoqun Lin, Zhifeng Yang, Qipeng Liu, Yuanlong Yu, and Qifeng Lin. Run and chase: Towards source-free domain adaptive object detection. In ICME, 2023. 3
[33] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. ArXiv, abs/1711.05101, 2017. 6
[38] 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. 6