Unsupervised Prompt Tuning for Text-Driven Object Detection

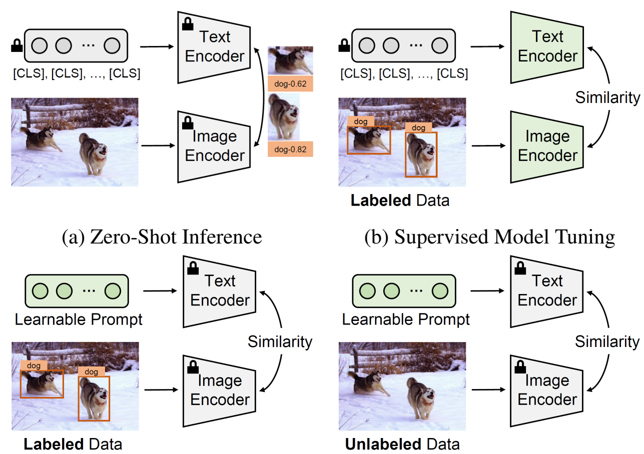
Weizhen He[1*†], Weijie Chen[1,2,3†‡], Binbin Chen[2], Shicai Yang[2],
Di Xie[2,3‡], Luojun Lin[4], Donglian Qi[1], Yueting Zhuang[1‡]
[1]Zhejiang University, [2]Hikvision Research Institute,
[3]Key Laboratory of Peace-building Big Data of Zhejiang Province, [4]Fuzhou University
{hewz, chenweijie, qidl, 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-



(a) Zero-Shot Inference (b) Supervised Model Tuning

(c) Supervised Prompt Tuning (d) Unsupervised Prompt Tuning

Figure 1: Illustration of GLIP and three optimization manners to adapt downstream object detection tasks. (a) Zero-shot inference, which locates the target objects with the pre-defined prompts using category name. (b) and (c) are two supervised tuning manners using the labeled downstream data. (d) is the proposed unsupervised prompt tuning method, which exploits the zero-shot results as the prior supervision cues of the unlabeled downstream data.

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 task-specific 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

*Internship in Hikvision Research Institute.
†Equal contributions
‡Corresponding authors

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 overfit on 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.

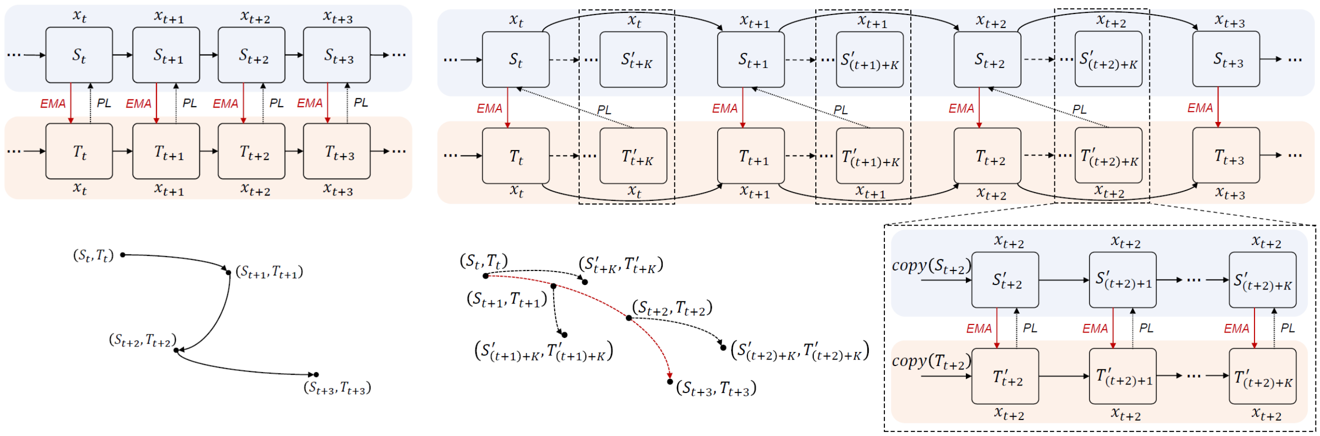
These two mean teaching mechanisms are orthogonal to each other. We build the Unsupervised Prompt Tuning framework (UPT) by integrating them, where the Dual Complementary Teaching process is optimized in a nested annotation manner. Extensive ablation studies and experiments are carried out on multiple 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], EgoHands [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,



(a) Mean Teaching

(b) Nested Mean Teaching

Figure 2: Illustration of Mean Teaching and Nested Mean Teaching (*PL* is short for *pseudo labeling*). (a) Conventional Mean Teaching uses the teacher to generate pseudo boxes to supervise the student in an EMA manner. (b) In contrast, the proposed Nested Mean Teaching consists of a nested inner loop and an outer loop. The inner loop uses k-step ghosted teacher-student mutual learning to achieve better pseudo boxes. The outer loop uses these better pseudo boxes to drive the optimization.

