Fully Test-time Adaptation for Object Detection

Xiaoqian Ruan and Wei Tang
University of Illinois Chicago
{xruan9,tangw}@uic.edu

Abstract

Though the object detection performance on standard benchmarks has been improved drastically in the last decade, current object detectors are often vulnerable to domain shift between the training data and testing images. Domain adaptation techniques have been developed to adapt an object detector trained in a source domain to a target domain. However, they assume that the target domain is known and fixed and that a target dataset is available for training, which cannot be satisfied in many real-world applications. To close this gap, this paper investigates fully test-time adaptation for object detection. It means to update a trained object detector on a single testing image before making a prediction, without access to the training data. Through a diagnostic study of a baseline self-training framework, we show that a great challenge of this task is the unreliability of pseudo labels caused by domain shift. We then propose a simple yet effective method, termed the IoU Filter, to address this challenge. It consists of two new IoU-based indicators, both of which are complementary to the detection confidence. Experimental results on five datasets demonstrate that our approach could effectively adapt a trained detector to various kinds of domain shifts at test time and bring substantial performance gains. Code is available at https://github.com/XiaoqianRuan1/IoU-filter.

1. Introduction

Object detection is a fundamental task in computer vision that deals with recognizing and locating objects in an image. Though deep learning approaches [9, 21, 27, 32] have drastically pushed forward the state-of-the-art object detection performance on standard benchmarks, current object detectors are often vulnerable to domain shifts between the training data and testing images, *e.g.*, unseen styles, weather, lighting conditions, and noise.

Domain adaptation techniques have been developed to adapt an object detector trained in a source domain to a target domain so that it will be robust to the domain shift. Unsupervised domain adaptation (UDA) methods [5, 20, 23, 25] require both labeled source data and unlabeled target data. This is undesirable as the source data are often unavailable for privacy and profit concerns. Source-free domain adaptation (SFDA) methods [1, 6, 10, 16, 31] have been developed to overcome this limitation: a detector trained on the source data is adapted to the target domain without access to the source data.

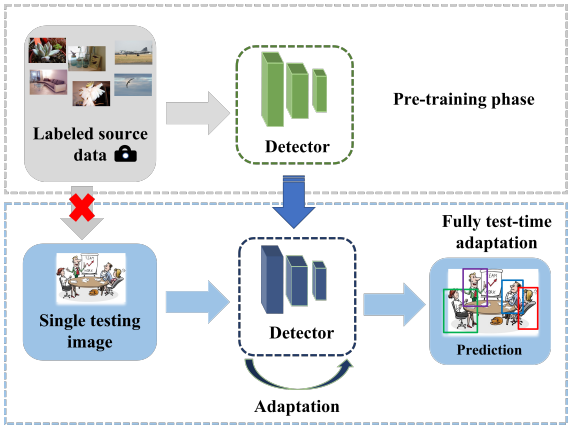


Figure 1. The task of fully test-time adaptation for object detection means to update a trained object detector on a single testing image before making a prediction, without access to the training data.

Both UDA and SFDA assume that the target domain is known and fixed and that a target dataset sampled from this domain is available for training. However, the real world is complex and non-stationary, which is unlikely to be covered by any fixed dataset. The detector must adapt itself *on the fly* to the unknown and varying domain shift at test time. This is desired in many real-world applications, from intelligent assistants that help visually impaired people read images and social media that automatically tag photos uploaded by users, to autonomous vehicles that drive safely as the place, weather, and pedestrian density change.

Test-time adaptation [4, 13, 26] has been developed to address this challenging but important problem. It does not anticipate the distribution shift, but instead learns from it at test time: the trained model is updated based on a single

test sample before making a prediction. However, this line of work focuses on image classification and requires access to the source data. Recently, TENT [29] addresses fully test-time adaptation, which is source-free, but it relies on a batch of test samples to estimate the normalization statistics and still focuses on image classification.

To close this gap, this paper investigates fully test-time adaptation for object detection. As illustrated in Fig. 1, it means to update a trained object detector, *e.g.*, Faster-RCNN [21], on a single testing image before making a prediction, without access to the training data. Compared with UDA and SFDA, we neither assume a stationary and known target domain nor have a target dataset. It will facilitate many applications, *e.g.*, image understanding systems for social media and visually impaired people, where the target domain differs from image to image, hence adaptation can be learned only from one sample.

We first introduce a baseline approach for this task, built on the classical self-training framework. It iteratively obtains pseudo labels on the testing image based on the detection confidence and uses pseudo labels to update the detector. Finally, the detector from the last iteration makes a prediction on the testing image. Our diagnostic study shows that this framework is promising, but its performance is largely bottlenecked by the low-quality pseudo labels, caused by the domain shift. The pseudo labels are very noisy even at a high confidence threshold.

