Source-free Domain Adaptation for Video Object Detection Under Adverse Image Conditions

Xingguang Zhang
Purdue University
West Lafayette, In 47907, USA
zhan3275@purdue.edu

Chih-Hsien Chou
Futurewei Technologies, Inc.
Santa Clara CA 95050, USA
cchou@futurewei.com

Abstract

When deploying pre-trained video object detectors in real-world scenarios, the domain gap between training and testing data caused by adverse image conditions often leads to performance degradation. Addressing this issue becomes particularly challenging when only the pre-trained model and degraded videos are available. Although various source-free domain adaptation (SFDA) methods have been proposed for single-frame object detectors, SFDA for video object detection (VOD) remains unexplored. Moreover, most unsupervised domain adaptation works for object detection rely on two-stage detectors, while SFDA for one-stage detectors, which are more vulnerable to fine-tuning, is not well addressed in the literature. In this paper, we propose Spatial-Temporal Alternate Refinement with Mean Teacher (STAR-MT), a simple yet effective SFDA method for VOD. Specifically, we aim to improve the performance of the one-stage VOD method, YOLOV, *under adverse image conditions, including* noise, air turbulence, *and* haze. *Extensive experiments on the ImageNetVOD dataset and its degraded versions demonstrate that our method consistently improves video object detection performance in challenging imaging conditions, showcasing its potential for real-world applications.*

1. Introduction

Object detection in images and videos represents a pivotal task in computer vision, primarily owing to its extensive range of applications across diverse scenarios, such as intelligent surveillance systems and automated driving vehicles. Video object detection (VOD) aims to predict the bounding box and category information of all targeting objects in all video frames. Compared to single-frame object detection tasks, VOD enjoys the advantage of accessing additional information from the temporal dimension [34, 45], which often contains consistent semantics and multiple views of the same target to help enrich the feature space and facilitate superior performance.

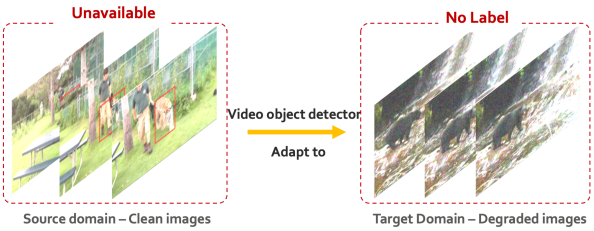


Figure 1. The scope of this work: we aim to adapt the video object detection model trained on clean image sequences to degraded image sequences under the condition that the data from the source domain and ground truth labels of the target domain are unavailable during the adaptation.

When deploying object detectors in real-world settings, adverse imaging conditions caused by rain, haze, low light, or air turbulence often lead to a notable domain gap. These adverse conditions frequently result in a significant reduction in performance, underscoring the necessity for domain adaptation to bridge the gap. Typical domain adaptation (DA) methods require access to data and labels in both source and target domains, with labels in the target domain being relatively scarce [28]. Due to the prohibitive cost of labeling target domain data, unsupervised domain adaptation (UDA) is sometimes required to facilitate effective fine-tuning on the target domain without any labels [27]. In most DA and UDA cases for image detection, source domain data is provided to serve as anchors for adaptation. However, in some UDA scenarios, source data may not be available due to storage or privacy constraints, necessitating the optimization of detection networks under such limitations. To address this challenge, source-free domain adaptation (SFDA) is gaining increasing attention in the field of object detection, as it enables adaptation to target domains without relying on source domain data.

