DESA-TTA: DYNAMIC EMA AND SOURCE ANCHORING FOR TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language object detectors (VLODs) achieve strong zero-shot performance but remain vulnerable to distribution shifts during deployment. Mean-teacher methods for test-time adaptation (TTA) can improve robustness by updating a student model using teacher-generated pseudo-labels. However, mean-teacher TTA is highly sensitive to the choice of a fixed exponential moving average (EMA) coefficient for teacher updates, and repeated optimization with noisy pseudo-labels can cause cumulative student drift. We propose Dynamic EMA and Source Anchoring for TTA (DESA-TTA), a low-overhead method that jointly regulates teacher updates and student drift through dynamic temporal averaging and source anchoring. Dynamic temporal averaging estimates teacher uncertainty from pseudo-label confidence and box density and uses it to select a sample-wise EMA coefficient within bounds determined by teacher parameter drift. Source anchoring partially restores the updated student parameters toward their pretrained values, with the anchoring strength increasing according to student drift. Experiments across diverse distribution shifts and two VLOD architectures show consistent improvements over existing TTA methods. On VOC-C, DESA-TTA improves AP_{50} by 14.5 points over zero-shot inference while achieving 55% higher inference throughput than the previous state-of-the-art TTA method for YOLO-World. Our code is included in the supplementary material and will be made publicly available.

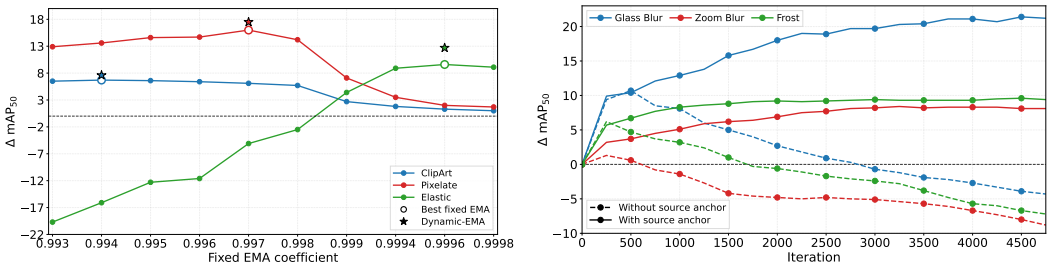
1 INTRODUCTION

VLODs enable open-vocabulary object detection by aligning visual regions with textual class descriptions (Liu et al., 2024a; Cheng et al., 2024). Unlike conventional closed-set detectors (Ren et al., 2015), VLODs can recognize categories specified at inference time, making them particularly attractive for deployment in dynamic real-world environments. However, their strong zero-shot (ZS) performance can degrade substantially when the deployment distribution differs from the pretraining distribution. To address this limitation, TTA has gained increasing attention, as it enables a pretrained model to adapt online during deployment using only unlabeled test data.

State-of-the-art TTA approaches for VLODs primarily follow two paradigms: entropy minimization (Belal et al., 2026) and mean-teacher adaptation (Gao et al., 2025). The entropy-minimization approach performs episodic adaptation by resetting the model parameters after each test batch (Belal et al., 2026). Although this prevents erroneous updates from accumulating, it also limits the model's ability to retain useful knowledge from previous samples (Niu et al., 2023). Mean-teacher adaptation instead enables cumulative adaptation by maintaining a temporally averaged teacher that generates pseudo-labels for updating a student model. Given its ability to accumulate target-specific knowledge across the test stream, we focus on mean-teacher adaptation in this work.

Despite its ability to accumulate knowledge over a test stream, mean-teacher adaptation has two key limitations for TTA of VLODs. First, the teacher relies on a fixed EMA coefficient to balance the preservation of its previous state against the incorporation of the adapted student state. As shown in Figure 1(a), the preferred coefficient varies substantially across domains. A smaller coefficient incorporates recent student updates more rapidly, enabling faster adaptation but increasing sensitivity to errors induced by noisy pseudo-labels. Conversely, a larger coefficient provides greater stability by preserving the previous teacher state but may make adaptation overly conservative. Because the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Sensitivity to the fixed EMA coefficient

(b) Effect of source anchoring over long streams

Figure 1: **Motivation for DESA-TTA. (a)** The performance gain from mean-teacher TTA is highly sensitive to the fixed EMA coefficient, whose best-performing value varies substantially across domains. Stars denote dynamic temporal averaging and are placed at the best fixed-EMA coefficient only for visual comparison. **(b)** Without source anchoring, the initial adaptation gains diminish as pseudo-label updates accumulate, whereas source anchoring preserves these gains over long target streams.

target domain is unknown during deployment, selecting a suitable domain-specific coefficient is impractical. Moreover, a fixed coefficient cannot account for variations in pseudo-label reliability or model state over the course of adaptation.

