Navigating the Unknown: Continual Test-Time Adaptation for Unknown-Aware Object Detection

Shilei Cao[*,1,4], Yan Liu[*,2], Junyu Liu[1], Qingmei Li[3], and Juepeng Zheng[†,1,4]
[1]Sun Yat-sen University, [2]University of Science and Technology of China,
[3]Tsinghua University, [4]National Supercomputing Center in Shenzhen

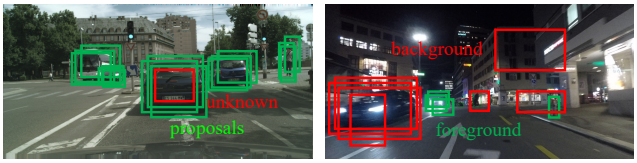
***Abstract*—Existing continual test-time adaptation frameworks operate under a restrictive closed-set assumption, presuming all objects belong to known training categories. This limitation renders them fragile in open-world scenarios where novel, unknown objects inevitably appear. In this paper, we introduce the Continual Test-Time Adaptive Unknown-aware Object Detection setting, where models must maintain performance on both unknown and known classes under continuous domain shifts. We identify two primary challenges in this setting: the ambiguity of unknowns, where parts of known objects are misclassified as unknown, and the suppression of foregrounds, where valid objects missed by the teacher model are erroneously suppressed as background. Therefore, we propose Unknown-Aware Adaptation (UAA), which incorporates IoU-based Unknown Awareness (IUA) to geometrically distinguish true unknowns from object fragments, and a Foreground Elimination Strategy (FES) that leverages the source model to recover missed detections, preventing the suppression of valid foregrounds. Extensive experiments demonstrate that UAA significantly outperforms baselines, offering robust unknown identification without compromising known-class stability.**

***Index Terms*—Continual Test-Time Adaptation, Unknown.**

I. Introduction

Deep learning models have demonstrated immense potential across various modalities, such as image [1] and language [2]. However, the performance of these models often degrades significantly when confronted with domain shift, i.e., target testing data diverges statistically from the source training data [3]–[6]. This challenge is particularly pronounced in object detection, where models trained in static environments are deployed in dynamic, uncontrolled real-world settings.

To mitigate this, Test-Time Adaptation (TTA) has emerged as a promising paradigm, which allows models to adapt directly to unlabeled test samples online during inference [7], [8]. Building upon this, Continual Test-Time Adaptation (CTTA) extends the scope to dynamic, non-stationary environments, which aims to enable models to adapt sequentially to a stream of changing target domains [9], [10]. Despite these advancements, existing CTTA frameworks typically rely on a restrictive "closed-set" assumption, presuming all test object categories were seen during training. This fails in open-world scenarios such as autonomous driving, where novel objects such as exotic animals or novel vehicle types inevitably appear [11], [12]. Forcing unknowns into known classes or ignoring them entirely risks dangerous errors or missed obstacles.



(a) UA (b) CTTA

Fig. 1: Failure modes in existing methods during CTUAOD. (a) UA Challenge: Existing selection rules incorrectly identified as a novel "unknown" sample. (b) CTTA Challenge: Missing pseudo-labels suppresses valid foreground proposals during training, degrading the model's recall.

Inspired by the cognitive principle that identifying unknown is a fundamental prerequisite for curiosity and learning [13], [14], we introduce the setting of Continual Test-time Adaptive Unknown-aware Object Detection (CTAUOD). CTAUOD assumes that unknown objects may appear unannotated during training and will persist in the test stream. The objective is: 1) maintain high performance on known classes under continuous domain shift, and 2) treat all novel instances as a unified "unknown" category to prevent misclassification. However, CTAUOD presents unique challenges that arise from the intersection of Unknown Awareness (UA) and the CTTA. Since adaptation relies on pseudo-labels rather than ground truth, the model must simultaneously distinguish between known objects, unknown objects, and background noise without explicit guidance. This leads to two specific hurdles: the **ambiguity of unknowns** and the **suppression of foregrounds**.