Although multiple approaches, such as those proposed

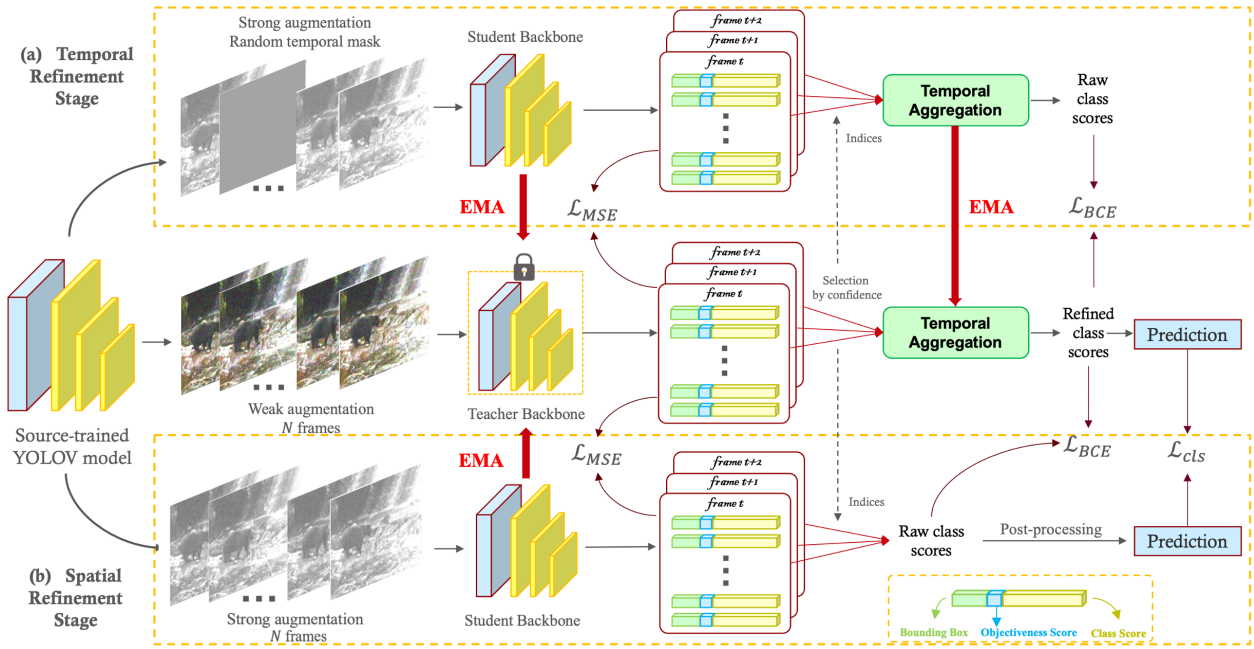


Figure 2. Overview of the proposed STAR-MT for source-free adaptive video object detection. The domain adaptive fine-tuning alternately operates in two stages: (a) Temporal Refinement Stage (TRS) and (b) Spatial Refinement Stage (SRS).

in [3, 7, 17, 18, 20, 23, 36], have been developed recently to address SFDA for object detection, they predominantly focus on two-stage object detectors [30]. To date, no prior work has specifically targeted SFDA for one-stage object detectors [29], as their weights are more sensitive to fine-tuning, and intermediate features are highly abstract. Moreover, no SFDA method has been proposed for the video object detection task, primarily due to the scarcity of cross-domain video object detection datasets. However, given the unsupervised nature of SFDA, it is possible to explore source-free domain adaptation methods for video object detection using synthetic data. The scope of this work is illustrated in Fig. 1. One can expect those methods to be reliably applied to real-world scenarios where source domain data may not be readily available or accessible.

In this paper, we conduct analysis and experiments about basic SFDA techniques for the video object detection task and propose a novel SFDA method for the one-stage video object detector YOLOV [34]. Specifically, we aim to adapt YOLOV [34] to challenging adverse image conditions by alternately fine-tuning the video object detection model in the teacher-student learning framework. We summarize the contributions of this paper as follows:

- We conduct the pioneering study exploring source-free unsupervised domain adaptation for video object detection (VOD). Specifically, we aim to adapt the YOLOV detector to adverse image conditions.
- We introduce a novel SFDA algorithm for VOD, termed Spatial-Temporal Alternate Refinement with Mean Teacher (STAR-MT). It alternately trains in two stages: the Temporal Refinement Stage works in a traditional mean-teacher learning scheme, while the Spatial Refinement Stage leverages temporally enhanced features to guide the single-frame backbone module.
- We demonstrate the effectiveness of our method through experiments on different synthetic adverse image conditions, including noise, air turbulence, and haze. Given its unsupervised nature, STAR-MT is anticipated to yield reliable performance boost for video object detectors in real-world and unseen scenarios.

2. Related works

2.1. Video object detection.

Object detection, one of the most fundamental problems in computer vision, aims to predict the location and class of objects of interest within an input image. Neural networks for object detection can be roughly categorized into one-stage and two-stage detectors [46]. Two-stage detectors, represented by Faster RCNN [30] and FPN [21], predict the feature proposals and detection results in a two-step, coarse-to-fine manner. These detectors generally exhibit good performance and are relatively easy to train. In contrast, one-stage detectors, such as YOLO [29], SSD [24], and DETR [4], predict all detections in a single inference stage, making them faster and easier to deploy in real-world applications. However, training one-stage detectors often requires more tricks, and they may struggle when detecting dense

and small targets.

Video object detection (VOD) presents a unique set of challenges distinct from still image object detection, primarily due to the dynamic nature of video content. In VOD, temporal consistency across frames in the same sequence can be exploited to enhance the robustness of features and reduce the potential ambiguity of the object information in a single image. Given an image sequence $\mathbf{I} \in \mathbb{R}^{H \times W \times T}$ where H, W and T are image height, image width, and temporal length of the sequence. Most recent VOD algorithms predict the object location $\{y^{loc}\}$ and category $\{y^{cls}\}$ by extracting spatial features via a single-frame backbone and temporal aggregation [39]. Those solutions are customized for two-stage image object detectors [14, 30] or transformers [11, 38, 44]. These detectors, while effective, often incur high computational costs due to their model size or complex processing pipeline [18].