We propose a new method, termed the IoU (Intersection over Union) Filter, to obtain higher-quality pseudo labels in the presence of domain shift. It consists of two new IoU-based indicators, both of which are complementary to the detection confidence. The first indicator, IoU between Consecutive Iterations (IoU-CI), matches object detections at the current self-training iteration with those at the previous iteration based on their classes and locations. Then, the IoU between these matched detections is used to select pseudo labels. The second indicator, IoU between Overlapped Detections (IoU-OD), removes the duplicate detections of the same instance as different classes, which is caused by the classification ambiguity under domain shift. Our statistical results indicate that both indicators increase the percentage of correct pseudo labels and thus significantly improve the object detection performance at test time.

It is worth noting that our task setting is different from the one-shot unsupervised cross-domain detection (OSHOT) [2] and the online domain adaptive object detection (ODA) [28].

OSHOT [2] performs unsupervised adaptation across domains by solving a self-supervised auxiliary task (i.e., rotation classification) on only one target sample seen at test time. However, it needs to add an auxiliary prediction head to the detection model and learn the self-supervised task on the training data. Thus, it is not source-free. Nevertheless, we show that our proposed approach is also effective under this setting, which demonstrates its versatility.

ODA [28] adapts a detector to a target dataset in an online manner. Each sample arrives sequentially and updates the model continuously. Testing and evaluation are performed after the source model has been trained on all samples in the target dataset. In addition, the core of their approach is a novel memory module (MemXformer) that stores prototypical patterns of the target distribution to avoid forgetting. This added MemXformer is pretrained on the source data and thus is not source-free.

The contribution of this paper is summarized as follows.

- To our knowledge, this is the first work on fully test-time adaptation for object detection. Compared with the popular UDA and SFDA, it neither assumes a stationary and known target domain nor requires access to a target dataset. This is desired in many image understanding applications, where the target domain is unknown a prior and differs from image to image.
- Through a diagnostic study of a baseline self-training framework, we show that a great challenge of this task is the unreliability of pseudo labels caused by domain shift. We propose a simple yet effective method, *i.e.*, IoU Filter, to address this challenge. It includes two new IoU-based indicators and selects higher-quality pseudo labels in the presence of domain shift.
- Experimental results on five datasets demonstrate that our approach could effectively adapt a trained detector to various kinds of domain shifts at test time and bring substantial performance gains.

2. Related Work

2.1. Test-time Adaptation

Test-time adaptation or training [26] aims at updating a trained model on a single unlabeled test sample before making a prediction to increase its robustness to distribution shift. Sun et al. [26] create a self-supervised auxiliary task (rotation classification) to train the model on this single test sample. Chen et al. [4] propose AdaContrast, based on self-supervised contrastive learning and an online pseudo-labeling scheme. Kim et al. [13] focus on the test-time adaptation of event-based object recognition, by leveraging the temporal structure of events. However, these methods require access to the training data. To address this limitation, TENT [29] introduces fully test-time adaptation, which directly minimizes the entropy of a model's predictions at test time. Recently, Wang et al. [30] extend TENT to continually changing environments. They apply weight-averaged and augmentation-averaged predictions to reduce the error accumulation and stochastically restore weights to avoid catastrophic forgetting. All these test-time adaptation methods focus on classification tasks.

2.2. Domain-adaptive Object Detection

A variety of methods have been developed to adapt an object detector trained in a source domain to a target domain [20], based on adversarial feature learning [5, 23, 25], self-training [11, 12, 22], image-to-image translation [3, 34], and domain randomization [14]. However, they generally require the source data, which are unavailable in some practical scenarios. This limitation motivates the work on source-free unsupervised domain-adaptive object detection. Some methods are built on the self-training framework. Li et al. [18] treat the prediction uncertainty as self-entropy and propose a new metric called self-entropy descent (SED) to search the optimal confidence threshold. Ahmed et al. [1] introduce a Negative Ensemble Learning (NEL) technique for noise filtering and pseudo label refinement, which tackles noisy pseudo labels by enhancing the diversity in ensemble members. Lee et al. [17] propose the Joint Model-Data Structure (JMDS) score, including a Log Probability Gap (LPG) and a Model Probability of Pseudo-Label (MPPL) score, to measure the importance of samples. However, this line of work assumes that the target domain is known and fixed and that a target dataset sampled from this domain is available for training.

Different from this line of existing work, we neither assume a stationary and known target domain nor require access to a target dataset. Instead, we aim to update a trained object detector on a single testing image before making a prediction, without access to the training data. In addition, we propose a new method, *i.e.*, the IoU Filter, to effectively address this challenging but important task.