Firstly, current methods for identifying unknown samples during training often rely on heuristic selection rules, such as identifying proposals with high objectness scores but low overlap with known class labels [12], [15]. As illustrated in Fig. 1 (a), this heuristic is brittle. It frequently misclassifies parts of known objects (e.g., a fragment of a car) as "unknown" simply because they have high objectness but do not match the full car's bounding box. Furthermore, in CTTA, models typically use Mean Teacher frameworks [16] where high-confidence predictions from a teacher model supervise the student [9], [10]. When introducing unknowns, the criteria for "known" objects become stricter to avoid confusion. Conse-

[*]These authors contributed equally to this work.
[†]Corresponding author: zhengjp8@mail.sysu.edu.cn

This work was supported in part by National Natural Science Foundation of China under Grant 42401415; Shenzhen Science and Technology Program under Grant KCXFZ20240903093759004 and Grant KJZD20230923115106012; Guangdong Science & Technology Program under Grant 2025B0101080001; and Tsinghua SIGS KA Cooperation Fund.

quently, the teacher often fails to generate pseudo-labels for hard-to-detect known objects. In standard detection losses, the proposals related to unlabeled regions are implicitly treated as background (Fig. 1 (b)). Therefore, valid objects missed by the teacher are penalized as background, suppressing the model's ability to recall them and destabilizing the adaptation.

To address these challenges in CTAUOD, we propose Unknown-Aware Adaptation (UAA) to enhance both unknown recall and adaptation stability, which is founded on two core components: IoU-based Unknown Awareness (IUA) and Foreground Elimination Strategy (FES). Instead of relying solely on objectness scores, IUA exploits the geometric relationship between proposals. By analyzing the pairwise Intersection-over-Union (IoU) density, IUA distinguishes true unknown objects for training from the fragmented parts of known objects, significantly reducing false positive unknown detections. Complementing this, FES acts as a safeguard against the "silence" of the teacher model. It utilizes the frozen source model as a stable reference to identify potential foreground regions that the adapting teacher missed. By filtering these regions out of the background loss calculation, FES prevents the model from incorrectly learning to suppress valid objects. The key contributions of our work are summarized as follows:

- We formulate and explore the CTAUOD setting, a more practical setting that accounts for both environmental shifts and the presence of unknown objects in dynamic streams.
- We propose UAA, a unified approach designed to rectify pseudo-label noise for both known and unknown categories, ensuring robust adaptation in real-world applications.
- Extensive experiments across object detection benchmark datasets demonstrate that UAA significantly outperforms baselines in identifying unknowns while maintaining superior stability on known classes in dynamic environments.

II. Method

A. Problem Statement

In this section, we formally define the CTAUOD setting. We consider a sequence of domains $\mathcal{D} = \{D_n\}_{n=0}^{N}$. Let $D_0 = \{(x_i^{D_0}, y_i^{D_0})\}_{i=1}^{N^{D_0}}$ denote the source domain, where $x_i^{D_0}$ represents the input images and $y_i^{D_0}$ denotes the corresponding labels (bounding boxes and class labels). The subsequent domains $D_n = \{x_i^{D_n}\}_{i=1}^{N^{D_n}}$ (where $n > 0$) serve as the unlabeled target domains. During source training, only objects from the known classes $\mathcal{K} = \{1, 2, \ldots, C\} \subseteq \mathbb{N}^+$ are annotated, where $\mathbb{N}^+$ denotes the set of positive integers. We also assume that there exists a set of unknown classes $\mathcal{U} = \{C+1, \ldots\}$, with these unknown categories represented as a united category $C_{unknown}$. A source object detector F_{θ_0}, parameterized by θ_0, is pre-trained on D_0 to recognize the known classes $\mathcal{K}$ and the unknown class $C_{unknown}$. The goal of CTAUOD is to improve the performance of the source-trained model F_{θ_0} across dynamic target domains during inference time for both known and unknown classes, without using the source data. At each time step t, the model receives the unlabeled test image x^t from the current target domain, where we omit the domain-specific subscript for simplicity. The model is expected to produce predictions $\hat{y}_t = F_{\theta_{t-1}}(x^t)$ at each step t, where θ_{t-1} is the model parameter updated from previous inputs $(x_1, x_2, \ldots, x_{t-1})$. Subsequently, $\hat{y}_t$ serves as the evaluation output at time step t, and the model will adapt itself toward x^t as F_{θ_t}, a change that influences only future predictions.