In contrast, YOLOV [18] integrates the one-stage object detector YOLOX [12] as its spatial backbone. This configuration benefits from a cost-effective temporal aggregation module, which significantly enhances YOLOX's performance, endowing YOLOV with both superior performance and efficiency. YOLOV's methodology involves selecting key regions from the dense prediction map produced by the detection head, minimizing the processing of numerous low-quality candidates. Furthermore, it assesses the affinity between extracted features from both target and reference frames, facilitating a lightweight feature aggregation process. This strategy presents an efficient alternative to more cumbersome methods, particularly advantageous in scenarios demanding real-time responsiveness.

2.2. Source-free domain adaptation

Domain adaptation for object detection involves data from two domains with different data distributions: the source domain, in which the detector is initially trained, and the target domain, where the detector will be ultimately deployed. Typically, labeling in the target domain is relatively scarce [28]. In addition to supervised domain adaptation, methods for semi-supervised [15, 32] and unsupervised domain adaptation [2, 6, 33, 37] for object detection have been studied based on the availability of labels. These methods have achieved significant success in domain adaptation, regardless of whether labeling is available in the target domain data. However, they all require access to source domain data, which may not always be available due to privacy or storage constraints. To address this issue, methodologies for source-free domain adaptation (SFDA) have been recently proposed [16, 19, 25, 35, 40–42].

SFDA aims to adapt the detector to the target domain using only the pre-trained model and target domain data, without requiring access to the source domain data, making it a promising approach for real-world applications where source data may be unavailable or inaccessible. Initial SFDA strategies have harnessed self-supervised techniques and pseudo-labeling [19]. The mean-teacher method [35] employs a student-teacher paradigm where the teacher model's parameters are an exponential moving average of the student model's parameters. This approach has shown effectiveness in stabilizing training and improving robustness as the teacher model accumulates and refines knowledge over time, aiding in generating more reliable pseudo-labels. Such methods underscore the essence of SFDA: leveraging target domain intrinsic properties while circumventing the need for source data, thereby aligning domain-specific feature distributions.

Mean-teacher has been a fundamental technique in SFDA for object detection. Most existing works [3, 7, 10, 17, 20, 23, 36] employ the mean-teacher as part of their frameworks. Besides mean-teacher, other methodologies include self-entropy descent and pseudo-label refinement [18], style enhancement and graph alignment constraint [17], adversarial alignment [10], instance relation graph [36] and contrastive representation learning [3, 36]. However, most existing algorithms are designed for two-stage detectors, particularly the Faster RCNN [30], and cannot be directly applied to the domain adaptation for one-stage detectors such as the YOLO series. This is partially because the region proposals in two-stage detectors could provide high-quality semantic information for additional feature alignment, providing meaningful additional training signals for SFDA. Another reason is that one-stage detectors usually need complicated training tricks; their feature space is more intractable and vulnerable to fine-tuning. Recently, YOLOV [34] provided an efficient feature selection and fusion mechanism for the one-stage detector YOLOX [12] among multiple frames, SFDA for YOLOV would provide valuable experience in both domain adaptation for one-stage detectors and VOD tasks.

3. Method

In this section, we detail our STAR-MT method and the domain adaptation benchmark. The overall scheme of the proposed solution is illustrated in Fig.2.

3.1. Mean-teacher for domain adaptive VOD

In developing our method, we leverage the advanced unsupervised domain adaptation strategies found in the mean-teacher self-training approach [35]. We introduce the implementation of this method in this paradigm.

As a class of student-teacher training approach, the mean-teacher method keeps two identical networks: the student network and the teacher network. They are initialized by the weights trained on the source domain. During training, the weights of the teacher model are fixed, while the student model is trained with the supervision signal from the

prediction output and features generated from the teacher model. On the other hand, the teacher model takes the exponential moving average (EMA) of consecutive student models for its parameter update:

$$\theta_{\mathcal{T}}^{t} \leftarrow \alpha\theta_{\mathcal{T}}^{t-1} + (1-\alpha)\theta_{\mathcal{S}}^{t-1}, \tag{1}$$

where the $\theta_{\mathcal{T}}$ and $\theta_{\mathcal{S}}$ denote the weights of teacher and student models, t denotes the training iteration, and $\alpha \in (0, 1)$ is the momentum coefficient which is usually set close to 1 for a smooth temporal ensemble [3].

3.2. Spatial-Temporal Alternate Refinement

YOLOV utilized the pre-trained backbone of YOLOX as its frame-wise feature extractor, followed by feature selection and affinity measurement that identifies features from the same object among frames to guide temporal aggregation. However, training the spatial backbone and temporal aggregation module simultaneously on the video object detection dataset is suboptimal because they require different training schemes. Hence, we propose to adapt the YOLOV in a two-stage alternate optimization manner, consisting of the temporal refinement stage (TRS) and spatial refinement stage (SRS).