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 $Enc_I(\cdot)$ and the text encoder $Enc_L(\cdot)$ to extract region features $feat_I$ and token features $feat_L$:

$$feat_I = Enc_I(x),\ feat_L = Enc_L(p), \tag{1}$$

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 $Enc_F(\cdot,\cdot)$ to achieve the results of object classification o_{cls} and box regression o_{reg}:

$$o_{cls}, o_{reg} = Enc_F(feat_I, feat_L), \tag{2}$$

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

$$\mathcal{L} = \mathcal{L}_{cls}(o_{cls}, y_{cls}) + \mathcal{L}_{reg}(o_{reg}, y_{reg}), \tag{3}$$

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:

$$p_{t+1} \leftarrow \min{}_{p_t} \ \mathcal{L}(\mathcal{F}(x, p_t), y) \tag{4}$$

where $y = (y_{cls}, y_{reg})$ is the ground-truth label and t denotes the learning iteration. $\mathcal{F}(\cdot,\cdot)$ 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 ($p + \Delta p$) and apply $\mathcal{L}2$ regularization to the residual prompt Δp to avoid mode diffusion:

$$\Delta p_{t+1} \leftarrow \min{}_{\Delta p_t} \ \mathcal{L}(\mathcal{F}(x, p + \Delta p_t), y) + w\|\Delta p_t\|_2 \tag{5}$$

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 ΔS) is trained on the unlabeled data x with the pseudo boxes $\hat{y}$ predicted by the teacher (with a momentum prompt ΔT) via strong-weak data augmentation. In turn, the teacher is evolved gradually by updating the prompt from the student in an EMA manner:

$$\begin{aligned}
&\hat{y} = \mathcal{A}_w^{-1}(\mathcal{F}(\mathcal{A}_w(x), p + \Delta T_t), \tau_1) \\
&\Delta S_{t+1} \leftarrow \min{}_{\Delta S_t} \ \mathcal{L}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta S_t), \mathcal{A}_s(\hat{y})) \\
&\Delta T_{t+1} = \mu \Delta T_t + (1-\mu)\Delta S_{t+1}
\end{aligned} \tag{6}$$

where μ is the momentum coefficient. τ_1 is the confidence threshold to filter out the pseudo boxes. $\mathcal{A}_s(\cdot)$ and $\mathcal{A}_w(\cdot)$ denote the strong-weak data augmentation. And $\mathcal{A}_w^{-1}(\cdot)$ is the inverse operation of $\mathcal{A}_w(\cdot)$ to map the pseudo boxes of $\mathcal{A}_w(x)$ 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:

$$\begin{aligned}
&\hat{y} = \mathcal{A}_w^{-1}(\mathcal{F}(\mathcal{A}_w(x), p + \Delta T'_{t+k}), \tau_1) \\
&\Delta S'_{t+k+1} \leftarrow \min{}_{\Delta S'_{t+k}} \mathcal{L}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta S'_{t+k}), \mathcal{A}_s(\hat{y})) \\
&\Delta T'_{t+k+1} = \mu \Delta T'_{t+k} + (1-\mu)\Delta S'_{t+k+1}
\end{aligned} \tag{7}$$

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:

$$\begin{aligned}
&\hat{y} = \mathcal{A}_w^{-1}(\mathcal{F}(\mathcal{A}_w(x), p + \Delta T'_{t+K}), \tau_1) \\
&\Delta S_{t+1} \leftarrow \min{}_{\Delta S_t} \ \mathcal{L}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta S_t), \mathcal{A}_s(\hat{y})) \\
&\Delta T_{t+1} = \mu \Delta T_t + (1-\mu)\Delta S_{t+1}
\end{aligned} \tag{8}$$