B. Overview

To address the challenges of CTAUOD, we propose the Unknown-Aware Adaptation (UAA) framework. As illustrated in Figure 2, UAA consists of IoU-based Unknown Awareness (IUA) and Foreground Elimination Strategy (FES) components during the source training and CTAUOD process. We adopt Faster R-CNN [17] as the base detector, as recent works have demonstrated that two-stage architectures can achieve stronger results for the UA task [11]. During adaptation, we employ the Mean Teacher paradigm [16]. This framework involves a student model supervised by a teacher model, where the teacher model is an exponential moving average of the student.

C. IoU-based Unknown Awareness

The core challenge in identifying unknown objects is the ambiguity of unknowns. Heuristic methods often rely on high objectness scores combined with low known-class probabilities [12], [15]. However, this frequently selects discriminative parts of known objects (e.g., a car door) as novel unknown objects, as these parts exhibit high objectness but do not align with the full object's class semantics. To resolve this confusion, we introduce IUA. We posit that a valid "unknown" object should be geometrically distinct from high-confidence "known" predictions. A proposal that heavily overlaps with a known detection is likely a fragmented part of that object.

Specifically, we first employ a burn-in period of I_a iterations to allow the model to stabilize its known-class predictions. Subsequently, for every iteration, we extract the set of region proposals $\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$ generated by the Region Proposal Network (RPN). Let $\mathcal{B}_{known}$ denote the bounding boxes set of the training labels. We exclude any proposal p_i that has a high overlap with a known object:

$$\mathcal{P}' = \{p_i \in \mathcal{P} \mid \max_{b \in \mathcal{B}_{known}} \text{IoU}(p_i, b) < \mu_s\}. \tag{1}$$

where μ_s is the spatial suppression threshold. To distinguish unknown objects from background clutter within the remaining set $\mathcal{P}'$, we compute the pairwise IoU matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{P}'| \times |\mathcal{P}'|}$ between all remaining proposals:

$$\mathbf{M}_{i,j} = \text{IoU}(p_i, p_j), \quad \forall p_i, p_j \in \mathcal{P}'. \tag{2}$$

Intuitively, true objects generate a tight cluster of proposals with high mutual overlap. We quantify this density for each proposal p_i by computing an IoU-based density score s_i, which sums its overlap with all other candidate proposals: $s_i = \sum_{j=1}^{|\mathcal{P}'|} \mathbf{M}_{i,j}$. We select the top-$k$ proposals from $\mathcal{P}'$ with the highest density scores. This isolates the most salient regions in the image. From these k candidates, we select the top-m proposals that possess the highest RPN objectness scores. These final m proposals are treated as pseudo-ground truth for the "unknown" category during RPN training. Subsequently,