3.2.1 Temporal Refinement Stage (TRS).

In the TRS, the entire teacher model, including the frame-wise backbone and temporal aggregation module, is updated via EMA. In the beginning, both teacher and student models are initialized the same. Like a typical mean-teacher-based algorithm, the same image sequences with different augmentations are fed into those models. The teacher model processes the weakly augmented images, and the heavily augmented images are fed into the student model. Moreover, we randomly mask out $r\%$ frames and enforce the student model to produce the same output with fewer frames than the teacher model. This masking mechanism can supposedly enhance the generalization capability of temporal aggregation. The student model is trained by aligning frame-wise features and soft pseudo labels with the features and predictions of the teacher model. The loss in this stage is defined as:

$$\mathcal{L} = \mathcal{L}_{MSE}(f_{\mathcal{T}}, f_{\mathcal{S}}) + \mathcal{L}_{BCE}(y_{\mathcal{T}}^{cls}, y_{\mathcal{S}}^{cls}), \tag{2}$$

where the first term is the mean square error between the feature maps $f_{\mathcal{T}}$ and $f_{\mathcal{S}}$, produced by the backbone module of the teacher and student models, respectively. The term $\mathcal{L}_{BCE}$ denotes the binary cross entropy loss. $y_{\mathcal{T}}^{cls}$ refers to the top-k classification prediction after the temporal aggregation of the teacher model, and $y_{\mathcal{S}}^{cls}$ refers to that of the student model. k is the number of proposals in the feature selection module before the temporal aggregation. We set $k = 30$ following the default setting of YOLOV. We do not particularly compute the loss of objectiveness and bounding box prediction because they are unchanged in the temporal aggregation module.

3.2.2 Spatial Refinement Stage (SRS).

TAM consists of two key components: a Feature Selection Module, which selects high-quality prediction proposals, and a Feature Aggregation Module, which fuses these proposals across multiple frames. However, due to the inconsistency between the training pipelines of the single-frame detection head (backbone) and the TAM, the TRS, which mostly follows the training setting of the TAM, may lead to suboptimal adaptation on the backbone side. Recognizing that the TAM can reliably improve prediction quality, we propose using the output class score of YOLOV, instead of YOLOX, in the teacher model as higher-quality pseudo labels to guide the fine-tuning of the detection head of YOLOX in the student model. In the SRS, only the backbone of the teacher model is updated via EMA, ensuring that the adaptation focuses on the spatial feature extraction process while leveraging the enhanced temporal information from the TAM. The loss is given as follows:

$$\mathcal{L} = \mathcal{L}_{MSE}(f_{\mathcal{T}}, f_{\mathcal{S}}) + \mathcal{L}_{BCE}(y_{\mathcal{T}}^{cls}, y_{\mathcal{S}}^{cls}) + \gamma\mathcal{L}_{cls}, \tag{3}$$

where γ is the weighting factor. The new loss term $\mathcal{L}_{cls}$ is the certainty-aware binary cross entropy loss between the filtered class score from the teacher and student model:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i}^{N} p_{\mathcal{S}}^{i}\Big[\frac{1}{n_c}\sum_{c}^{n_c}\Big(s_{\mathcal{T}}^{i,c}\log(s_{\mathcal{S}}^{i,c}) + (1-s_{\mathcal{T}}^{i,c})\log(1-s_{\mathcal{S}}^{i,c})\Big)\Big], \tag{4}$$

where c is the index of the category, $n_c = 30$ is the number of classes, and i and N are the index and number of detected objects in the sequence. $s_{\mathcal{S}}^{i,c}$ and $s_{\mathcal{T}}^{i,c}$ are the i-th output scores of class c for the student and teacher models, respectively. $p_{\mathcal{S}}^{i} \in (0, 1)$ is the normalized objectiveness score in the student model output, serving as the weight of the pseudo-label. It can be viewed as the certainty measurement of the object's existence; the greater $p_{\mathcal{S}}^{i}$ indicates the higher confidence of the particular pseudo label.

3.2.3 Alternate Refinement.

STAR-MT training is periodical, with the TRS and SRS having identical iterations τ in each period. Given k the index of the period, TRS is executed in iterations $[2k\tau, 2k\tau + \tau)$ and SRS in iterations $[2k\tau+\tau, 2k\tau+2\tau)$. During the experiment, it was observed that the order of those two stages only had a trivial impact on the overall performance.