Second, the student is updated cumulatively using pseudo-labels generated throughout the target stream. Since these pseudo-labels can be noisy or miscalibrated, repeated optimization can lead to error accumulation (Wang et al., 2022). As each adapted student state is subsequently incorporated into the teacher, the accumulated errors can affect the pseudo-labels generated for later samples. As shown in Figure 1(b), adaptation without source anchoring initially improves performance but subsequently deteriorates as more target samples are processed. This behavior indicates that repeated pseudo-label optimization can cause the student to drift progressively away from the pretrained ZS detector, resulting in long-term performance degradation.

To address these limitations, we propose **DESA-TTA**: **D**ynamic **EMA** and **S**ource **A**nchoring for TTA of VLODs. DESA-TTA jointly regulates temporal averaging in the teacher and constrains student drift during cumulative adaptation. To reduce sensitivity to a fixed EMA coefficient, DESA-TTA dynamically adjusts the teacher update coefficient according to teacher uncertainty and teacher parameter drift. Teacher uncertainty is estimated from the mean pseudo-label confidence and normalized box density, while teacher drift determines the admissible EMA range. Dense, low-confidence predictions produce greater uncertainty and therefore a smaller coefficient, allowing the teacher to incorporate the adapted student state more rapidly. Conversely, lower uncertainty produces stronger temporal averaging to preserve teacher stability. As shown in Figure 1(a), dynamic temporal averaging outperforms the best domain-specific fixed-EMA setting on all three displayed domains without requiring prior knowledge of the target distribution.

To mitigate cumulative student drift, DESA-TTA introduces source anchoring after each pseudo-label-based gradient update. The pretrained ZS detector provides a fixed reference, while the adapted student captures target-specific information from the test stream. After each update, the student parameters are partially restored toward their source initialization by linearly interpolating the adapted parameters with their pretrained values using a drift-dependent anchoring coefficient. The anchoring remains weak when the student is close to its initialization and progressively strengthens as student drift increases. This mechanism preserves the student's capacity to acquire target-specific information while constraining excessive deviation from the pretrained detector. As illustrated in Figure 1(b), source anchoring preserves the initial adaptation gains and substantially mitigates the long-term degradation observed without anchoring.

The two components address complementary failure modes. Source anchoring constrains the adapted student state before it is incorporated into the teacher, while dynamic temporal averaging controls the extent of this incorporation. Together, they enable cumulative acquisition of target-specific knowledge while limiting error propagation from repeated pseudo-label optimization.

The contributions of this paper are summarized as follows. (1) We identify and empirically characterize two key limitations of mean-teacher TTA for VLODs: the domain sensitivity of fixed EMA updates and cumulative student drift caused by repeated pseudo-label optimization. **(2)** We intro-

duce DESA-TTA, an efficient method that combines sample-wise dynamic temporal averaging with drift-dependent source anchoring to jointly regulate teacher updates and student drift. **(3)** Extensive experiments on benchmarks covering style shifts, autonomous-driving scenarios, illumination shifts, and common corruptions show that DESA-TTA consistently outperforms state-of-the-art TTA methods across diverse distribution shifts.

2 Related Work

Vision-Language Object Detection. VLODs extend conventional closed-set detectors by aligning visual regions with textual representations, enabling detection beyond a fixed training vocabulary. Early open-vocabulary detection methods leverage image-text pretraining to transfer semantic knowledge from pretrained vision-language models (Radford et al., 2021) to region-level recognition by distilling CLIP representations into detection heads (Gu et al., 2022; Zhong et al., 2022), expanding the detector vocabulary with image-level supervision (Zhou et al., 2022), or adapting vision-language transformers for open-vocabulary localization (Minderer et al., 2022). Grounding-based models unify object detection and phrase grounding by learning region-text alignment from large-scale grounding data (Li et al., 2022; Liu et al., 2024a). More recently, YOLO-World introduces an efficient real-time VLOD by integrating region-text contrastive learning into the YOLO family (Cheng et al., 2024). Despite their strong ZS detection performance, these detectors remain vulnerable to distribution shifts. In this paper, we study the online adaptation of pretrained VLODs using unlabeled target data.