3. Method

3.1. Problem Setting

We formally introduce fully test-time adaptation for object detection. At test time, we are provided with a trained object detector, *e.g.*, Faster RCNN, with parameters $\boldsymbol{\theta}_0$ and a single testing image $\boldsymbol{I}$. There is no access to the source data where the detector was originally trained nor a target *dataset* sampled from a known target domain. Then, we will adapt the detector on $\boldsymbol{I}$ and obtain updated parameters $\boldsymbol{\theta}_T$. Following the setting of test-time adaptation for classification [29], we allow the model to be updated multiple iterations on this single testing image. Finally, we will use the updated detector $\boldsymbol{\theta}_T$ to make a prediction on $\boldsymbol{I}$.

3.2. Self-training Baseline: A Diagnostic Study

We will first introduce a baseline approach built on the classical self-training framework [15, 19, 22], as it has been shown to be effective in learning from unlabeled data. Then, we will present an empirical study on its effectiveness in our task.

Iteration	Method	Comic	Clipart	Watercolor
$t = 0$		18.45	28.01	43.83
$t = 1$	All detections	19.83	28.65	45.42
	Detection confidence	18.58	28.42	45.23
$t = 5$	All detections	20.17	29.57	35.79
	Detection confidence	19.11	30.49	46.21

Table 1. Performance of fully test-time adaptation obtained by the self-training baseline on three datasets. $t = 0$ means the original Faster RCNN detector trained on the Pascal VOC dataset. The pseudo labels are simply all object detections or filtered by the detection confidence.

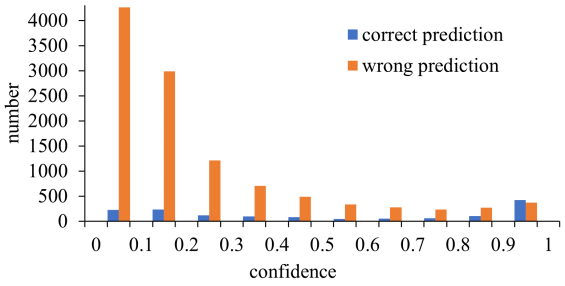


Figure 2. The numbers of correct predictions and wrong predictions at different detection confidence intervals, obtained on the Comic2k dataset.

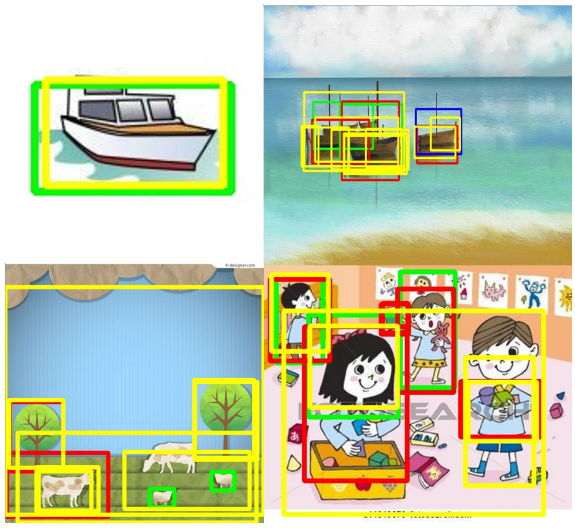
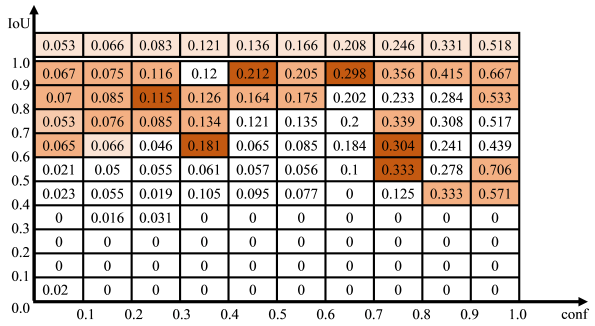
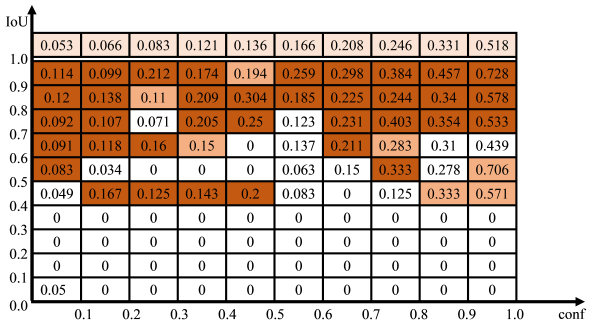


Figure 3. Illustration of consistent and inconsistent object detections between two consecutive self-training iterations. The IoU threshold is set to 0.9. Green boxes are consistent and correct detections. Red boxes are consistent but wrong detections. Yellow boxes are inconsistent and wrong detections. Blue boxes are inconsistent but correct detections. Green and yellow boxes are desired but red and blue boxes are not.