Although early stopping is not explicitly implemented in our approach, we utilize the mean self-entropy [18] of the class score from the teacher model as a performance criterion for all output checkpoints. This mean self-entropy, denoted as H, serves as a measure of reliability for pseudo labels; a lower H indicates greater confidence in the teacher model in guiding the student. The checkpoint corresponding to the minimal value of H is selected as our output model. The formula to compute H is as follows:

$$H = -\frac{1}{Nn_c}\sum_{i}^{N}\sum_{c}^{n_c} s_{\mathcal{T}}^{i,c}\log(s_{\mathcal{T}}^{i,c}). \tag{5}$$

4. Experiment

4.1. Adverse image condition synthesis

Real-world VOD often faces the challenge of domain gaps caused by adverse image conditions. Because of the complexity of image degradation, testing domain adaptation algorithms under various conditions is desired. However, appropriate datasets for testing the domain adaptation algorithm for VOD models trained on the ImageNetVOD dataset are unavailable. In this work, we synthesized videos in three common imaging conditions: noise, air turbulence, and haze. Each video has its distinct degradation parameter, with the profile varies temporally. A sample image sequence from the original dataset and the associated three degraded sequences are shown in Fig. 3. The unsupervised property of the algorithm guarantees that our method will be effective in real-world unknown degradations. The simulation of the three adverse image conditions is described below:

Noise. Noise is the predominant degradation in the low-light conditions. The noise in our experiment is modeled with:

$$\widetilde{I}(h,w,t) = I(h,w,t) + n(h,w,t), \tag{6}$$

where I is the input image sequence, $\widetilde{I}$ is the degraded image sequence, h, w, and t are the sequence's height, width, and frame indices. $n(h,w,t) \sim \mathcal{N}(0,\sigma^2)$ is the Gaussian noise. We randomly sample the variance of the noise in $\sigma^2 \in [10/255, 50/255]$. Each sequence has its distinct variance.

Air Turbulence. In long-range imaging conditions, air turbulence may affect the performance of computer vision models significantly [22]. The air turbulence primarily causes random pixel displacement and spatially varying blur on the image [5]. We utilized the popular P2S simulator [8, 26] to synthesize the degraded video:

$$\widetilde{I}(h,w,t) = \text{P2S}(I(h,w,t); n(h,w,t)). \tag{7}$$

The P2S simulator converts the Gaussian random seed n to spatially varying pixel displacement and blur. Inspired by [9, 43], we applied the temporal correlation and varying kernel size to improve the diversity and fidelity of the synthetic turbulence. Each image sequence has a distinct turbulence strength and profile.

Haze. Haze is a crucial adverse image condition in VOD application scenarios, especially in surveillance systems and automated driving. The hazy video can be modeled with the transmission function [1]:

$$\widetilde{I}(h,w,t) = I(h,w,t)e^{-\beta d(h,w,t)} + A(1 - e^{-\beta d(h,w,t)}), \tag{8}$$

where $e^{-\beta d(h,w,t)}$ is the transmission rate, β is the scattering coefficient, $A = 255$ is the maximum intensity of a pixel, and $d(h,w,t)$ is the relative depth value measured by [13]. Like other degradations, each sequence has its own scattering coefficient randomly sampled from a uniform distribution $\beta \in [0.5, 1.5]$.

4.2. Dataset and baselines

In Section 4.1, for each degradation type identified, we generated a corresponding synthetic target-domain dataset utilizing the ImageNetVID dataset [31] as the source. Comprising 30 classes set against diverse natural backdrops, ImageNetVID provides over 1 million training frames and more than 100,000 validation frames. We used all frames from this source dataset to synthesize the target-domain datasets. Consequently, each domain can retain the same set of labels. Fig 4 shows a snippet of the testing set of our synthetic degradations, along with the visualization of the detection results before and after the domain adaptation. The visual comparison proves the efficacy of the proposed method.

Following its publicly available codebase, we trained the YOLOV in all three scales — small (S), large (L), and extra-large (X) — using the source dataset. In our experiment, the post-processing method was omitted as it does not pertain to our algorithms. These models were then directly tested on target domains, with the findings detailed in Table 1. The Average Precision at 50% threshold (AP50) on the source domain is registered as 77.3%, 83.6%, and 85.0% for YOLOV-S, YOLOV-L, and YOLOV-X, respectively. The significant degradation in performance, when we test the source-train model on the target domain dataset, indicates that challenging image conditions markedly reduce the performance of the VOD models.

In addition to the initial training, we used a supervised approach to fine-tune the source-trained YOLOV models on the target domain datasets. This supervised fine-tuning serves as a theoretical benchmark for the upper limit of performance achievable through unsupervised adaptation. Deviating from the original pipeline, which involves training the base detector prior to the temporal aggregation module, we discovered that directly fine-tuning the temporal aggregation module leads to improved outcomes. Therefore, our

Figure 3. A snippet of the ImageNetVOD dataset and three forms of degradation. The original frames are from the testing video *ILSVRC2015_test_00028000.mp4* and $t = 32$.