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 ΔT_t, the performance of ΔT_{t+1} in Fig. 2b will be better than ΔT_{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

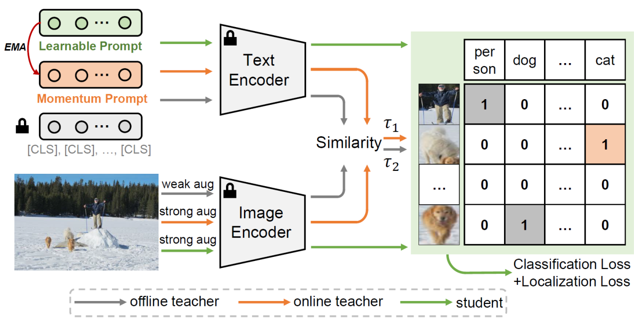


Figure 3: The overview of Dual Complementary Teaching: Firstly, the offline teacher (with a frozen pre-defined prompt composed by the category name) and the online teacher (with a momentum prompt) generate pseudo boxes using the weakly-augmented and the strongly-augmented images, respectively. Then, the pseudo boxes are merged and fed to supervise the student (with a learnable prompt) using the strongly-augmented images.

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 τ_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 τ_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:

$$\hat{y} = \mathcal{M}(\mathcal{A}_w^{-1}(\mathcal{F}(\mathcal{A}_w(x), p), \tau_2),\ \mathcal{A}_s^{-1}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta T_t), \tau_1)) \tag{9}$$

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 ΔT_t in Eq.9 to $\Delta T'_{t+k}$ for Eq.7) and Eq.8 (change ΔT_t in Eq.9 to $\Delta T'_{t+K}$ 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-

---

Algorithm 1 Unsupervised Prompt Tuning (UPT)

Input: Pre-trained GLIP $\mathcal{F}$, unlabeled training data $\mathcal{D}_x$, pre-defined prompt p "[CLS],[CLS],...,[CLS]", two confidence thresholds τ_1 and τ_2 for Dual Teachers, strong-weak data augmentation strategies $\mathcal{A}_s$ and $\mathcal{A}_w$, EMA rate μ, inner loop steps K

Output: Momentum prompt ΔT

Initialize $\Delta S = \vec{0}$ for the student model
Initialize $\Delta T = \vec{0}$ for the online teacher
for each batch x in $\mathcal{D}_x$ **do**
  $\Delta S', \Delta T' = copy(\Delta S), copy(\Delta T)$
  // inner loop
  for k in $1, ..., K$ **do**
    $y_{off} = \mathcal{A}_w^{-1}(\mathcal{F}(\mathcal{A}_w(x), p), \tau_2)$
    $y_{on} = \mathcal{A}_s^{-1}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta T'), \tau_1)$
    $y = NMS(y_{off}, y_{on})$
    $\Delta S' \leftarrow \min_{\Delta S'} \mathcal{L}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta S'), \mathcal{A}_s(y))$
    $\Delta T' = \mu \Delta T' + (1 - \mu)\Delta S'$
  end for
  // outer loop
  $y_{off} = \mathcal{A}_w^{-1}(\mathcal{F}(\mathcal{A}_w(x), p), \tau_2)$
  $y_{on} = \mathcal{A}_s^{-1}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta T'), \tau_1)$
  $y = NMS(y_{off}, y_{on})$
  $\Delta S \leftarrow \min_{\Delta S} \mathcal{L}(\mathcal{F}(\mathcal{A}_s(x), p + \Delta S), \mathcal{A}_s(y))$
  $\Delta T = \mu \Delta T + (1 - \mu)\Delta S$
end for

---

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 vali-

Datasets	Pre-defined Text Prompt	Train / Test
Cityscapes	truck. car. rider. person. train. motorcycle. bicycle. bus.	2685/2685⋆
Foggy Cityscapes	truck. car. rider. person. train. motorcycle. bicycle. bus.	8055/1407
BDD100K	truck. car. rider. person. motorcycle. bicycle. bus.	36596/5258
WaterColor	person. bird. car. cat. bicycle. dog.	2000/2000⋆
Pascal VOC	aeroplane. bicycle. ··· . tvmonitor	5717/5823
MS-COCO	person. bicycle. ··· . toothbrush.	118287/5000
KITTI	car.	6684/6684⋆
Sim10K	car.	9975/9975⋆
EgoHands	hands.	3840/480
Pistols	pistol.	2971/2971⋆

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.

date the proposed method. We evaluate the performance using the mean average precision (mAP) with an intersection-over-union (IoU) threshold of 0.5.