(a) Without the IoU-OD filter



(b) With the IoU-OD filter

Figure 4. In each table, the first row is the probabilities of correct pseudo labels conditioned on different detection confidence intervals. The second to last rows are the probabilities of correct pseudo labels conditioned on both different detection confidence intervals and different IoU-CI thresholds. The pseudo labels have been filtered by the IoU-OD filter in (b). In both grids, highlighted cells (light/dark brown) in the second to last rows mean their values are higher than the corresponding values in the first row. Dark brown means a higher value between two corresponding cells in (a) and (b).

3.2.1 Self-training Baseline

It is an iterative algorithm. At the tth iteration ($t \in \{1, \dots, T\}$), the current detector $\boldsymbol{\theta}_{t-1}$ makes a prediction $\mathcal{D}_t = \{(\boldsymbol{b}_{t,i}, \boldsymbol{p}_{t,i}) : \forall i\}$ on $\boldsymbol{I}$, where $\boldsymbol{b}_{t,i}$ is the bounding box of the ith object instance and $\boldsymbol{p}_{t,i} \in [0,1]^K$ is the probability distribution of the K classes. The maximum probability within $\boldsymbol{p}_{t,i}$ and its index respectively define the detection confidence $c_{t,i} \in [0,1]$ and the object class $y_{t,i} \in \{1, \dots, K\}$. We then collect confident detections as pseudo labels: $\mathcal{P}_t = \{(\boldsymbol{b}_{t,i}, y_{t,i}) : c_{t,i} > \lambda^{\text{conf}}\}$, where λ^{conf} is the confidence threshold. Finally, we tune the current detector $\boldsymbol{\theta}_{t-1}$ on the pseudo labels via a gradient descent step and obtain the updated model $\boldsymbol{\theta}_t$.

At the first iteration, *i.e.*, $t = 1$, the current detector $\boldsymbol{\theta}_{t-1}$ is initialized as the model $\boldsymbol{\theta}_0$ trained on the source data. After the last iteration, *i.e.*, $t = T$, the model $\boldsymbol{\theta}_T$ will be used to make a final prediction on $\boldsymbol{I}$. Obviously, this self-training framework does not modify the network architecture and is source-free.

3.2.2 Diagnostic Study

We validate the effectiveness of this baseline in our task. The source-detector is a Faster RCNN trained on the Pascal VOC dataset [8]. It performs fully test-time adaptation on each individual testing image from three datasets of different domains. The RoI classification loss is not used, which leads to better performance. Tab. 1 shows the results obtained using the optimal confidence threshold.

We have two observations. First, the baseline consistently improves the performance of the original detector. This demonstrates the potential of the self-training framework in our task. Second, in most scenarios, using detection confidence to select pseudo labels leads to similar performance as using all detections as pseudo labels. Meanwhile, Fig. 2 shows that the pseudo labels are noisy even at a high confidence threshold. These observations motivate us to hypothesize that the low quality of pseudo labels caused by domain shift is the main challenge faced by the self-training framework in our task.

3.3. IoU Filter

We introduce a new method, termed the IoU Filter, to obtain higher-quality pseudo labels in the presence of domain shift. As will be described below, it consists of two new IoU-based indicators that are complementary to the detection confidence.

3.3.1 IoU between Consecutive Iterations (IoU-CI)

This indicator stems from our observation that object detections which are consistent over two consecutive self-training iterations are more likely to be correct than those that are inconsistent, as illustrated in Fig. 3.

Formally, at the tth iteration ($t \in \{2, \dots, T\}$), $\mathcal{D}_t = \{(\boldsymbol{b}_{t,i}, \boldsymbol{p}_{t,i}) : \forall i\}$ denotes the prediction made by the current detector $\boldsymbol{\theta}_{t-1}$ on $\boldsymbol{I}$. For every object instance in $\mathcal{D}_t$, we match it to an instance in $\mathcal{D}_{t-1}$ with the same class and minimum IoU. $\mathcal{D}_{t-1}$ is the prediction made in the previous iteration. Then, the IoU-CI score of an instance in $\mathcal{D}_t$ is defined as the IoU between itself and its matched instance in $\mathcal{D}_{t-1}$. For instances without a match in the previous iteration, their IoU-CI scores are zero.

Fig. 4a compares the percentages of correct pseudo labels at different detection confidence intervals and different IoU-CI score intervals. We can see that when the IoU-CI score is higher than 0.7, it improves the quality of pseudo labels at most detection confidence intervals. Thus, we filter pseudo labels with an IoU-CI threshold $\lambda^{\text{IoU-CI}}$.

