SFUOD: Source-Free Unknown Object Detection

Keon-Hee Park
Kyung Hee University
Yongin, Republic of Korea
pgh2874@khu.ac.kr

Seun-An Choe
Kyung Hee University
Yongin, Republic of Korea
dragoon0905@khu.ac.kr

Gyeong-Moon Park*
Korea University
Seoul, Republic of Korea
gm-park@korea.ac.kr

Abstract

*Source-free object detection adapts a detector pre-trained on a source domain to an unlabeled target domain without requiring access to labeled source data. While this setting is practical as it eliminates the need for the source dataset during domain adaptation, it operates under the restrictive assumption that only pre-defined objects from the source domain exist in the target domain. This closed-set setting prevents the detector from detecting undefined objects. To ease this assumption, we propose **S**ource-**F**ree **U**nknown **O**bject **D**etection (**SFUOD**), a novel scenario which enables the detector to not only recognize known objects but also detect undefined objects as unknown objects. To this end, we propose **CollaPAUL** (**Colla**borative tuning and **P**rincipal **A**xis-based **U**nknown **L**abeling), a novel framework for SFUOD. Collaborative tuning enhances knowledge adaptation by integrating target-dependent knowledge from the auxiliary encoder with source-dependent knowledge from the pre-trained detector through a cross-domain attention mechanism. Additionally, principal axes-based unknown labeling assigns pseudo-labels to unknown objects by estimating objectness via principal axes projection and confidence scores from model predictions. The proposed CollaPAUL achieves state-of-the-art performances on SFUOD benchmarks, and extensive experiments validate its effectiveness. Our code is available at SFUOD.*

1. Introduction

Domain adaptive object detection [1, 5, 6, 12, 16, 29] aims to transfer knowledge from a source to a target domain and is widely studied for its practical applications. Especially, source-free object detection (SFOD) [3, 10, 14, 15, 18, 23, 30] focuses on adapting source-trained models to the unlabeled target domain without requiring access to labeled source data. SFOD is a realistic scenario that addresses data privacy and storage constraints. However, SFOD assumes a closed-set scenario, where the source and target domains share the same set of class labels, struggling to detect objects not defined in the source domain. In real-world applications (e.g., self-driving), detectors must recognize unknown objects to handle unforeseen events (*e.g.*, identifying pedestrians or animals on the road to prevent accidents), but SFOD limits the detector from detecting unknown objects.

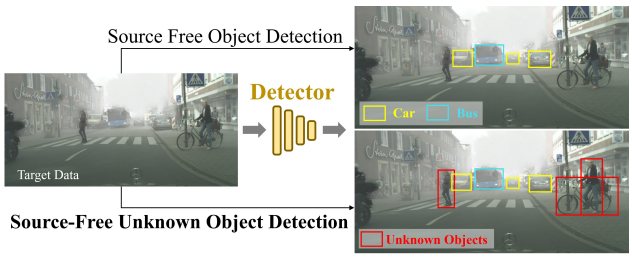
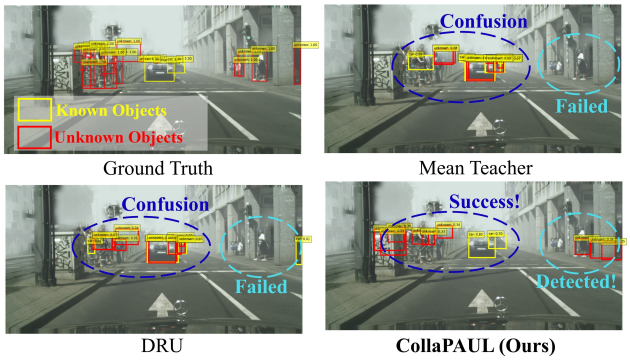


Figure 1. Description of the proposed SFUOD. Yellow and sky-blue boxes denote known objects (*e.g.*, "Car", "Bus"). Red boxes denote unknown objects (*e.g.*, "Person", "Bicycle").

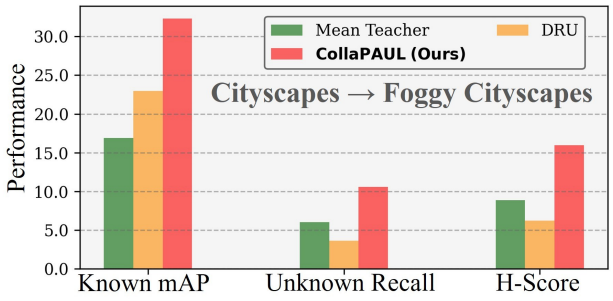
To address the limitation of detecting unknown objects in source-free object detection, we propose a novel Source-Free Unknown Object Detection (SFUOD) scenario. The SFUOD scenario requires not only adapting a source-trained detector to the target domain for known objects but also detecting undefined objects as unknown. As shown in Figure 1, traditional source-free object detection adapts a detector to recognize only pre-defined objects from the source domain (*e.g.*, "Car", "Bus"). In contrast, the detector can also detect unknown objects (*e.g.*, "Person", "Bicycle") in our new scenario. Since unknown objects frequently appear in the real world, the proposed SFUOD is a realistic and challenging scenario as it simultaneously requires knowledge transfer for known objects and detecting unknown objects.

In the SFUOD scenario, existing SFOD methods struggle with both adapting to known objects and detecting unknown ones. Existing approaches like DRU [14] use the Mean Teacher (MT) framework [28], where a teacher model generates pseudo-labels for a student model, and the teacher is updated via exponential moving average (EMA) [28] to mitigate domain shift by transferring source-dependent knowledge into the target domain. However, since the teacher model trained in the source domain lacks knowledge

*Corresponding author.