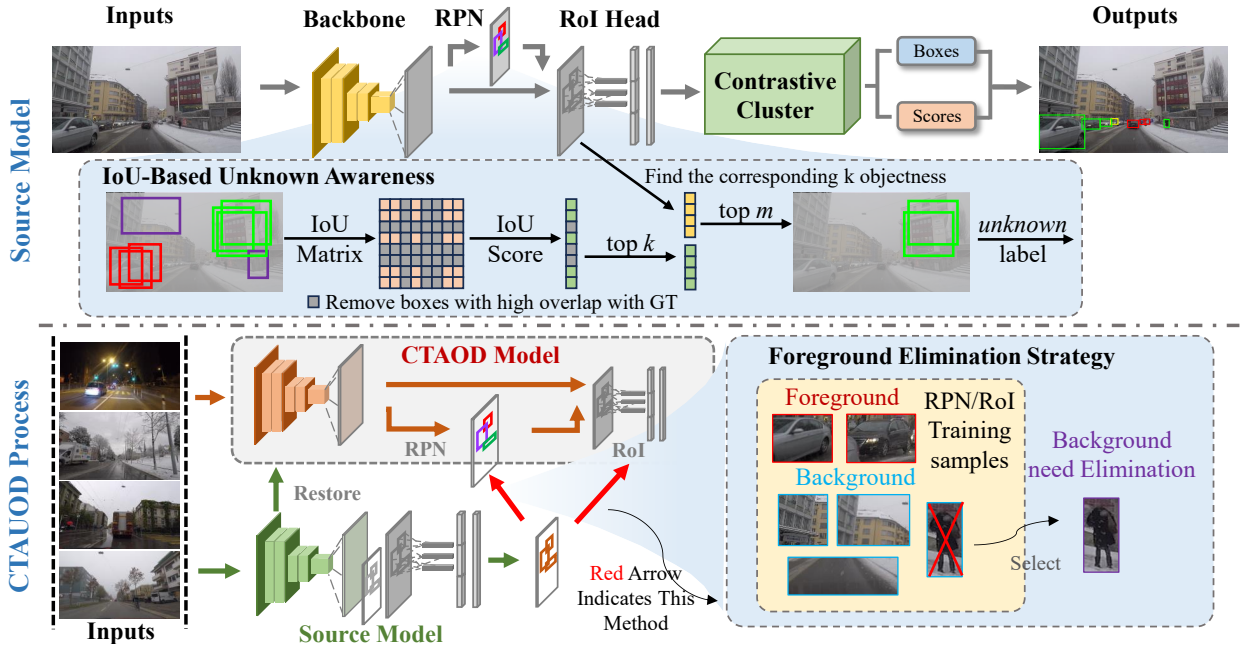


Fig. 2: Overview of UAA. 1) IUA identifies valid unknown objects by computing the pairwise IoU density of proposals. It filters out proposals with high overlap with known classes and selects the top-k density candidates to generate unknown pseudo-labels. 2) FES, which utilizes the frozen source model to detect missed valid foregrounds. These "silent" foregrounds are removed from the background negative samples in both the RPN and RoI heads to prevent from suppressing valid objects.

a contrastive clustering method [12] is applied to separate features of known and unknown classes in the latent space.

D. Foreground Elimination Strategy

In CTTA, the teacher model effectively filters noise by using a high confidence threshold to generate pseudo-labels for known classes. However, this rigorous filtering creates a side effect: valid objects with slightly lower confidence (often due to domain shift) are missed. In standard object detection training, any region not covered by a ground-truth box is implicitly treated as background. This affects training as it relies on positive and negative sampling based on their matching with training labels. Consequently, missed detections are penalized, forcing the model to learn to suppress these valid foregrounds. To counteract this, we propose FES, which utilizes the frozen source model F_{θ_0} as a "safety anchor". Since F_{θ_0} retains the original source knowledge, it often recalls objects that the adapting teacher might temporarily forget.

Specifically, we maintain the pre-trained source model F_{θ_0} throughout the adaptation. For each input image x, we utilize F_{θ_0} to generate a set of "pseudo-foreground" proposals $\mathcal{P}_f$. To ensure reliability, we select only the top-r most confident predictions: $\mathcal{P}_f = \{p_1, p_2, \ldots, p_r\}$. During student model adaptation, to prevent false background classification, we ignore background proposals that show significant overlap with any pseudo-foreground instance in $\boldsymbol{P}_f$, i.e., background proposals with $IoU \geq \mu_{t1}$ are eliminated in the RPN stage, while those with $IoU \geq \mu_{t2}$ are eliminated in the RoI head. By eliminating these regions from the background loss component, FES ensures that the student model is not penalized for detecting valid foregrounds that the teacher missed, thereby stabilizing the adaptation process and preserving recall.

III. Experiments

A. Datasets

In these experiments, we empirically validate our methodology across four datasets: Cityscapes [18], Cityscapes-C [19], ACDC [20], and SHIFT [21]. The source model is pre-trained on the Cityscapes and subsequently adapted to three domain-shifted datasets: Cityscapes-C, ACDC, and SHIFT. These benchmarks exhibit distinct distribution shifts: Cityscapes-C comprises algorithmically corrupted data, SHIFT features synthetic driving scenarios, and ACDC consists of real-world imagery captured under diverse adverse weather conditions.

B. Evaluation protocol

For the adaptation tasks, we partition base and novel classes based on *semantic overlap* and *instance frequency* to reflect diverse real-world scenarios [22], [23], resulting in four sub-tasks: 1) heterogeneous semantics (**het-sem**), where there is no semantic overlap between base and novel classes (e.g., car and person); 2) homogeneous semantics (**hom-sem**), which features semantic overlaps (e.g., car and truck); 3) frequency decrease (**freq-dec**), where base objects appear more frequently than novel counterparts; and 4) frequency increase (**freq-inc**), where novel objects outnumber base ones. Detailed results for the **freq-dec** task are presented herein, while full results for the remaining tasks are provided in the Appendix.