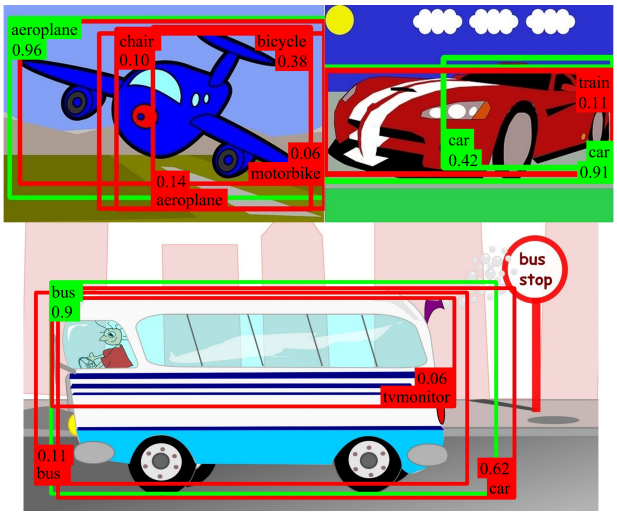


Figure 5. Illustration of repeated detections of the same instance as different object classes in the presence of domain shift. The detection with the highest confidence is most likely to be correct.

3.3.2 IoU between Overlapped Detections (IoU-OD)

We find that the detector tends to repeatedly detect the same instance as different object classes in the presence of domain shift, as illustrated in Fig. 5. In this scenario, the detection with the highest confidence is most likely to be correct, while the others are false positives. This observation motivates us to define an IoU-OD filter.

At the tth iteration ($t \in \{1, \ldots, T\}$), we have the collection of all object detections $\mathcal{D}_t$. For an object instance i in $\mathcal{D}_t$, we compare its detection confidence c_i with those of instances whose IoU w.r.t. instance i are higher than a threshold $\lambda^{\text{IoU-OD}}$. If c_i is the highest among them, instance i passes the IoU-OD filter, otherwise, it is excluded from the pseudo labels. The IoU-OD filter is like *class-agnostic* non-maximum suppression (NMS). It is designed to handle ambiguous object classification caused by domain shift. In contrast, current object detectors apply NMS to the detections of each class separately because they usually classify objects unambiguously when there is no domain shift.

Fig. 4b compares the percentages of correct pseudo labels in different confidence and IoU-CI intervals after applying the IoU-OD filter. We can see that it further improves the quality of pseudo labels.

4. Experiments

4.1. Experimental Setting

4.1.1 Datasets

We use five datasets of different domains for evaluation. Clipart1k, Comic2k, and Watercolor2k are three artistic media datatsets [11] and include artistic images in a variety of styles. They have been commonly used to benchmark domain adaptive object detection methods when the source dataset is Pascal-VOC [8]. Clipart1k includes the same 20 object categories as Pascal-VOC. Both Comic2k and Watercolor2k include 6 classes, which are a subset of the 20 classes of Pascal-VOC [8]. Each of them consists of 1000 training images and 1000 testing images. Foggy Cityscapes [24] and Rainy Cityscapes [25] are two datasets by adding different levels of synthetic fog and rain to original Cityscapes images [7], which are the source data. Only the highest level of fog and rain is considered. The Foggy Cityscapes dataset contains 492 images for evaluation while Rainy Cityscapes dataset has 99 testing images. Only the test sets of these five datasets are used for evaluation. Mean average precision (mAP) at the IoU threshold 0.5 is used as the performance metric.

4.1.2 Implementation Details

The object detector is Faster-RCNN with a RseNet50 backbone pre-trained on ImageNet, without any modification to its network architecture. Its region proposal network (RPN) produces 300 top proposals after non-maximum-suppression (NMS), based on anchors at three scales (128, 256, 512), and three aspect ratios (1:1, 1:2, 2:1). The source-trained model is trained for 70k iterations on Pascal-VOC, and 30k iterations on Cityscapes, respectively, using SGD with momentum set at 0.9, the initial learning rate set as 0.001. The batch size is set as 1. For each single testing image, we update this trained model for 5 iterations, using SGD with momentum set at 0.9, and learning rate 0.001. For all five testing datasets, the detection confidence threshold is set as 0.6, the IoU-CI threshold is set as 0.6, and the IoU-OD threshold is set as 0.9. We only use the detection confidence threshold to filter detections that occupy almost the entire image, *i.e.*, more than 90% pixels, which empirically leads to better performance.

4.1.3 Benchmark Methods

We will compare our proposed approach with CoTTA [30], a fully test-time adaptation method for classification, and the self-training baseline introduced earlier, which only uses detection confidence to select pseudo labels. Recently, Xu et al. [33] demonstrate that the bounding box uncertainty helps remove low-quality pseudo labels in semi-supervised object detection. We integrate it into the self-training baseline and compare it with our approach.