(a) Qualitative comparison in the source-free unknown object detection.



(b) Quantitative comparison of Known mAP, U-Recall, and H-Score.

Figure 2. Figure 2a presents a qualitative comparison of our method with existing SFOD approaches in Cityscapes → Foggy Cityscapes. Prior methods often misclassify known and unknown objects (blue-colored section) and fail to detect unknown objects (sky blue-colored section). Yellow and red boxes indicate known and unknown objects, respectively. Figure 2b provides a quantitative comparison of our method with prior SFOD methods.

of unknown objects, it cannot successfully generate pseudo-labels for unknown objects. As a result, it struggles to learn unknown objects effectively and fails to detect unknown objects. Additionally, applying source-dependent knowledge to unknown objects leads to knowledge confusion between known and unknown classes, further exacerbating domain shift and hindering adaptation to known objects.

To validate this, we tested SFOD methods (*e.g.*, MT [28], DRU [14]) to the proposed SFUOD scenario. The detector was trained on Cityscapes [4] as the source domain and adapted to the unlabeled Foggy Cityscapes [26] as the target domain. Following the SFUOD setting, the model was trained to classify only vehicles (e.g., Car, Truck, and Bus) using labeled source data, while the target domain remained unlabeled. To assign pseudo-labels to unknown objects, we employed confidence-based pseudo-labeling, a common approach in unsupervised domain adaptation [2, 7, 25]. As shown in Figure 2a, existing SFOD methods misclassify known objects as unknown and vice versa due to knowledge confusion, often failing to detect unknown objects due to ineffective pseudo-labeling. Furthermore, Figure 2b illustrates that knowledge confusion and inaccurate unknown pseudo-labels lead to low mAP for known objects and low detection for unknown objects, respectively. To sum up, SFUOD has two main challenges: 1) *knowledge confusion* due to source-dependent knowledge hinders effective knowledge transfer, and 2) *ineffective unknown pseudo-labeling* fails to detect unknown objects.

To this end, we propose **CollaPAUL** (**Colla**borative tuning and **P**rincipal **A**xis-based **U**nknown **L**abeling) for the new SFUOD scenario, employing the mean teacher approach to effectively adapt knowledge of known objects while learning unknown objects. CollaPAUL comprises two key components. First, collaborative tuning integrates source-dependent knowledge from the student model with target-dependent knowledge from an auxiliary target encoder. To extract target-dependent knowledge, the target encoder applies truncated reconstruction via Singular Value Decomposition (SVD), and cross-domain attention is then employed to fuse source and target features, allowing the student model to learn richer representations through collaborative tuning. Second, Principal Axis-based Unknown Labeling (PAUL) estimates the principal axes of known objects to compute objectness scores, enabling more accurate pseudo-label assignment for unknown objects. Since known and unknown object embeddings share inherent properties (*e.g.*, objectness) absent in non-object proposals, we assume that unknown proposals projected onto the principal axes of known proposals exhibit similarity, whereas non-object proposals remain dissimilar. Based on this assumption, PAUL leverages these principal axes to identify object proposals and assign pseudo-labels to unknown objects. Through the experiments shown in Figure 2, the proposed CollaPAUL can mitigate misclassification by knowledge confusion and detect unknown objects from the background, enhancing both known mAP and unknown recall performances (*i.e.*, achieving the best H-Score).

Our contributions can be summarized as follows:

- We propose **SFUOD**, a new scenario where a source-trained detector adapts its knowledge to a target domain for known objects while simultaneously detecting unknown objects in the target domain. To this end, we introduce **CollaPAUL**, a novel framework combining collaborative tuning and principal axis-based unknown labeling.
- We propose *collaborative tuning* to alleviate knowledge confusion by source-dependent knowledge. Through cross-domain attention, it integrates both source- and target-dependent knowledge, mitigating source reliance by collaboratively tuning target-dependent features.
- We propose a *principal axis-based unknown labeling* method to assign pseudo-labels to unknown objects. It achieves this by projecting embeddings onto the principal axes of known objects to estimate objectness and computing confidence scores for unknown classes.
- We develop two SFUOD benchmarks, where the proposed CollaPAUL achieves state-of-the-art. We further validate its effectiveness through extensive experiments.

2. Related Work

2.1. Source-Free Object Detection

Source-Free Object Detection (SFOD), introduced in SED [18], addresses domain adaptation with data privacy constraints, where the model adapts to the target domain without access to source data. Recent SFOD methods rely on the mean teacher framework for its robustness and stability, but noisy pseudo-labels from the teacher model can lead to the model collapse representing inaccurate student learning and improper teacher updates. To address this, PET [23] introduces a multi-teacher framework with static and dynamic teachers, stabilizing training by periodically exchanging the teacher and student models. SF-UT [10] uses fixed pseudo-labels from the initial teacher model with AdaBN [20] to prevent improper teacher updates. DRU [14] dynamically retrains the student model and updates the teacher accordingly. While these methods focus on stabilizing teacher updates for source-free domain adaptation, we propose collaborative tuning, which improves student model learning by integrating source- and target-dependent representation knowledge for more effective adaptation.

2.2. Open Recognition in Object Detection

Open recognition has gained significant attention in object detection due to its importance in real-world applications. Unknown object detection [9, 21] trains a detector to classify annotated known objects while identifying unannotated novel objects as unknown. Open-world object detection (OWOD) [8, 13, 27, 33] extends this by incorporating continual learning, requiring the detector to sequentially learn annotated known objects and detect unknowns. In domain adaptive object detection, SOMA [17] introduces Adaptive Open-set Object Detection (AOOD), which adapts a source-trained detector to a target domain, enabling it to classify known objects from the source while detecting unknowns. However, AOOD relies on source data for adaptation, raising privacy concerns. To address this, we propose Source-Free Unknown Object Detection (SFUOD), which adapts a source-trained model to the target domain without requiring source data. SFUOD enables the model to recognize known objects defined in the source domain while detecting novel objects as unknowns in the target domain.