To comprehensively evaluate the performance of unknown and known categories, we adopt the average recall of unknown

TABLE I: Experimental results (mAP@0.5) of **Cityscapes-to-Cityscapes-C** short-term CTUAOD task under **freq-dec** setting. We evaluate by continually adapting the source model to the twelve corruptions with the largest corruption severity level 5.

Time	$t \longrightarrow$												
Condition	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	Jpeg	Mean
CoTTA	5.01	6.79	6.82	0.30	0.05	1.67	17.42	20.77	1.39	42.18	17.66	17.45	11.46
SVDP	5.03	6.24	7.12	0.51	0.02	1.77	17.86	21.43	1.22	42.11	18.15	17.39	11.57
IRG	5.11	6.14	6.12	0.21	0.01	1.39	17.05	20.77	1.24	42.00	17.64	17.20	11.24
MemCLR	5.07	6.05	5.94	0.23	0.01	1.44	17.31	19.98	1.20	41.73	17.82	16.91	11.14
Source	5.43	6.56	6.47	0.27	0.03	1.15	14.00	20.41	0.71	42.24	18.18	17.30	11.06
Source+CTTA	5.14	6.95	6.21	0.32	0.03	1.62	17.62	21.78	1.36	42.17	18.83	17.29	11.61
Source+UA	5.66	8.14	6.83	0.53	0.03	2.02	19.89	24.82	1.70	**44.46**	20.90	19.19	12.85
UAA	**6.22**	**10.42**	**8.77**	**1.64**	**0.36**	**8.23**	**34.86**	**36.16**	**8.86**	44.11	**29.90**	**28.31**	**18.15**

TABLE II: Experimental results (mAP@0.5) of **Cityscapes-to-SHIFT** short-term CTUAOD task under **freq-dec** setting.

Time	$t \longrightarrow$				
Condition	Cloudy	Overcast	Rainy	Foggy	Mean
CoTTA	26.18	27.14	25.37	27.29	26.75
SVDP	26.25	27.02	25.78	27.96	26.75
IRG	24.12	23.29	20.69	16.16	21.07
MemCLR	24.45	23.11	20.73	17.00	21.32
Source	24.60	23.75	21.06	16.99	21.60
Source+UA	25.77	25.34	22.11	17.83	22.76
Source+CTTA	27.39	29.07	29.90	27.82	28.54
UAA	**28.20**	**30.47**	**31.15**	**29.92**	**29.94**

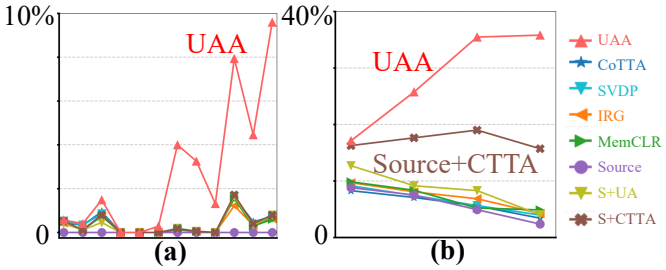


Fig. 3: Experimental results (ARu@0.5) of **Cityscapes-to-Cityscapes-C** (a) and **to-SHIFT** (b) short-term CTUAOD task under **freq-dec** setting.

(ARu) objects as the indicator. In addition, we also report the overall mean average precision (mAP) during testing. Both metrics are calculated under an IoU threshold of 0.5.

C. Implementation details

We employ the Faster R-CNN [17] with a ResNet-50 [1] backbone as the base detector. During source model training, annotations of the selected unknown classes are discarded. For testing, all unknown categories are uniformly mapped to a single index to support consistent evaluation. All algorithms are implemented based on the Detectron2 framework [24].

D. Baselines and Compared Approaches

We compare UAA with 7 baselines, including Source [17], Source+CTTA, Source+UA, CoTTA [9], SVDP [25], IRG [26] and MemCLR [27]. Specifically, Source is the base model without adaptation or unknown awareness; Source+CTTA adds adaptation, and Source+UA enables unknown awareness. We extract unknown objects from models without UA by applying a threshold-based method. To tackle Source-Free Domain Adaptation (SFDA), MemCLR integrates cross attention with CL, and IRG incorporates the instance relation graph and CL. CoTTA utilizes a fixed threshold for pseudo-labeling and random neuron recovery to tackle CTTA, while SVDP explores sparse visual prompts for CTTA dense prediction.