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 w of 0.25. The EMA rate μ 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.

Methods	Cityscapes								
	tru.	*car*	*rid.*	*per.*	*tra.*	*mot.*	*bic.*	*bus*	**Avg.**
GLIP-T [23]	15.7	55.6	10.6	39.3	19.7	44.4	41.6	43.2	33.8
GLIP-T [23] + TTA	15.2	64.0	11.9	**48.0**	22.4	47.5	**48.0**	46.4	37.9
Unbiased Teacher [30]	15.7	66.9	35.5	42.8	24.0	46.8	44.2	43.7	40.0
UPT (ours)	**20.8**	**71.1**	**36.5**	46.7	**30.3**	**47.9**	43.7	**47.7**	**43.1**
Oracle	38.2	75.8	45.1	58.3	45.1	51.2	48.7	53.1	51.9

Table 2: Performance comparison on Cityscapes.

Methods	Foggy Cityscapes								
	tru.	*car*	*rid.*	*per.*	*tra.*	*mot.*	*bic.*	*bus*	**Avg.**
GLIP-T [23]	23.1	51.8	14.9	35.6	4.9	30.5	40.9	50.3	31.5
GLIP-T [23]+TTA	23.2	58.1	15.7	41.2	4.9	33.4	**46.4**	50.5	34.2
Unbiased Teacher [30]	20.0	57.0	32.7	33.5	**12.0**	36.1	42.6	50.6	35.6
UPT (ours)	**26.3**	**61.7**	**36.5**	**41.3**	7.2	**37.3**	44.0	**53.1**	**38.4**
Oracle	36.5	67.8	44.4	48.8	34.7	39.4	50.2	57.4	47.4

Table 3: Performance comparison on Foggy Cityscapes.

Methods	BDD100K							
	tru.	*car*	*rid.*	*per.*	*mot.*	*bic.*	*bus*	**Avg.**
GLIP-T [23]	35.3	50.0	4.1	42.7	31.9	34.5	42.9	34.5
GLIP-T [23]+TTA	36.1	56.3	4.1	46.3	35.3	37.3	46.7	37.4
Unbiased Teacher [30]	35.5	60.6	7.9	47.0	38.6	41.7	39.7	38.7
UPT (ours)	**46.6**	**67.0**	**10.4**	**49.6**	**42.4**	**42.8**	**48.4**	**43.9**
Oracle	50.9	76.5	36.6	64.2	42.4	48.8	53.7	53.3

Table 4: Performance comparison on BDD100K.

Methods	WaterColor						
	per.	*bir.*	*car*	*cat*	*bic.*	*dog*	**Avg.**
GLIP-T [23]	71.5	49.2	60.6	19.3	82.2	24.9	51.3
GLIP-T [23]+TTA	71.1	53.4	50.6	**26.4**	**92.6**	17.4	51.9
Unbiased Teacher [30]	78.2	52.0	61.5	15.6	87.0	19.2	52.2
UPT(ours)	**79.1**	**54.0**	**62.4**	17.8	86.7	**26.7**	**54.5**
Oracle	82.6	62.8	67.2	29.9	91.6	32.5	61.1

Table 5: Performance comparison on WaterColor.

Methods	COC.	VOC.	KIT.	SIM.	Ego.	Pis.	Avg.
GLIP-T [23]	63.5	82.7	61.2	66.3	65.5	71.5	68.5
GLIP-T [23]+TTA	**65.0**	**85.2**	64.2	**69.9**	64.8	74.9	70.7
Unbiased Teacher [30]	63.8	83.1	57.2	62.6	**74.9**	63.9	67.6
UPT(ours)	64.1	83.8	**67.6**	68.9	69.5	**82.0**	**72.7**
Oracle	65.9	86.8	87.4	73.7	94.7	88.0	82.8

Table 6: Performance comparison on MS-COCO, Pascal VOC, KITTI, SIM10K, EgoHands, and Pistols.

We set τ_1 0.5 for the online teacher, and set τ_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 reformed. 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-

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

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.