3. Method

We propose CollaPAUL, which incorporates collaborative tuning to mitigate knowledge confusion during knowledge transfer and principal axes-based unknown labeling to handle unknown objects. In Sec. 3.1, we first define the source-free unknown object detection problem. Sec. 3.2 describes the base model pipeline, followed by Sec. 3.3, where we introduce collaborative tuning that fuses source- and target-dependent knowledge through the cross-domain attention, improving adaptation to the target domain. Finally, in Sec. 3.4, we present principal axes-based unknown labeling, assigning pseudo-labels to unknown objects by leveraging the principal axes of known objects to estimate the objectness. Figure 3 shows an overview of the CollaPAUL.

3.1. Problem Definition

In this section, we define the problem setting for the proposed SFUOD. Let $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{|D_s|}$ denote the labeled source dataset and $D_t = \{(x_t^i)\}_{i=1}^{|D_t|}$ the unlabeled target dataset. Each source label $y_s^i = \{(b^j, c^j)\}_{j=1}^{J_i}$ contains the number of J_i annotations for known objects, where b^j represents the bounding box and $c^j \in \mathcal{Y}_s$ denotes the object category for the j-th annotation. Unlike SFOD, which assumes a shared class space between source and target domains, SFUOD defines a set of known classes $\mathcal{Y}_s = \{1, 2, \ldots, K\}$ pre-defined in the source dataset and novel classes $\mathcal{Y}_t = \{K+2, K+3, \ldots, K+K'\}$ that are undefined in the source dataset. In the source domain, all labeled objects belong to the base classes $c^j \in \mathcal{Y}_s$, while the target domain may contain both base and novel classes. SFUOD aims to adapt the detector pre-trained on D_s to the target domain using only D_t, enabling it to detect base classes while identifying novel classes as a single 'unknown' category, represented as the class $K+1$ in the target domain.

3.2. Pipeline of the Mean Teacher-based Approach

We utilize the Mean Teacher (MT) framework [28], which consists of a teacher model and a student model sharing the same architecture and initialization. The student model is trained using pseudo-labels generated by the teacher model, while the teacher model is updated via EMA from the student model. During training, the student and teacher models receive strongly and weakly augmented images, respectively. The updates for both models are as follows:

$$\mathcal{L}_{det} = \mathcal{L}_{cls}(\hat{c}_s, \hat{c}_t) + \mathcal{L}_{\mathrm{L}_1}(\hat{b}_s, \hat{b}_t) + \mathcal{L}_{giou}(\hat{b}_s, \hat{b}_t), \quad (1)$$

$$\bar{\theta}_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \quad (2)$$

where $\mathcal{L}_{cls}$ denote the classification loss [22], $\mathcal{L}_{\mathrm{L}_1}$ and $\mathcal{L}_{giou}$ represent the regression losses [24], and α denotes a hyperparameter for weight update. The predicted class and bounding boxes from the student model θ_s and teacher model θ_t are denoted as $\hat{c}_s, \hat{c}_t$ and $\hat{b}_s, \hat{b}_t$, respectively. The teacher model $\bar{\theta}_t$ is updated using EMA and the student model θ_s is trained via $\mathcal{L}_{det}$. The MT framework ensures consistent predictions during training, facilitating knowledge transfer to the target domain.

Most SFOD approaches leverage the MT framework to enhance adaptation while addressing improper model updates caused by noisy pseudo-labels. However, in the proposed SFUOD, the MT framework faces knowledge confusion, as both the student and teacher models, initialized