Degradation	Noise			Air Turbulence			Haze		
Model	YOLOV-S	YOLOV-L	YOLOV-X	YOLOV-S	YOLOV-L	YOLOV-X	YOLOV-S	YOLOV-L	YOLOV-X
Source-only	38.0	57.0	60.4	63.2	72.7	73.9	57.2	70.7	73.0
PL *w.* SE [18]	47.8	61.4	62.3	64.3	73.2	74.1	61.1	73.2	74.5
Basic MT	54.8	70.3	70.6	64.1	74.2	74.4	64.7	75.3	76.9
STAR-MT	**57.6**	**71.4**	**71.5**	**65.2**	**75.0**	**75.7**	**68.1**	**78.0**	**78.9**
Oracle	61.0	72.5	72.7	66.7	76.4	78.3	69.7	79.6	80.2

Table 1. Performance comparison on AP50(%). The larger, the better. "PL" refers to the pseudo-label method, and "Source-only" refers to the models trained by only using labeled source domain data.

fine-tuning focuses solely on the temporal aggregation module for the target domains, as detailed in [34]. The results of this supervised fine-tuning, labeled as "oracle", are also presented in Table 1.

Before implementing the mean-teacher-based methods, we conducted a preliminary experiment with the pseudo-label (PL) algorithm. In this approach, models trained on the source domain are employed to process all videos in the target domain's training set, generating initial predictions. They are then filtered by threshold 0.5 on the product of objectiveness and the maximal class scores to generate pseudo labels. Since fine-tuning the single-frame backbone always causes catastrophic failure, we fixed the parameters in the backbone module and only trained the temporal aggregation module with pseudo labels. After training, we utilized the self-entropy [18] as the indicator to select the potential best model. The result is also demonstrated in Table 1.

4.3. Implementation details of STAR-MT

Adhering closely to the YOLOV codebase, we maintained most of the original settings unaltered. For hyperparameter configuration, we empirically set the teacher model's smoothing coefficient, α, to 0.9995 and the weighting factor, γ, of the $\mathcal{L}_{cls}$ to 0.2. The model training was executed using Stochastic Gradient Descent (SGD) with a batch size of 1 over 10,000 iterations. We initialized the learning rate at 2×10^{-4} and applied a cosine annealing scheduler, tapering it down to 1×10^{-4}. In the evaluation phase, only the teacher model was utilized for inference. The mean Average Precision (mAP) was calculated with an IoU threshold of 0.5. All experiments are conducted on NVIDIA 3080 Ti and V100 GPUs.

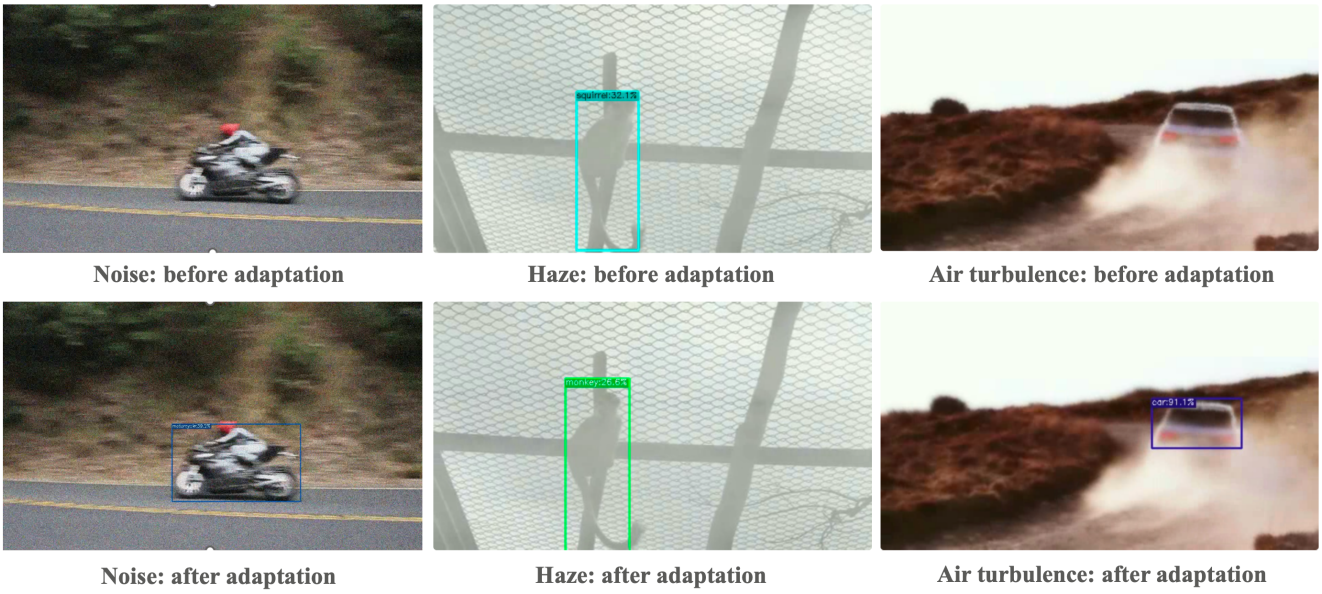
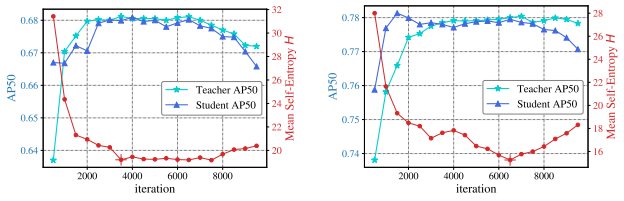


Figure 4. Visual comparison before and after the SFDA by STAR-MT. All experiments are conducted with YOLOV-S.