E. Experimental results

Short-term CTUAOD Task. We evaluate on two short-term CTUAOD tasks: Cityscapes-to-Cityscapes-C (Table I, Figure 3a, Table III) and Cityscapes-to-SHIFT (Table II, Figure 3b, Table III). The source model, augmented with Source+CTTA and Source+UA demonstrating notable improvements over the baseline. Other methods exhibit low recall for unknown classes, lacking mechanisms for handling unseen objects, as shown in Figure 3. While IRG and MemCLR, tailored for static target domains, exhibit poor performance due to conservative teacher model updates. While CoTTA and SVDP achieve relatively high accuracy, their performance remains suboptimal compared to our method. Under all unknown settings shown in Table III, our model demonstrates both improved accuracy and unknown detection capability.

Long-term CTUAOD Task. We further assess two through long-term Cityscapes-to-Cityscapes-C (Table IV, Table III) and Cityscapes-to-ACDC (Table V, Table III) CTUAOD Tasks by adapting the model to the domain group for 10 cycles. MemCLR and IRG suffer from pronounced performance degradation after multiple adaptation rounds. Although CoTTA employs stochastic restoration to mitigate forgetting and SVDP leveraging prompt learning, their effectiveness wanes. All methods perform poorly in unknown object recall. Conversely, UAA mitigates model degradation, evidenced by minimal mAP fluctuations with increasing rounds.

F. Ablation study

Component. We conduct an ablation study to empirically assess the impact of the proposed IUA and FES. As shown in Table VI, incorporating the IUA module significantly improves the ARu metric, validating our hypothesis that better unknown localization during source training enhances unknown detection performance. This improvement also contributes to a higher overall mAP, indicating a stronger synergy between unknown awareness and continual adaptation. Adding the FES module further boosts mAP. When both modules are

TABLE III: Experimental results (average mAP@0.5 across all rounds) for four CTTA tasks under different unknown settings.

Task	to-Cityscapes-C Short			to-SHIFT			to-Cityscapes-C Long			to-ACDC		
	het-sem	**hom-sem**	**freq-inc**	**het-sem**	**hom-sem**	**freq-inc**	**het-sem**	**hom-sem**	**freq-inc**	**het-sem**	**hom-sem**	**freq-inc**
CoTTA	11.1	9.7	8.8	26.3	23.5	22.6	14.4	13.7	13.3	24.8	17.6	15.4
SVDP	11.3	9.2	8.2	25.5	23.2	22.7	14.8	14.3	13.2	25.1	18.1	14.5
IRG	11.1	9.1	8.1	21.2	21.1	20.3	14.0	13.7	13.2	24.0	17.6	14.2
MemCLR	11.0	9.0	8.9	21.2	21.0	20.0	14.3	13.9	13.2	24.7	17.8	14.1
Source	11.0	9.5	8.4	21.2	21.3	20.1	10.0	10.4	9.1	20.2	18.1	13.0
Source+CTTA	11.1	9.9	8.2	22.3	21.0	20.6	10.6	10.8	9.3	20.1	18.8	13.6
Source+UA	12.6	10.4	11.8	27.5	26.0	25.2	19.8	17.2	15.7	25.4	20.9	15.8
UAA	**17.6**	**15.4**	**14.7**	**28.2**	**27.3**	**26.2**	**22.6**	**20.8**	**18.1**	**27.5**	**22.5**	**18.1**

TABLE IV: Experimental results (mAP@0.5 and ARu@0.5) of **Cityscapes-to-Cityscapes-C** long-term CTUAOD task under **freq-dec** setting at specific adaptation rounds.