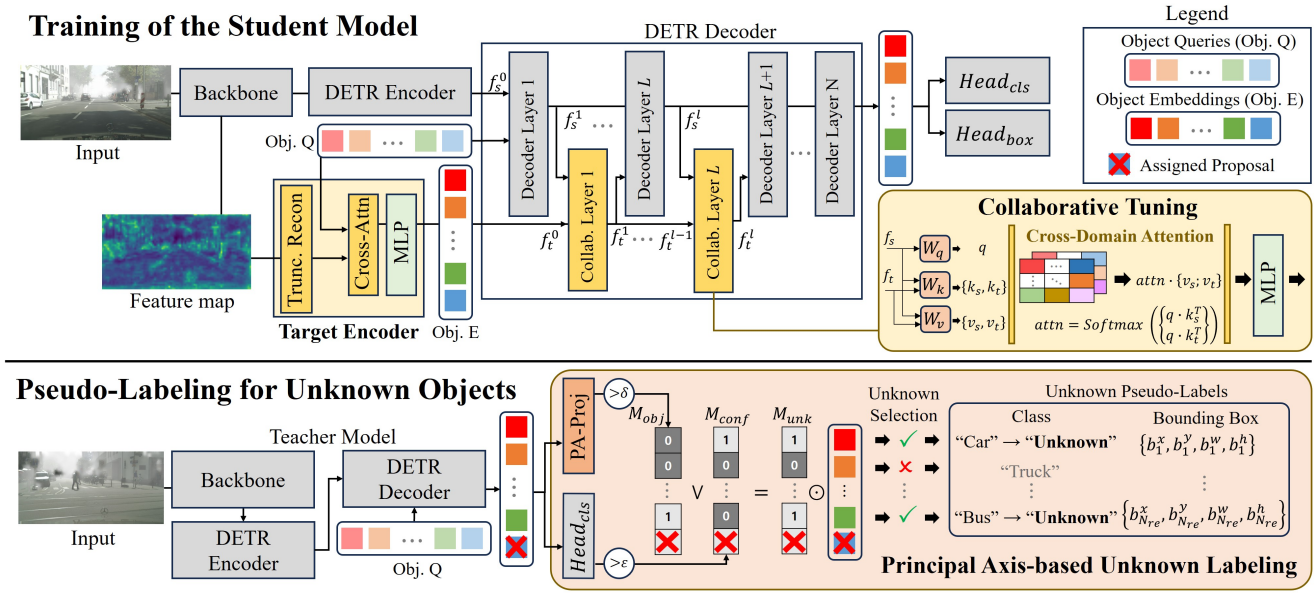


Figure 3. Overview of the proposed CollaPAUL framework. CollaPAUL consists of collaborative tuning and principal axis-based unknown labeling. The proposed collaborative tuning employs an auxiliary target encoder to extract target-dependent knowledge, integrated with source-dependent knowledge from the DETR encoder via cross-domain attention. A collaborative layer is inserted between the decoder layers to convey the integrated knowledge to the decoder. The proposed principal axis-based unknown labeling generates pseudo-labels for unknown objects by creating an objectness mask and confidence mask, selecting feature proposals representing unknown objects, and assigning the unknown label to them.

with source-trained weights, struggle to learn representation knowledge for unknown objects in the target data. Unlike SFOD, SFUOD requires the model to not only adapt source knowledge for known objects but also learn effective representation knowledge for unknown objects in the target domain.

3.3. Collaborative Tuning

The proposed SFUOD scenario requires the model to transfer the learned source knowledge to the target domain and simultaneously learn novel classes as unknown objects. Since unknown objects are not observed in the source domain, the adapted source knowledge hinders the model from learning representation knowledge for unknown objects, causing knowledge confusion between known and unknown objects. To this end, we propose collaborative tuning that aims to incorporate source-dependent knowledge extracted by the student model and target-dependent knowledge by an auxiliary target encoder to preserve transferable source knowledge while capturing representation knowledge for the target domain.

We first construct the target encoder, which consists of truncated reconstruction, a cross-attention layer, and an MLP layer. Given a C-dimensional feature map f with spatial size $h \times w$ from the backbone (*e.g.*, Resnet [11]), we calculate the magnitude of the feature activation $A \in \mathbb{R}^{h \times w}$ by averaging the feature map f over the channels C as in OW-DETR [8]. Then, we select top-k activated features $f_a \in \mathbb{R}^{k \times C}$ from the A. We apply Singular Value Decomposition (SVD) like $f_a = U\Sigma V^T$, where $U \in \mathbb{R}^{k \times k}$, $\Sigma \in \mathbb{R}^{k \times C}$, and $V^T \in \mathbb{R}^{C \times C}$ denote the decomposed components, respectively, and reconstruct the feature map $f'_a = U'\Sigma' V'^T$, where $U' \in \mathbb{R}^{k \times r}, \Sigma' \in \mathbb{R}^{r \times r}$ and $V'^T \in \mathbb{R}^{r \times C}$, using the top-$r$ decomposed components to explore latent knowledge of the target domain. Since the feature f'_a is reconstructed using only the principal components that encapsulate sufficient knowledge, it effectively reveals latent representations. The reconstructed feature map f'_a then undergoes cross-attention with object queries q_{obj} and MLP layers to extract target domain representations f_t and align the dimension of the feature vector for the decoder layers. Since the target encoder operates independently from the source-trained encoder in the student model, it learns target-dependent, source-independent knowledge, capturing effective representations for the target domain.

To integrate target-dependent knowledge with source-dependent knowledge, we feed target-dependent features f_t into the decoder layers of the student model, defined as $\theta_s^{dec} = \{\theta_s^{dec_l}\}_{l=1}^{|N_{dec}|}$, where $|N_{dec}|$ denotes the number of decoder layers. To enable collaborative learning between the source-dependent knowledge f_s from the student model and the target-dependent knowledge f_t from the target encoder, we introduce collaborative layers, consisting of a cross-domain attention layer and an MLP layer. The cross-domain attention computes attention scores by aligning source and target features using identical object

queries, facilitating effective knowledge integration. Given $q = f_s \cdot W_q^T$, $k = [f_s; f_t] \cdot W_k^T$, and $v = [f_s; f_t] \cdot W_v^T$, where $q \in \mathbb{R}^{N_q \times D}$, $k \in \mathbb{R}^{2N_q \times D}$, and $v \in \mathbb{R}^{2N_q \times D}$ denote the query, key, and value for the attention layer, respectively, N_q denotes the number of object queries, and D denotes the dimension of the feature vectors, the cross-domain attention is calculated as follows:

$$attn = \sigma\left(\begin{Bmatrix} q \cdot k_s^T \\ q \cdot k_t^T \end{Bmatrix}\right) \in \mathbb{R}^{N_q \times 2 \times N_q}, \tag{3}$$

$$f_{cd} = attn \cdot \{v_s; v_t\} \in \mathbb{R}^{N_q \times D}, \tag{4}$$

where $\sigma(\cdot)$ and f_{cd} denote the softmax function and the output of the cross-domain attention layer. In Eq. (3), We compute the softmax score of each query on the corresponding source- and target-dependent features and concatenate $attn \in \mathbb{R}^{N_q \times 2N_q}$ to compute the attention value, as in Equation (4). Then, we feed f_{cd} into the MLP layers to refine the integrated knowledge.