Test-Time Adaptation for Object Detection. TTA for object detection (OD-TTA) aims to improve detector robustness by updating the model on unlabeled target images during inference. Prior OD-TTA methods use EMA-based pseudo-label self-training with feature-alignment regularization (Chen et al., 2023), IoU-based pseudo-label filtering for fully test-time adaptation (Ruan & Tang, 2024), and multi-level feature alignment during online adaptation (Liu et al., 2024b). Cao et al. (2025) further study continual OD-TTA under changing target domains. These methods focus on conventional closed-set detectors and generally assume a shared category space between the source and target domains, limiting their applicability to VLODs.

For VLODs, VLOD-TTA (Belal et al., 2026) introduces a detection-aware entropy objective based on dense proposal overlap but does not explicitly accumulate adaptation over a target stream. TTAOD-F (Gao et al., 2025) adopts a mean-teacher framework with multi-modal prompt adaptation and instance dynamic memory to improve pseudo-label quality but relies on an additional foundation model, increasing inference cost. In contrast, DESA-TTA enables low-overhead cumulative mean-teacher adaptation by dynamically regulating teacher updates and constraining the adapted model toward its pretrained initialization.

Teacher Stability and Drift Control in TTA. The mean-teacher framework maintains an EMA teacher whose predictions provide more stable supervision than those of the online student. Prior TTA methods improve this framework through robust consistency learning (Döbler et al., 2023), timeliness-aware sample reweighting (Yuan et al., 2023), and parameter-selective teacher updates (Tian & Lyu, 2024). However, these methods use a fixed EMA coefficient and therefore cannot adjust the teacher update rate to variations in teacher prediction reliability over the course of adaptation. Moreover, repeated optimization using noisy pseudo-labels can progressively move the student away from its pretrained initialization. To mitigate this drift, CoTTA randomly restores parameters to their source values (Wang et al., 2022), Brahma & Rai (2023) selectively restore parameters according to their estimated importance, Niu et al. (2022) regularize source-important parameters, and Song et al. (2023) constrain adapted features using a frozen source model. These methods are primarily developed for classification and rely on stochastic restoration, parameter-importance estimation, or additional regularization objectives. Chakrabarty et al. (2023) anchors classification adaptation through source-prediction distillation, while Kim et al. (2025) adjusts EMA momentum using prediction divergence. He et al. (2025) divide detector parameters into two groups and update them using two predefined, fixed EMA coefficients. In contrast, DESA-TTA anchors detector parameters through drift-dependent interpolation and selects EMA using detection confidence and box density within drift-dependent bounds.

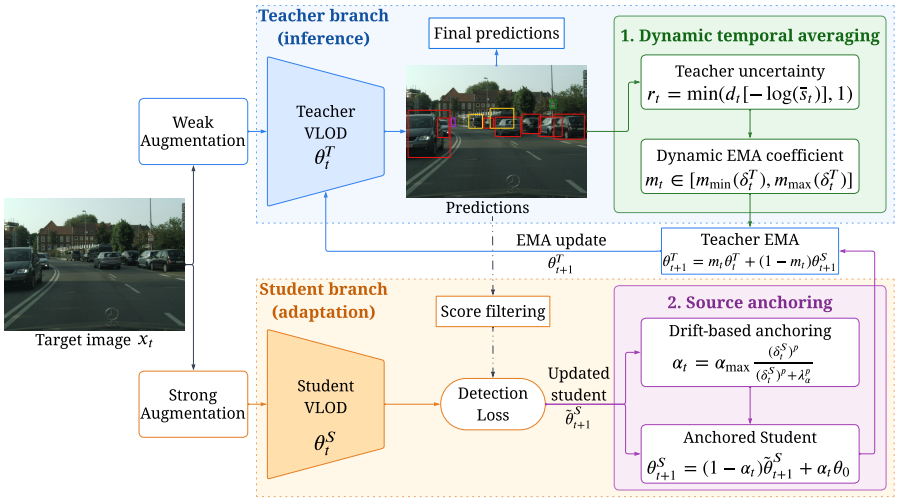


Figure 2: **Overview of DESA-TTA.** Given an unlabeled target image, the teacher processes a weakly augmented view to generate pseudo-labels, while the student is optimized using a strongly augmented view. **(1) Dynamic temporal averaging** computes a teacher-uncertainty score from the mean pseudo-label confidence and box density, and uses it to select m_t within a teacher-drift-dependent EMA range. **(2) Source anchoring** partially restores the gradient-updated student parameters $\tilde{\theta}_{t+1}^S$ toward the pretrained source parameters θ_0 using the drift-dependent coefficient α_t, producing the anchored student θ_{t+1}^S. The anchored student is subsequently used to update the teacher, whose predictions are used for inference.