	Online	Offline	mAP
Strong Augmentation	×	×	40.5
	×	✓	40.3
	✓	×	**42.4**
	✓	✓	<u>41.7</u>

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

	DCT	NMT	mAP
Zero-shot	-	-	33.8
Unbiased Teacher	-	-	40.0
UPT (Ours)	✓	-	<u>42.4</u>
	✓	✓	**43.1**

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

tection method, Unbiased Teacher [30], in the setting of unsupervised prompt tuning for performance comparison. In addition, GLIP-T with pre-trained model is served as a zero-shot performance baseline, and we also introduce an ensemble zero-shot performance via test-time augmentation (TTA) as a strong baseline for performance comparison, which can further reflect the superiority of the proposed method. Moreover, we use the ground-truth labels to conduct prompt tuning in a supervised manner ("Oracle"), which serves as the performance upper bound.

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**$_{50}$ 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

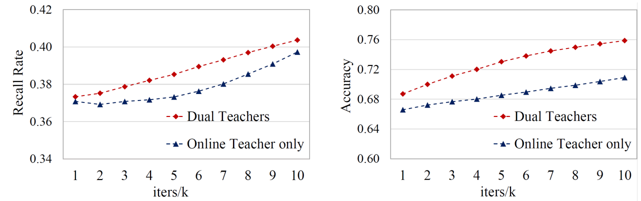


Figure 4: The quality of pseudo boxes during training. We report the results of recall rate and accuracy of "Dual Teachers" and "Online Teacher only" on Cityscapes dataset.

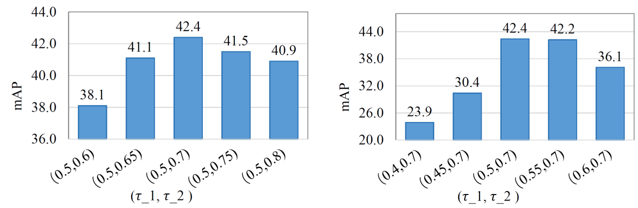
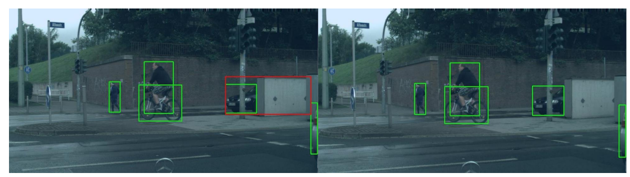
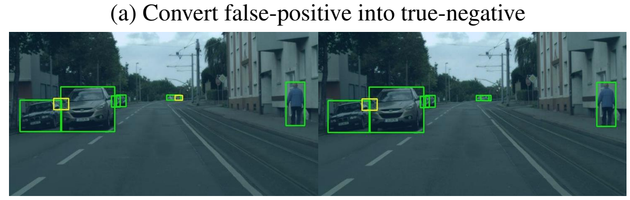


Figure 5: Ablation study on the confidence threshold τ_1 in Online Teacher and τ_2 in Offline Teacher (Cityscapes). *Left*: Fix τ_1 as 0.5 and tune τ_2 from 0.6 to 0.8. *Right*: Fix τ_2 as 0.7 and tune τ_1 from 0.4 to 0.6.



(a) Convert false-positive into true-negative



(b) Recall false-negative into true-positive

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**.

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

K	0	1	2	3	4
mAP	42.4	42.8	43.0	**43.1**	42.7

Table 10: Ablation study on k-step in the inner loop of the Nested Mean Teaching on Cityscapes dataset.

Method	GLIP	MT	MT-V
mAP	**33.8**	12.8	26.5

Table 11: Performance of model tuning (MT) and visual model tuning (MT-V) on Cityscapes dataset.

"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 τ_1 for online teacher and τ_2 for offline teacher. As shown in Fig. 5, we can achieve the best result when (τ_1, τ_2) are set (0.5, 0.7). Here τ_2 is set 0.7 so as to achieve low-noisy true-positives with high confidence. If τ_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.