As shown in Figure 3, to convey integrated knowledge by the collaborative (collab.) layer to the target decoder, we insert a collaborative layer between the decoder layers. Specifically, the collaborative layer is inserted into the first L decoder layers, excluding the initial layer, as its output feature is used for source-dependent representations. Here, L is a hyperparameter that determines the number of inserted layers. We set $L = 3$ based on empirical analysis. Using the first output from the decoder layer $f_s^1 = \theta_s^{dec_1}(f_s^0)$ and the output of target encoder f_t^0, where f_s^0 denotes the output of the student encoder, propagations of the $(l+1)$-th decoder and the l-th collaborative layer for collaborative tuning are calculated as follows:

$$f_t^l = \theta^{clb_l}(f_s^l, f_t^{l-1}), \tag{5}$$

$$f_s^{l+1} = \theta_s^{dec_{l+1}}([f_s^l; f_t^l]), \tag{6}$$

where θ^{clb_l} and $\theta^{dec_{l+1}}$ denote weights for the collaborative layer and decoder layer, respectively. During the decoder layer, which consists of self-attention and cross-attention, we apply prefix tuning [19] to feed f_t^l into the self-attention layer, leveraging its effectiveness in knowledge adaptation. Through the repeated propagation of the decoder and collaborative layers, the collaborative tuning trains the collaborative layer to integrate source- and target-dependent knowledge, while the decoder extracts enhanced representations using features from both the previous layer and the collaborative layer. The collaborative tuning enhances representation learning for the target domain and facilitates knowledge transfer from source to target domain.

3.4. Principal Axis-based Unknown Labeling

The SFUOD scenario requires the model to not only classify known objects but also identify novel objects as unknown, which are not defined in the source domain. However, due to the absence of knowledge of unknown objects in the source-trained model, the teacher model in the mean teacher framework struggles to assign reliable pseudo-labels to unknown objects. As a result, the detector often fails to detect unknown objects in SFUOD. To address this, we propose Principal Axis-based Unknown Labeling (PAUL), which estimates the objectness of proposal features and leverages an unknown confidence score from the prediction to effectively select proposal features for assigning unknown pseudo-labels.

The input data x_t is fed into the teacher model to extract N_q proposal features and generate predictions, including class labels and bounding boxes. First, pseudo-labels are assigned to N_k known proposals using a confidence-based approach leveraging the predicted confidence scores with the threshold, *e.g.*, we use 0.3 for all experiments. Then, to assign unknown pseudo-labels among the $N_q - N_k$ remaining proposals, the proposed PAUL is applied to filter unknown proposals and assign pseudo-labels accordingly.

Since the principal axes, computed from assigned known proposal features, represent the direction that preserves representation knowledge, we assume that unknown proposals projected onto the principal axes of known proposals exhibit similarity to projected known proposals due to sharing objectness, while non-object proposals remain dissimilar. We project both known and remaining proposals onto the principal axes to compute the objectness scores via cosine similarity. Given the $\bar{f}_{kn} = f_{kn} \cdot P^T$, $\bar{f}_{re} = f_{re} \cdot P^T$, where k_{kn}, f_{re}, and P denote known proposals, remaining proposals, and principals axes, respectively, the objectness score is formulated as follows:

$$s_{kn}^i = \frac{1}{N_k - 1} \sum_{j=1, j \neq i}^{N_k} d(\bar{f}_{kn}^i, \bar{f}_{kn}^j), \tag{7}$$

$$s_{re}^i = \frac{1}{N_{re}} \sum_{j=1}^{N_{re}} d(\bar{f}_{re}^i, \bar{f}_{kn}^j), \tag{8}$$

where $\bar{f}_{kn}^i$ and $\bar{f}_{re}^i$ represent the i-th known and remaining proposals projected onto the principal axes, respectively, $N_{re} = N_q - N_k$ denotes the number of remaining proposals, and $d(\cdot, \cdot)$ represents cosine similarity. The threshold δ is set as the average of the known objectness scores $S_{kn} = \{s_{kn}^1, s_{kn}^2, \ldots, s_{kn}^{N_k}\}$, and the objectness mask M_{obj} is constructed by comparing the remaining scores with the threshold δ. Additionally, a confidence mask M_{conf} is generated based on the confidence scores of the teacher model, using a confidence threshold ϵ. Finally, an unknown mask M_{unk} is generated by merging the objectness and confidence masks to select reliable unknown proposals. This mask ensures that selected proposals exhibit shareable objectness while representing unknown classes. The process of selecting unknown proposals is as follows:

$$M_{obj} = \{m_{obj}^i | m_{obj}^i = \mathbb{I}(s_{un}^i \geq \delta)\}_{i=1}^{N_{re}}, \tag{9}$$

Setting	Methods	Cityscape → Foggy Cityscape					
		Car	Truck	Bus	Known mAP	U-Recall	H-Score
	Source only	43.20	12.05	24.43	26.56	0.00	0.00
SFOD	Mean Teacher [28]	<u>50.20</u>	0.00	0.54	16.91	<u>6.02</u>	<u>8.88</u>
	PET [23]	36.38	1.45	0.00	12.61	2.76	4.53
	DRU [14]	41.14	9.65	18.12	22.97	3.60	6.22
	SF-UT [18]	39.82	<u>13.83</u>	<u>24.90</u>	<u>26.18</u>	0.00	0.00
SFUOD	**CollaPAUL (Ours)**	**52.10**	**16.49**	**28.37**	**32.32**	**10.59**	**15.95**

Table 1. Experimental results on the weather adaptation benchmark (Cityscapes → Foggy Cityscapes). "Source only" refers to the model trained solely on the source domain. The best and second-best performances are highlighted in bold and underlined, respectively.