Round	1		4		7		10		Mean	
Metrix	mAP	ARu	mAP	ARu	mAP	ARu	mAP	ARu	mAP	ARu
CoTTA	9.02	0.00	14.34	0.01	18.62	0.05	11.20	0.08	13.65	0.03
SVDP	11.01	0.00	19.32	0.04	20.77	0.05	17.16	0.16	15.97	0.08
IRG	11.64	0.00	18.03	0.00	20.96	0.00	16.63	0.00	15.68	0.00
MemCLR	11.19	0.00	18.88	0.00	20.14	0.00	17.76	0.00	15.91	0.00
Source	9.27	0.00	9.27	0.00	9.27	0.00	9.27	0.00	9.27	0.00
Source+UA	10.16	0.22	10.16	0.22	10.16	0.22	10.16	0.22	10.16	0.22
Source+CTTA	11.26	0.00	19.64	0.03	21.51	0.12	19.29	0.09	19.28	0.07
UAA	**13.57**	**0.55**	**23.56**	**4.92**	**25.26**	**7.98**	**22.86**	**5.69**	**22.86**	**5.69**

TABLE V: Experimental results (mAP@0.5 and ARu@0.5) of **Cityscapes-to-ACDC** long-term CTUAOD task under **freq-dec** setting at specific adaptation rounds.

Round	1		4		7		10		Mean	
Metrix	mAP	ARu	mAP	ARu	mAP	ARu	mAP	ARu	mAP	ARu
CoTTA	26.09	0.00	32.27	0.02	33.24	0.08	34.54	4.67	31.77	1.09
SVDP	26.95	0.00	33.98	0.01	34.02	0.11	34.54	5.39	33.10	1.21
IRG	26.77	0.00	32.88	0.00	27.33	0.27	26.38	4.02	28.63	1.07
MemCLR	25.51	0.00	31.69	0.00	29.08	0.14	28.06	3.84	29.42	0.94
Source	24.41	0.00	24.41	0.00	24.41	0.00	24.41	0.00	24.41	0.00
Source+UA	24.37	1.96	24.37	1.96	24.37	1.96	24.37	1.96	24.37	1.96
Source+CTTA	26.86	0.00	34.75	0.00	34.90	0.19	34.54	5.89	33.69	1.17
UAA	**29.50**	**3.82**	**38.31**	**18.55**	**37.09**	**18.26**	**36.19**	**15.56**	**35.93**	**15.62**

TABLE VI: Alation on components. All experiments are done on long-term **Cityscapes-to-Cityscapes-C** under **freq-dec**.

	IUA	FES	mAP	ARu
1			18.82	0.88
2	✓		19.66	4.07
3		✓	20.21	2.54
4	✓	✓	**21.81**	**4.29**

TABLE VII: Alation on IUA threshold. Experiments are done on long-term **Cityscapes-to-ACDC** tasks under **freq-dec**.

ARu m \ k	5	10	15	20
1	14.19	**15.62**	10.48	10.21
2	14.01	15.18	10.52	10.16
3	12.67	13.29	10.41	10.24
4	12.21	13.32	10.33	10.51

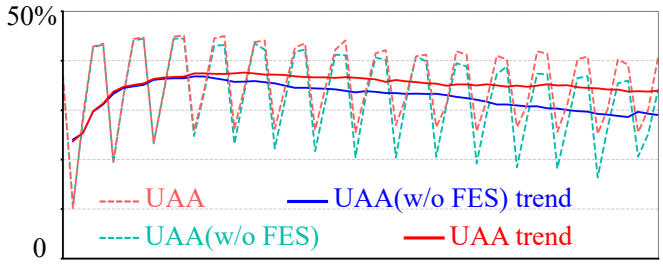


Fig. 4: Experimental results (mAP@0.5) and trend lines for the 30-round **Cityscapes-to-ACDC**.

applied, the model achieves the best overall performance. To verify FES's effectiveness in mitigating model degradation, we extend the evaluation to 20 adaptation steps. As illustrated in Figure 4, the mAP drops more slowly with FES, confirming its stabilizing effect. We present ablation results on the CTTA and UA components of the UAA model in Tables I, II, IV, V and Figure 3. The results clearly demonstrate the effectiveness of the CTTA module in improving detection accuracy, and the UA module in enhancing the recall of unknown objects.

IUA threshold. Recall that in method, we select proposals with top-k IoU-based score and top-m objectness score as unknown. The ablation on m and k are given in table VII. As k increases, the model progressively relies more exclusively on objectness scores for unknown. Meanwhile, m governs the model's convergence rate - excessively large values of m tend to introduce more noise into the learning process.