(a) YOLOV-S (b) YOLOV-L

Figure 5. The teacher model's AP50 and mean self-entropy H variation in the STAR-MT training of YOLOV-S and YOLOV-L. Both experiments are conducted on clean $\rightarrow$ haze. The H indicating the best teacher model are marked in the figures with "+".

For each sequence in our domain adaptation experiments, 32 frames are loaded. Mosaic augmentation was disabled for all these experiments. However, we have retained both random flip and perspective transformations, applying these consistently to both weakly and strongly augmented sequence pairs. The key distinction between weak and strong augmentation lies in the strength of random chromatic transformation. Random erasing is involved only in the strong augmentation. In the TRS, random masking is applied to restrict the temporal information the student model can access, compelling it to enhance the temporal aggregation capability. The masking rate is $r\%$ where r is randomly sampled from $[0, 75]$.

The performance of the STAR-MT method is demonstrated in Table 1. Our method shows a significant improvement in the SFDA for VOD under all three adverse image conditions. It also demonstrates a clear advantage over conventional methods like pseudo-labeling and basic mean-teacher learning. Notably, although the method seems straightforward and not complicated, *the performance of our method closely approaches that of supervised fine-tuning*. To illustrate the correlation between mean self-entropy H and the performance of SFDA, we drew the variations in model performance alongside the change in H value on the evaluation set, as shown in Fig 5. The H values were calculated using a sliding window average over 100 iterations. We can observe the lowest value of H aligns well with the peak performance of the model.

4.4. Ablation study

Efficacy of alternate refinement. One major novelty in this paper is the spatial refinement stage, as an alternately updated module in addition to the normal mean teacher learning framework. The key insights behind this are 1) temporally enhanced features of the teacher model can be used to generate reliable pseudo labels for the training of the single-frame detection head, and 2) training the single-frame detection head under the YOLOV setting is suboptimal. Thus, it needs additional guidance. From the comparison between the basic mean-teacher method (TRS only) and the proposed STAR-MT in Table 1, we can verify the efficacy of alternate refinement with the spatial refinement stage. To demonstrate that the pseudo labels from the YOLOV are of higher quality, we conducted source-free domain adaptation for the YOLOX. We follow the mean-teacher framework and use the pseudo labels generated by YOLOX and YOLOV to guide the training of the student network. The result is shown in Table 2. The model guided

Model	YOLOX-S	YOLOX-L
Source-only	35.9	56.6
PL guided by YOLOX	49.2	61.3
PL guided by YOLOV	51.0	62.9
Oracle	56.7	66.0

Table 2. The efficacy of YOLOV as the teacher model for the SFDA of the single-frame detection backbone. All experiments are conducted on clean → noise. The metric is AP50(%), the larger, the better.

$\mathcal{L}_{MSE}$	$\mathcal{L}_{BCE}$	$\mathcal{L}_{cls}$	YOLOV-S	YOLOV-L
✗	✓	✓	62.1	71.5
✓	✓	✗	67.6	77.4
✓	✗	✓	67.8	77.3
✓	✓	✓	**68.1**	**78.0**

Table 3. The efficacy of losses. All experiments are conducted on clean → haze. The metric is AP50(%), the larger, the better.

Model	$\tau = 50$	$\tau = 100$	$\tau = 200$	$\tau = 500$
YOLOV-S	68.0	**68.1**	67.6	67.8
YOLOV-L	76.9	77.5	**78.0**	77.7
YOLOV-X	78.2	78.4	**78.9**	78.8

Table 4. The impact of the number of iterations in each stage. All experiments are conducted on clean → haze. The metric is AP50(%), the larger, the better.

by the temporally refined labels in YOLOV gets better performance, which provides evidence for the efficacy of SRS.

Efficacy of losses. In our study, the three utilized loss functions are categorized as feature alignment loss $\mathcal{L}_{MSE}$ and pseudo-label based losses ($\mathcal{L}_{BCE}$ and $\mathcal{L}_{cls}$). We experimented with various reasonable combinations of these loss terms to assess their impact. In all combinations, at least one pseudo-label-based loss was maintained. The results are detailed in Table 3. Initially, we excluded the feature alignment loss $\mathcal{L}_{MSE}$ and observed a significant decline in adaptation performance. This indicates the model's high sensitivity to label quality and the importance of restricting the feature space generated by the detection head. Further, we excluded $\mathcal{L}_{MSE}$ and $\mathcal{L}_{cls}$ separately to evaluate their individual contributions. The results confirmed the effectiveness of both losses.