Setting	Methods	Cityscape → BDD100K					
		Car	Truck	Bus	Known mAP	U-Recall	H-Score
	Source only	51.38	13.73	8.13	24.41	0.00	0.00
SFOD	Mean Teacher [28]	**59.68**	<u>13.04</u>	6.55	26.42	6.08	<u>9.89</u>
	PET [23]	48.56	6.54	2.26	19.12	1.98	3.59
	DRU [14]	44.91	4.14	2.68	17.24	<u>6.86</u>	9.82
	SF-UT [18]	52.51	12.31	**15.21**	<u>26.68</u>	1.17	2.24
SFUOD	**CollaPAUL (Ours)**	<u>57.95</u>	**17.27**	<u>9.40</u>	**28.21**	**8.57**	**13.15**

Table 2. Experimental results on the cross-scene adaptation benchmark (Cityscapes → BDD100K)."Source only" refers to the model trained solely on the source domain. The best and second-best performances are highlighted in bold and underlined, respectively.

$$M_{conf} = \{m^j_{obj} | m^j_{obj} = \mathbb{I}(c^j_{un} \geq \epsilon)\}^{N_{re}}_{j=1}, \quad (10)$$

$$M_{unk} = M_{obj} \vee M_{conf}, \quad (11)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, c^j_{un} denotes the unknown confidence score predicted for the j-th remaining proposal, and δ and ϵ denote the objectness and confidence thresholds, respectively. We apply an element-wise product between the remaining proposals and the unknown mask to identify activated proposals as unknown objects. Pseudo-labels, including class labels and bounding boxes, are then assigned based on the predictions of the teacher model for the selected proposals.

To sum up, we propose the CollaPAUL framework, which integrates collaborative tuning and principal axis-based unknown labeling. Collaborative tuning mitigates knowledge confusion caused by source-dependent knowledge, which hinders effective knowledge transfer. Integrating source- and target-dependent knowledge, collaborative tuning enables the model to enhance knowledge adaptation and improves learning of representations of the target domain. Additionally, we introduce an effective unknown pseudo-labeling to address detection failures for unknown objects. This approach selects unknown proposals based on objectness and confidence scores, enabling the model to detect unknown objects more effectively during training.

4. Experiments

4.1. Experimental Settings

Benchmarks. To benchmark the proposed SFUOD, we utilized the existing SFOD benchmarks: weather adaptation and cross-scene adaptation. In weather adaptation, Cityscapes [4] served as the source domain, while Foggy Cityscapes [26] (fog density 0.02) was the target domain. Cityscapes consists of 3,475 annotated urban images, with 2,975 for training and 500 for evaluation, while Foggy Cityscapes is generated via fog synthesis. In cross-scene adaptation, the BDD100K [31] daytime subset was used as the target domain, containing 36,728 training and 5,258 validation images with annotations, while Cityscapes remained the source domain. For both benchmarks, vehicle classes (*e.g.*, "Car", "Truck", "Bus"]) were predefined as known classes, while other categories (*e.g.*, "Person", "Rider", "Motorcycle", "Train", "Bicycle") were treated as unknown objects.

Metrics. We validated CollaPAUL by evaluating both known and unknown classes. Known class performance was measured using mean average precision (known mAP), which quantifies mAP for predefined known classes in the target domain. Unknown class detection was assessed using recall (U-Recall). To evaluate overall performance across both categories, we computed the harmonic mean (H-score), which balances known mAP and U-Recall.

Implementation Details. We used DRU [14] as the base model, as it effectively manages student learning and teacher updates within the Mean Teacher framework. For the detector, we adopted Deformable-DETR [32] with a ResNet-50 backbone pre-trained on ImageNet. CollaPAUL was trained using the AdamW optimizer, with a known object pseudo-labeling threshold and an unknown threshold δ for the proposed PAUL set to 0.3. We set α to 0.99 for the EMA updates. Training was conducted on 4 NVIDIA RTX 3090 GPUs, with a batch size of 2 per GPU.

Ablation		Cityscapes → Foggy Cityscapes		
Collab	PAUL	Known mAP	U-Recall	H-Score
		22.97	3.60	6.22
✓		30.63	3.56	6.38
	✓	25.40	6.46	10.30
✓	✓	**32.32**	**10.59**	**15.95**

Table 3. Ablation study of CollaPAUL on weather adaptation. Collab and PAUL denote the proposed collaborative tuning and principal axes-based unknown labeling.

# of Collab Layers	Cityscapes → Foggy Cityscapes		
	Known mAP	U-Recall	H-Score
$L = 1$	22.63	7.23	10.96
$L = 2$	26.98	7.25	11.43
$\boldsymbol{L = 3}$	32.32	**10.59**	**15.95**
$L = 4$	**32.90**	7.44	12.14
$L = 5$	30.33	8.29	13.02

Table 4. Ablation study for the number of inserted collaborative layers on weather adaptation.

4.2. Experimental Results

In this section, we evaluated CollaPAUL on the Weather Adaptation (Cityscapes → Foggy Cityscapes) and Cross-Scene Adaptation (Cityscapes → BDD100K) benchmarks and compared its performance with recent source-free object detection methods using the Mean Teacher framework.

Weather Adaptation. Weather adaptation captures real-world variations in weather conditions. To assess detector robustness, we conducted experiments on Cityscapes → Foggy Cityscapes. As shown in Table 1, CollaPAUL outperformed all baselines across all metrics. Notably, the mean teacher framework outperformed existing SFOD methods based on the mean teacher. Since SFOD methods focus on stabilizing teacher updates, they update the teacher model slowly, limiting its ability to learn unknown objects and restricting the student model from effectively utilizing pseudo-labels. In contrast, CollaPAUL achieved performance gains of up to 6.14% in known mAP, 4.57% in U-Recall, and 7.07% in H-Score, demonstrating its robustness to weather changes.