4.1.4 OSHOT Setting

This is the setting of one-shot unsupervised cross-domain detection [2], which trains Faster-RCNN along with an auxiliary task (*i.e.*, image rotation classification). During train-

		Comic	Clipart	Watercolor
t	Method	mAP	mAP	mAP
0		18.45	28.01	43.83
1	CoTTA	18.57	28.28	44.44
	Self-training	18.58	28.42	45.23
	Self-training+U	18.23	28.37	**45.53**
	Ours	**19.06**	**28.58**	44.88
2	CoTTA	19.07	28.39	45.47
	Self-training	19.27	28.84	44.44
	Self-training+U	19.03	28.84	45.51
	Ours	**20.67**	**29.98**	**46.16**
3	CoTTA	19.15	27.83	46.06
	Self-training	20.24	28.02	46.61
	Self-training+U	20.49	29.43	46.03
	Ours	**21.54**	**31.32**	**47.29**
4	CoTTA	18.11	27.33	46.70
	Self-training	20.16	29.88	45.76
	Self-training+U	19.43	29.41	46.70
	Ours	**22.92**	**31.95**	**46.92**
5	CoTTA	18.53	26.92	45.97
	Self-training	19.11	30.49	46.21
	Self-training+U	19.98	30.13	45.96
	Ours	**22.51**	**32.56**	**46.48**

Table 2. Object detection results on three artistic media datasets under fully test-time adaptation setting. The source dataset is the Pascal-VOC dataset [8]. Self-training+U means the integration of the self-training baseline and uncertainty modeling [33].

		RainyCityscape	FoggyCityscape
t	Method	mAP	mAP
0		25.02	26.11
1	CoTTA	**26.16**	26.84
	Self-training	25.17	26.23
	Self-training+U	25.17	26.58
	Ours	26.12	**26.90**
2	CoTTA	26.75	26.97
	Self-training	25.70	26.44
	Self-training+U	24.77	26.70
	Ours	**28.58**	**27.71**
3	CoTTA	27.04	27.55
	Self-training	26.76	26.62
	Self-training+U	24.76	26.65
	Ours	**30.19**	**28.66**
4	CoTTA	28.48	28.41
	Self-training	25.94	27.03
	Self-training+U	25.02	26.79
	Ours	**32.70**	**30.08**
5	CoTTA	29.53	26.59
	Self-training	24.46	26.35
	Self-training+U	24.39	26.31
	Ours	**32.87**	**28.58**

Table 3. Results on RainyCityscape and FoggyCityscape datasets under the fully test-time adaptation setting. The source dataset is the original Cityscapes dataset. Self-training+U denotes the combination of the self-training baseline and uncertainty modeling [33].

ing, the weight of the auxiliary task is set as 0.05. At test time, the detector is updated based on both the detection loss and the self-supervised loss. For all three datasets, the weight of the detection loss is set as 1.0, and the weight of the self-supervised loss is 0.2.

As the network architecture is modified and the self-supervised head needs to be trained on the source data, this setting is not source-free.

4.2. Experiment Results

4.2.1 Fully Test-time Adaptation

Tab. 2 and Tab. 3 show the object detection results on the five testing datasets, under the fully test-time adaptation setting. Four methods are compared, including the CoTTA [30], self-training baseline, its integration with uncertainty modeling, and our proposed method. t is the number of self-training iterations. The best performance of each iteration is bold. Our method outperforms the other methods under most settings. From Tab. 2, the performance of each method does not necessarily improve with more iterations.

4.2.2 OSHOT

Tab. 4 and Tab. 5 show the results obtained under the OSHOT setting [2] (Sec. 4.1.4) with five testing datatsets. We can see that using our IoU Filter to select pseudo labels in [2] improves its performance under most of the settings. Comparing Tab. 3 and Tab. 5, we could observe that the performance obtained under the OSHOT setting is generally better than that obtained in fully test-time adaptation. This is expected as the training data are assumed available under the OSHOT setting and they could provide useful information for test-time adaptation.

4.3. Ablation study

We first validate the contribution of each model component (Sec. 4.3.1), then study the impact of different thresholds on the performance (Sec. 4.3.2), and finally show the results obtained after more than five self-training iterations (Sec. 4.3.3). All results are reported at the fifth self-training iteration (except for Sec. 4.3.3).

4.3.1 Component Analysis

Results are shown in Tab. 6. We can see that each of two indicators IoU-CI and IoU-OD improves the performance of

		Comic	Clipart	Watercolor
t	Method	mAP	mAP	mAP
0		17.59	28.02	43.76
1	OSHOT	19.55	28.04	45.43
	Ours	**21.59**	**29.36**	**47.10**
2	OSHOT	21.34	29.21	**47.36**
	Ours	**23.28**	**30.15**	46.96
3	OSHOT	22.25	30.03	47.75
	Ours	**24.17**	**30.80**	**47.79**
4	OSHOT	24.47	30.05	**48.40**
	Ours	**24.90**	**32.63**	47.76
5	OSHOT	24.59	30.86	47.80
	Ours	**25.23**	**32.80**	**48.03**

Table 4. Object detection results under the OSHOT setting based on three artistic datasets. The source dataset is the Pascal-VOC dataset. Ours: using the proposed IoU Filter to select pseudo labels in OSHOT [2].