Shot	**Cityscapes**								
	tru.	*car*	*rid.*	*per.*	*tra.*	*mot.*	*bic.*	*bus*	**Avg.**
1	19.9	53.0	19.4	40.6	29.0	45.8	39.3	48.6	37.0
2	21.9	62.5	22.2	36.9	23.8	45.6	43.6	46.9	37.9
5	27.4	70.0	26.3	50.7	30.1	43.3	45.4	46.9	42.5
10	25.9	73.7	29.0	54.5	42.3	46.6	46.6	44.9	45.4
20	25.3	73.7	35.7	54.7	42.3	48.7	46.9	43.1	46.3
all	38.2	75.8	45.1	58.3	45.1	51.2	48.7	53.1	51.9
UPT	20.8	71.1	36.5	46.7	30.3	47.9	43.7	47.7	43.1

Table 12: Performance comparison between UPT and few-shot prompt tuning on Cityscapes dataset.

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.

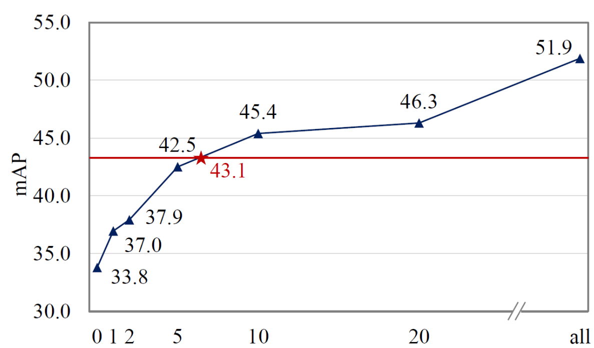


Figure 8: Data efficiency of few-shot prompt tuning on Cityscapes dataset, X-axis is the amount of task specific data (providing at least X examples per category), Y-axis is the average AP across 8 categories. We also mark the UPT performance in the figure for a clear comparison with the few-shot mAP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

[1] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14361–14370, 2022. 3

[2] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, 2022. 6

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 3

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 2, 5

[5] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 6

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 2, 5

[8] Qiang feng Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4079–4088, 2021. 3

[9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. Clip-adapter: Better vision-language models with feature adapters. *ArXiv*, abs/2110.04544, 2021. 3

[10] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *ArXiv*, abs/2001.01526, 2020. 3

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5

[12] Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020. 3

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3

[14] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6910–6920, 2021. 3

[15] Junchu Huang, Weijie Chen, Shicai Yang, Di Xie, Shiliang Pu, and Yueting Zhuang. Transductive clip with class-conditional contrastive learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3858–3862. IEEE, 2022. 2

[16] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models, 2022. 3

[17] Naoto Inoue, Ryosuke Furuta, T. Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5009, 2018. 2, 5

[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. 2

[19] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753, 2017. 2, 5

[20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021. 3, 6

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1

[22] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry Davis. Rethinking pseudo labels for semi-supervised object detection. In *AAAI*, 2022. 3

[23] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022. 1, 2, 3, 4, 5, 6

[24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3

[25] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1

[26] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2023. 3

[27] Luojun Lin, Zhifeng Yang, Qipeng Liu, Yuanlong Yu, and Qifeng Lin. Run and chase: Towards accurate source-free domain adaptive object detection. In *ICME*, 2023. 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5

[29] Qipeng Liu, Luojun Lin, Zhifeng Shen, and Zhifeng Yang. Periodically exchange teacher-student for source-free object detection. In *International Conference on Computer Vision*, 2023. 3

[30] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Péter Vajda. Unbiased teacher for semi-supervised object detection. *ArXiv*, abs/2102.09480, 2021. 2, 3, 6, 7

[31] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 6

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

[33] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 6

[34] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14054–14063, 2022. 3

[35] Longtian Qiu, Renrui Zhang, Ziyu Guo, Ziyao Zeng, Yafeng Li, and Guangnan Zhang. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv e-prints*, pages arXiv–2112, 2021. 1

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 2, 5

[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

[39] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 1

[40] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *ArXiv*, abs/2001.07685, 2020. 3

[41] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 3

[42] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv e-prints*, pages arXiv–2201, 2022. 1

[43] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[44] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3040–3049, 2021. 3

[45] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[46] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 2, 5

[47] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022. 1

[48] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Jiao Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *ArXiv*, abs/2111.03930, 2021. 3

[49] Wei Zhao, Binbin Chen, Weijie Chen, Shicai Yang, Di Xie, Shiliang Pu, and Yueting Zhuang. 1st place solution for eccv 2022 ood-cv challenge object detection track, 2023. 3

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022. 1, 3

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130:2337–2348, 2022. 1, 3