Cross-Scene Adaptation. Scene configurations dynamically change in different locations in the real world, especially in automated driving contexts. Hence, cross-scene adaptation is crucial in domain adaptation. As shown in Table 2, the proposed CollaPAUL surpassed the baselines across all the metrics. It gained performance enhancements of 1.53% in known mAP, 1.71% in U-Recall, and 3.26% in H-Score. This demonstrates the effectiveness of the proposed CollaPAUL in cross-scene adaptation.

4.3. Ablation Study

We conducted an ablation study on the weather adaptation benchmark to validate the performance improvements of the proposed method. Additionally, we analyzed the impact of collaborative tuning by varying the number of inserted collaborative layers and further examined the proposed principal axes-based unknown labeling.

PAUL		Cityscapes → Foggy Cityscapes		
M_{conf}	M_{obj}	Known mAP	U-Recall	H-Score
✓		30.63	3.56	6.38
	✓	29.88	9.65	14.59
✓	✓	**32.32**	**10.59**	**15.95**

Table 5. Ablation study of the proposed PAUL, which consists of confidence and objectness mask on weather adaptation.

Collab Tuning	Cityscapes → Foggy Cityscapes		
	Known mAP	U-Recall	H-Score
Baseline (only PAUL)	25.40	6.46	10.30
Prefix-tuning	28.43	8.07	12.57
Cross-domain (Ours)	**32.32**	**10.59**	**15.95**

Table 6. Analysis of the proposed cross-domain attention. Baseline denotes employing on PAUL.

Unknown Pseudo Label	Cityscapes → Foggy Cityscapes		
	Known mAP	U-Recall	H-Score
Confidence-based	30.63	3.56	6.38
Attention-driven	32.18	4.00	7.12
PAUL (Ours)	**32.32**	**10.59**	**15.95**

Table 7. Analysis of the principal axes-based unknown labeling. We used confidence-based unknown labeling as a baseline.

Ablation for the CollaPAUL. Collab and PAUL represent the proposed collaborative tuning and principal axes-based unknown labeling, with DRU as the baseline. As shown in Table 3, collaborative learning improved known mAP by 7.66%, while PAUL enhanced U-Recall by 2.86%. Their combination, CollaPAUL, achieved strong performance, showing that proposed components enhance knowledge transfer for known objects and facilitate unknown pseudo-labeling, leading to synergistic improvements.

Ablation for the Collaborative Tuning. We conducted an ablation study to determine the optimal number of collaborative layers for the proposed collaborative tuning. To ensure stable propagation through the decoders, we inserted these layers from the initial stages. As shown in Table 4, inserting three layers showed the best U-Recall and H-Score. While performance improved with up to three layers, adding more resulted in a declined performance.

Ablation for the PAUL. We further examined the effectiveness of the proposed principal axes-based unknown labeling, which includes confidence and objectness masks. Confidence-based labeling, widely used in open-set learning, served as the baseline. As shown in Table 5, employing the objectness mask significantly improved unknown object detection by 6.09%, while incorporating both masks achieved the best overall performance. This demonstrates that principal axes-based objectness estimation enhances unknown object detection.

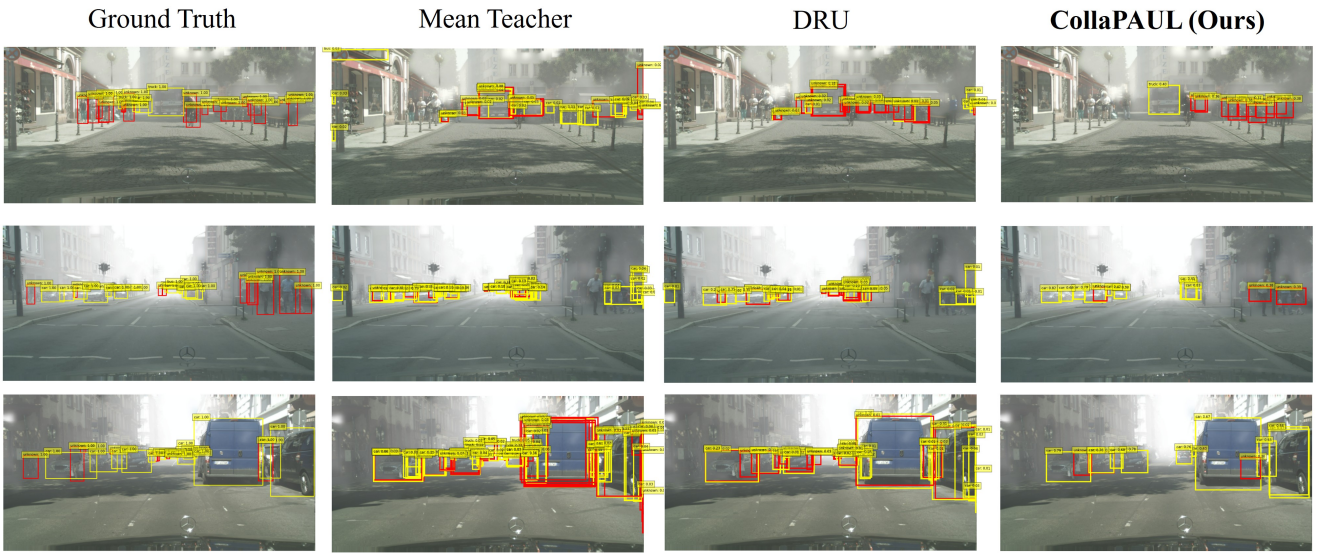


Figure 4. Qualitative results on weather adaptation. The visualizations compare the performance of Mean Teacher, DRU, and CollaPAUL, with known objects highlighted in yellow boxes and unknown objects in red boxes.