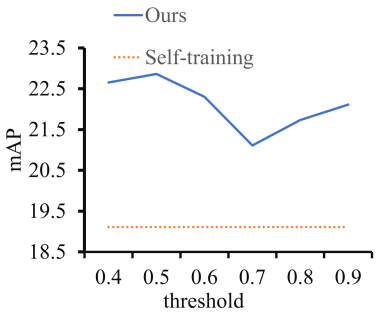
		RainyCityscape	FoggyCityscape
t	Method	mAP	mAP
0		24.15	25.73
1	OSHOT	24.07	25.91
	Ours	**24.71**	**28.12**
2	OSHOT	24.35	26.72
	Ours	**28.71**	**29.87**
3	OSHOT	24.18	27.32
	Ours	**31.62**	**30.68**
4	OSHOT	26.10	27.65
	Ours	**31.97**	**30.25**
5	OSHOT	26.43	27.91
	Ours	**33.84**	**30.46**

Table 5. Object detection results under the OSHOT setting [2] with RainyCityscape and FoggyCityscape datasets. The source dataset is the original Cityscapes dataset.

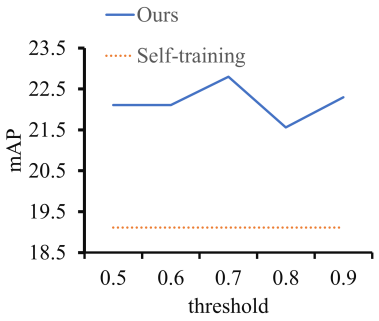
Method	Comic	RainyCityscape
Self-training Baseline	19.11	24.46
+IoU-CI	21.92	29.43
+IoU-OD	21.77	25.75
+IoU-CI + IoU-OD	**22.51**	**32.87**

Table 6. Effectiveness of each IoU-based indicator in our IoU Filter.

the self-training baseline as they could obtain higher-quality pseudo labels. In addition, these two indicators are complementary to each other and their integration leads to the best performance.



(a) IoU-CI Threshold



(b) IoU-OD Threshold

Figure 6. Performance obtained by the proposed IoU Filter (blue solid curves) on the Comic dataset after changing the thresholding values. The orange dashed lines indicate the performance of the self-training baseline.

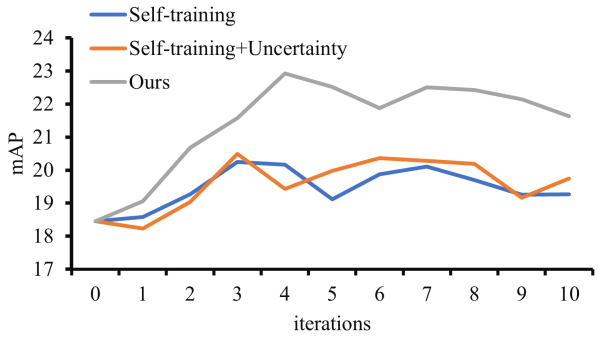


Figure 7. Performance changes as the number of iterations increases. Results are obtained on the Comic2k dataset.

4.3.2 Impact of Different Thresholds

We study the impact of different thresholds on performance, including the IoU-CI threshold and the IoU-OD threshold. We only change one threshold in every experiment and keep the others the same as described in Sec. 4.1.2. The results are displayed in Fig. 6. We can see that though there is some performance perturbation, the proposed method always outperforms the self-training baseline.