Number of iterations for each stage. We also evaluated the optimal number of fine-tuning iterations, τ, for each stage within a period. For this purpose, we conducted experiments that set τ to various values: 50, 100, 200, and 500, while maintaining 10,000 training iterations. The determination of optimal results within this range was based on the mean self-entropy H values, where the teacher model associated with the first local minima of H is selected. This assessment was carried out across all three model scales, and the findings are presented in Table 4. The results indicate that different scales of the model may require distinct hyperparameters to achieve optimal adaptation performance.

5. Conclusion

In this paper, we propose a pioneering approach to explore the source-free domain adaptation (SFDA) for video object detection (VOD). Specifically, we developed a novel SFDA method for a one-stage-based detector, YOLOV. The proposed STAR-MT technique significantly improves the performance of the video object detector in adverse image conditions without access to the target domain label or source domain data. Owing to its unsupervised nature, this work can be seamlessly applied to real-world scenarios requiring VOD models. The proposed method could serve as a baseline for future research in unsupervised domain adaptation for video object detection.

6. Acknowledgment

Futurewei Technologies Inc. funded this research while the first author was an intern there. We are grateful to the company's IC Lab for the research assistance.

References

[1] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 5

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 3

[3] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23839–23848, 2023. 2, 3, 4

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[5] Stanley H Chan and Nicholas Chimitt. Computational imaging through atmospheric turbulence. *Foundations and Trends® in Computer Graphics and Vision*, 15(4):253–508, 2023. 5

[6] 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. 3

[7] Zhihong Chen, Zilei Wang, and Yixin Zhang. Exploiting low-confidence pseudo-labels for source-free object detec-

tion. In *Proceedings of the ACM International Conference on Multimedia*, pages 5370–5379, 2023. 2, 3

[8] Nicholas Chimitt and Stanley H Chan. Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated zernike coefficients. *Optical Engineering*, 59(8): 83101–83101, 2020. 5

[9] Nicholas Chimitt, Xingguang Zhang, Zhiyuan Mao, and Stanley H Chan. Real-time dense field phase-to-space simulation of imaging through atmospheric turbulence. *IEEE Transactions on Computational Imaging*, 8:1159–1169, 2022. 5

[10] 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. 3

[11] Masato Fujitake and Akihiro Sugimoto. Video sparse transformer with attention-guided memory for video object detection. *IEEE Access*, 10:65886–65900, 2022. 3

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv:2107.08430*, 2021. 3

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 3828–3838, 2019. 5

[14] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal ROI align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AAAI)*, pages 1442–1450, 2021. 3

[15] 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

[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. 3

[17] 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 (CVPR)*, pages 8014–8023, 2022. 2, 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 and Interactive Digital Entertainment (AAAI)*, 2021. 2, 3, 5, 6

[19] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 3

[20] L. Lin, Z. Yang, Q. Liu, Y. Yu, and Q. Lin. Run and chase: Towards accurate source-free domain adaptive object detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 2453–2458. IEEE Computer Society. 2, 3

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[22] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, Yiyang Su, Pegah Varghaei, Kai Wang, Xingguang Zhang, Stanley Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, and Xiaoming Liu. Farsight: A physics-driven whole-body biometric system at large distance and altitude. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6227–6236, 2024. 5

[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 (ICCV)*, pages 6414–6424, 2023. 2, 3

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016. 2

[25] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 3

[26] Zhiyuan Mao, Nicholas Chimitt, and Stanley H. Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14759–14768, 2021. 5

[27] Poojan Oza, Vishwanath A Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1 – 24, 2023. 1

[28] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 1, 3

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[30] 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*, 2015. 2, 3

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 5

[32] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058, 2019. 3

[33] 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. 3

[34] Yuheng Shi, Naiyan Wang, and Xiaojie Guo. Yolov: making still image object detectors great at video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AAAI)*, pages 2254–2262, 2023. 1, 2, 3, 6

[35] 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. 3

[36] VS Vibashan, 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 (CVPR)*, pages 3520–3530, 2023. 2, 3

[37] 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. 3

[38] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In *European Conference on Computer Vision (ECCV)*, pages 732–747, 2022. 3

[39] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9217–9225, 2019. 3

[40] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *European Conference on Computer Vision*, pages 147–164. Springer, 2022. 3

[41] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8978–8987, 2021.

[42] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022. 3

[43] Xingguang Zhang, Zhiyuan Mao, Nicholas Chimitt, and Stanley H. Chan. Imaging through the atmosphere using turbulence mitigation transformer. *IEEE Transactions on Computational Imaging*, 10:115–128, 2024. 5

[44] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 7853 – 7869, 2022. 3

[45] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 408–417, 2017. 1

[46] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 2