4.4. Analysis

Analysis for the Cross-domain Attention. To validate the effectiveness of the proposed cross-domain attention, designed to integrate source- and target-dependent knowledge in the collaborative layer, we compared its performance against prefix-tuning, a widely used parameter-efficient tuning method. As shown in Table 6, with PAUL as the baseline, prefix-tuning, which integrates attention knowledge from all feature vectors, improved the performance of the source-trained model. However, our cross-domain attention, which focuses on integrating source- and target-dependent knowledge extracted from identical object queries, achieved superior results across all metrics, with increases of 3.89% in known mAP, 2.52% in U-Recall, and 3.38% in H-Score. This validates its effectiveness in integrating domain-dependent representation knowledge.

Comparison of the Unknown Labeling. To validate the effectiveness of the proposed principal axes-based unknown labeling (PAUL), we compared its performance with other unknown labeling methods. Specifically, we evaluated PAUL against confidence-based approaches (baseline) and attention-driven unknown labeling from OW-DETR [8]. As shown in Table 7, PAUL outperformed all methods, achieving significant improvements of 6.59% in U-Recall and 8.83% in H-Score. While attention-driven labeling from OW-DETR showed slight gains, PAUL demonstrated superior performance. This shows that PAUL effectively enhances unknown object detection by enabling the teacher model to generate high-quality unknown pseudo-labels.

Qualitative Results on Weather Adaptation. To demonstrate the real-world applicability of our method, we conducted qualitative experiments comparing CollaPAUL with the mean teacher framework and the DRU method. We visualized results from the weather adaptation benchmark, where yellow boxes indicate known objects (*e.g.*, "Car", "Truck", "Bus") and red boxes represent novel objects (*e.g.*, "Person", "Rider", "Motorcycle", "Train", "Bicycle") predicted as unknowns. As shown in Figure 4, the mean teacher and DRU methods struggled with knowledge confusion, misclassifying known objects as unknown and vice versa, and failed to distinguish unknown objects from the background. In contrast, CollaPAUL effectively mitigated knowledge confusion, enabling the detector to correctly classify known and unknown objects while accurately identifying unknowns against the background. These results show the superior performance of CollaPAUL in the proposed source-free unknown object detection.

5. Conclusion

In this paper, we propose a novel **S**ource-**F**ree **U**nknown **O**bject **D**etection (**SFUOD**) scenario, where the model recognizes known objects while detecting novel objects as unknowns. The proposed SFUOD presents a challenging yet realistic setting due to knowledge confusion and limited knowledge of unknowns. To address this, we propose **Colla**borative tuning and **P**rincipal **A**xis-based **U**nknown **L**abeling (**CollaPAUL**). The proposed CollaPAUL mitigates knowledge confusion through effective knowledge transfer via collaborative tuning and generates unknown pseudo-labels via principal axis-based unknown labeling, estimating objectness by utilizing principal axes of known objects. While there is room for improvement in SFUOD, CollaPAUL outperforms prior SFOD methods. We expect our research to offer novel insights into domain adaptive object detection with open-set recognition for more realistic applications.

Acknowledgments

This research was conducted with the support of the HANCOM InSpace Co., Ltd. (Hancom-Kyung Hee Artificial Intelligence Research Institute), and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), and RS-2024-00457882, AI Research Hub Project).

References

[1] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1

[2] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23943–23953, 2024. 2

[3] Qiaosong Chu, Shuyan Li, Guangyi Chen, Kai Li, and Xiu Li. Adversarial alignment for source free object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 452–460, 2023. 1

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 6

[5] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4091–4101, 2021. 1

[6] Jinhong Deng, Xiaoyue Zhang, Wen Li, Lixin Duan, and Dong Xu. Cross-domain detection transformer based on spatial-aware and semantic-aware token alignment. *IEEE Transactions on Multimedia*, 26:5234–5245, 2023. 1

[7] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XV 16*, pages 567–583. Springer, 2020. 2

[8] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9235–9244, 2022. 3, 4, 8, 1

[9] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9591–9600, 2022. 3, 1

[10] Yan Hao, Florent Forest, and Olga Fink. Simplifying source-free domain adaptation for object detection: Effective self-training strategies and performance insights. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024. 1, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 733–748. Springer, 2020. 1

[13] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021. 3

[14] Trinh Le Ba Khanh, Huy-Hung Nguyen, Long Hoang Pham, Duong Nguyen-Ngoc Tran, and Jae Wook Jeon. Dynamic retraining-updating mean teacher for source-free object detection. In *European Conference on Computer Vision*, pages 328–344. Springer, 2024. 1, 2, 3, 6

[15] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8014–8023, 2022. 1

[16] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5291–5300, 2022. 1

[17] Wuyang Li, Xiaoqing Guo, and Yixuan Yuan. Novel scenes & classes: Towards adaptive open-set object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15780–15790, 2023. 3, 1

[18] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8474–8481, 2021. 1, 3, 6

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 5

[20] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 3

[21] Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3230–3239, 2023. 3

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 3

[23] Qipeng Liu, Luojun Lin, Zhifeng Shen, and Zhifeng Yang. Periodically exchange teacher-student for source-free object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6414–6424, 2023. 1, 3, 6

[24] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 3

[25] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 153–168, 2018. 2

[26] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 2, 6

[27] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17302–17312, 2024. 3

[28] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 6

[29] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 1

[30] Vibashan VS, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3520–3530, 2023. 1

[31] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6

[32] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*. 6

[33] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2023. 3