G. Qualitative Study

Fig. 5 presents a qualitative comparison at the 10^{th} round of the Cityscapes-to-ACDC adaptation task. The baseline CoTTA suffers from severe catastrophic forgetting, failing to detect obvious known objects. In contrast, our method maintains robust stability on known classes due to the FES. Our method demonstrates superior stability, successfully recalling known instances even under severe domain shift. This robustness is attributed to the FES, which effectively utilizes the source model to recover missed foregrounds. Furthermore, our model accurately localizes novel unknown objects (highlighted in red) that the baseline completely ignores. This validates the efficacy of the IUA module in distinguishing true unknowns from background noise via geometric density.

IV. Conclusion

In this work, we address the challenge of novel category emergence during continual test-time adaptation by introducing a new problem setting: Continual Test-Time Adaptive Unknown-aware Object Detection (CTAUOD), along with comprehensive benchmarks and evaluation protocols. To tackle the key challenges of CTAUOD, which incorporates two key technical contributions. First, we introduce an IoU-based Unknown Awareness module to more accurately localize unknown objects, significantly improving unknown recall. Second, we design a Foreground Elimination Strategy to mitigate model degradation caused by self-supervised learning in object detection, especially under long adaptation sequences. Exten-

Fig. 5: Visualization samples in 10^{th} round of Cityscapes-to-ACDC long-term task, and compare among the ground truth (a), CoTTA [9] (b), and our model (c). We highlight the locations of unknown objects with red bounding boxes.

sive experiments across four different CTAUOD tasks and four various unknown settings demonstrate the effectiveness of UAA in both short-term and long-term scenarios.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019.

[3] Y. Liang, S. Cao, J. Zheng, X. Zhang, J. Huang, and H. Fu, "Low saturation confidence distribution-based test-time adaptation for cross-domain remote sensing image classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 139, p. 104463, 2025.

[4] H. Liang, S. Cao, Y. Lai, and J. Zheng, "Federated open-set domain generalization with adaptive adjustment boundary and weights," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[5] Z. Ye, G. Li, H. Liang, Z. Wang, S. Cao, Y. Lai, and J. Zheng, "Quantifying samples with invariance for source-free class incremental domain adaptation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

[6] Q. Li, Y. Zhang, P. Zhang, H. Fu, and J. Zheng, "Sage: Style-adaptive generalization for privacy-constrained semantic segmentation across domains," *arXiv preprint arXiv:2512.02369*.

[7] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[8] H. Liang, X. Zhang, S. Cao, G. Li, and J. Zheng, "Tta-feddg: Leveraging test-time adaptation to address federated domain generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 18, 2025, pp. 18 658–18 666.

[9] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.

[10] S. Cao, J. Zheng, Y. Liu, B. Zhao, Z. Yuan, W. Li, R. Dong, and H. Fu, "Exploring test-time adaptation for object detection in continually changing environments," *arXiv preprint arXiv:2406.16439*, 2024.

[11] A. Dhamija, M. Gunther, J. Ventura, and T. Boult, "The overlooked elephant of object detection: Open set," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1021–1030.

[12] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *The IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[13] J. A. Meacham, "Wisdom and the context of knowledge: Knowing that one doesn't know," *On the development of developmental psychology*, vol. 8, no. 111-134, p. 1, 1983.

[14] S. Engel, "Children's need to know: Curiosity in schools," *Harvard educational review*, vol. 81, no. 4, pp. 625–645, 2011.

[15] Y. Wang, Z. Yue, X.-S. Hua, and H. Zhang, "Random boxes are open-world object detectors," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[16] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[18] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[19] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[20] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 765–10 775.

[21] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu, "Shift: a synthetic driving dataset for continuous multi-task domain adaptation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[22] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?" 2021.

[23] W. Li, X. Guo, and Y. Yuan, "Novel scenes & classes: Towards adaptive open-set object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[24] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2.

[25] S. Yang, J. Wu, J. Liu, X. Li, Q. Zhang, M. Pan, Y. Gan, Z. Chen, and S. Zhang, "Exploring sparse visual prompt for domain adaptive dense prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 16 334–16 342.

[26] V. VS, P. Oza, and V. 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*, 2023, pp. 3520–3530.

[27] ——, "Towards online domain adaptive object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 478–488.