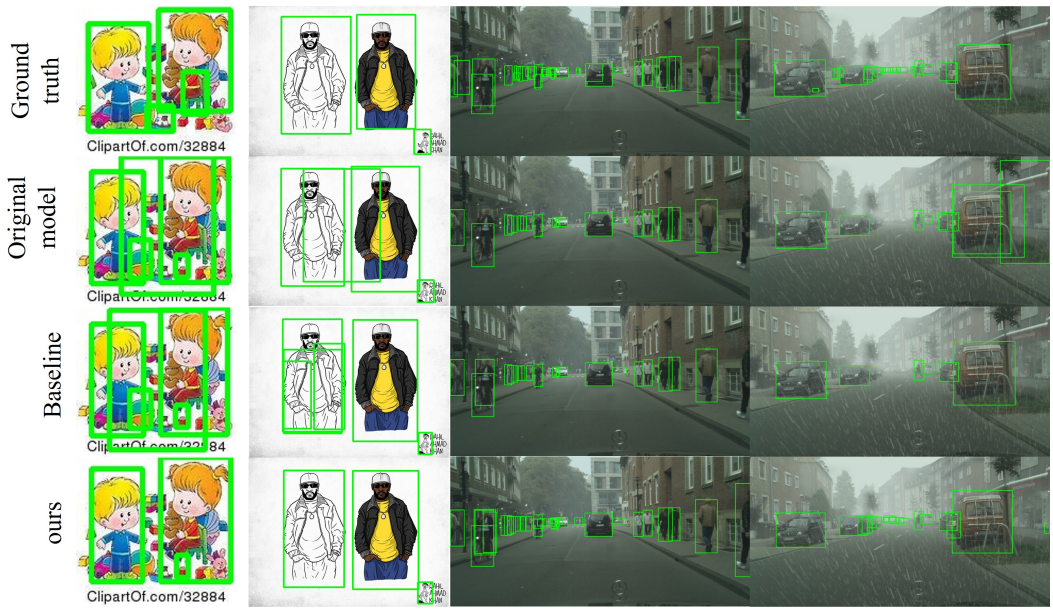


Figure 8. Qualitative results of fully test-time adaptation for object detection. The first row is the ground truth of each image, the second row is the results of the pre-trained detector, the third row is the results of the self-training baseline, and the last row is the results of our proposed method. The object detection results are localized as green boxes.

4.3.3 Self-training Iterations

Fig. 7 shows how the performance of each method changes in different iterations. It delineates that all these methods improve at the first 5 or 6 iterations, but degrade in more iterations and would continue this trend in the future. This could be attributed to two reasons. First, as there is only one testing image to perform adaptation, too many iterations could lead to overfitting. Second, detection errors could accumulate in the pseudo labels and adversely affect the test-time training.

4.4. Qualitative Results

Fig. 8 compares the qualitative results obtained by different methods as well as the original Faster RCNN. The detected targets are localized as green boxes. Incorrect classification and false negatives are the two major problems in the presence of domain shift. Our method could effectively address them. For example, as shown in the fourth column of Fig. 8, our proposed method could effectively detect the missing objects, such as the small and obscured objects.

5. Discussion of Limitation

The major limitation of our method is that the proposed IoU filter could exclude some correct detections from the pseudo labels in addition to the incorrect detections, *e.g.*, the blue box in Fig. 3. It increases the percentage of correct pseudo labels but decreases the absolute number of pseudo labels. Though our method could obviously improve object detection in the presence of domain shift, we believe the performance could be further improved if fewer correct pseudo labels are removed while increasing the quality of pseudo labels.

6. Conclusion

This paper presents the first approach to address fully test-time adaptation for object detection. Compared with current domain adaptive object detectors, it neither assumes a stationary and known target distribution nor requires access to a target dataset, which is desired in many applications. We first investigate a baseline self-training framework but find that its performance is bottlenecked by the low-quality pseudo labels, caused by the domain shift. To overcome this obstacle, we introduce the IoU Filter. It includes two IoU-based indicators that could select higher-quality pseudo labels in the presence of domain shift. Experimental results on three datasets demonstrate that our approach could effectively adapt a trained detector to various kinds of domain shifts at test time and bring substantial performance gains. Through a controlled ablation study, we show that each indicator is effective and they are complementary, the threshold values could impact the performance, and training too many iterations could degrade fully test-time adaptation.

Acknowledgements. This work was supported in part by Wei Tang's startup funds from the University of Illinois Chicago and the National Science Foundation (NSF) award CNS-1828265.

References

[1] Waqar Ahmed, Pietro Morerio, and Vittorio Murino. Cleaning noisy labels by negative ensemble learning for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1616–1625, 2022. 1, 3

[2] Francesco Cappio Borlino, Salvatore Polizzotto, Barbara Caputo, and Tatiana Tommasi. Self-supervision & meta-learning for one-shot unsupervised cross-domain detection. *Computer Vision and Image Understanding*, 223:103549, 2022. 2, 5, 6, 7

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 3

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1, 2

[5] 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, 3

[6] Tong Chu, Yahao Liu, Jinhong Deng, Wen Li, and Lixin Duan. Denoised maximum classifier discrepancy for source-free unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 472–480, 2022. 1

[7] 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. 5

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. 4, 5, 6

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[10] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021. 1

[11] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 3, 5

[12] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019. 3

[13] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2022. 1, 2

[14] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 3

[15] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021. 3

[16] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020. 1

[17] Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 12365–12377. PMLR, 2022. 3

[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. 3

[19] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. *Advances in Neural Information Processing Systems*, 34:22770–22782, 2021. 3

[20] Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *arXiv preprint arXiv:2105.13502*, 2021. 1, 3

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2

[22] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019. 3

[23] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 1, 3

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

[25] Vishwanath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 763–780. Springer, 2020. 1, 3, 5

[26] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 1, 2

[27] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 1

[28] Vibashan VS, Poojan Oza, and Vishal M Patel. Towards online domain adaptive object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 478–488, 2023. 2

[29] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 2, 3

[30] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 2, 5, 6

[31] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9010–9019, 2021. 1

[32] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3520–3529, 2021. 1

[33] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 5, 6

[34] Dan Zhang, Jingjing Li, Lin Xiong, Lan Lin, Mao Ye, and Shangming Yang. Cycle-consistent domain adaptive faster rcnn. *IEEE Access*, 7:123903–123911, 2019. 3