3 Proposed Methodology

3.1 Preliminaries

Vision-Language Object Detector. Let $\mathcal{C} = \{c_k\}_{k=1}^K$ denote the target vocabulary specified by textual class names. A VLOD comprises an image encoder, a text encoder, and a detection head. Given an input image x, the image encoder extracts visual features, and the detection head predicts candidate boxes $\{b_i\}_{i=1}^N$. The text encoder maps each class name c_k to a text embedding t_k, producing the set $\{t_k\}_{k=1}^K$. Each candidate region is represented by a visual embedding v_i, with class-wise scores computed as:

$$s_{i,k} = \sigma\left(v_i^\top t_k\right), \quad k = 1, \ldots, K, \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. After confidence filtering and non-maximum suppression, the detector outputs the final prediction set:

$$\mathcal{D}_\theta(x, \mathcal{C}) = \{(b_j, y_j, s_j)\}_{j=1}^M, \tag{2}$$

where θ denotes the detector parameters, M is the number of retained detections, and b_j, y_j, and s_j denote the predicted bounding box, class label, and confidence score of the j-th detection, respectively. In this work, a pretrained VLOD is deployed on an unlabeled target stream $\{x_t\}_{t=1}^T$ and adapted online without access to source data or target annotations.

Mean-Teacher Test-Time Adaptation. At adaptation step t, mean-teacher TTA maintains a student model $\mathcal{D}_{\theta_t^S}$ and a teacher model $\mathcal{D}_{\theta_t^T}$, both initialized from the pretrained detector $\mathcal{D}_{\theta_0}$. For each target image x_t, the teacher generates predictions from a weakly augmented view. After confidence filtering and non-maximum suppression, the retained predictions are used as pseudo-labels to supervise the student on a strongly augmented view through the detection loss, producing the gradient-updated student parameters $\tilde{\theta}_{t+1}^S$. The teacher is subsequently updated using an EMA update:

$$\theta_{t+1}^T = m\theta_t^T + (1-m)\tilde{\theta}_{t+1}^S, \tag{3}$$

where m is the EMA coefficient. In standard mean-teacher TTA, m remains fixed, making the teacher update sensitive to target domain characteristics, while repeated pseudo-label optimization can induce cumulative student drift.

3.2 DESA-TTA METHOD

An illustration of our DESA-TTA method is shown in Figure 2. It adapts a pretrained VLOD online through two coupled mechanisms. First, dynamic temporal averaging determines the EMA coefficient used to update the teacher. Second, source anchoring constrains the adapted student state before it is incorporated into the teacher. This joint regulation enables adaptation to the target distribution while reducing sensitivity to the EMA coefficient and limiting cumulative student drift.

Dynamic Temporal Averaging. The EMA coefficient m controls the trade-off between preserving the previous teacher state and incorporating the target-adapted student state. However, using a single value of m throughout adaptation is suboptimal because the quality of teacher-generated pseudo-labels can vary substantially across target samples and domains. When the teacher produces a large number of low-confidence predictions, its current state is less reliable for the target sample, and a larger contribution from the target-adapted student may be required. In contrast, when the teacher predictions are confident, stronger temporal averaging is preferable to preserve teacher stability.

For the target image x_t, let M_t denote the number of retained teacher pseudo-labels and let $\bar{s}_t$ denote their mean confidence. For $M_t > 0$, we define the normalized retained prediction density as $d_t = M_t/M_{\max}$, where $M_{\max}$ is a predefined maximum number of retained detections per image. The teacher-uncertainty score is defined as

$$r_t = \min\left(d_t\left(-\log \bar{s}_t\right), 1\right). \tag{4}$$

The score increases when the teacher produces a larger number of low-confidence predictions, indicating greater uncertainty in its current predictions. When $M_t = 0$, both the student adaptation and the teacher EMA update are skipped.

Although r_t characterizes the uncertainty of the current teacher predictions, it remains a sample-wise signal and does not reflect the extent to which the teacher has changed during adaptation. Consequently, the same uncertainty score may require different levels of temporal averaging depending on the accumulated teacher drift from the pretrained source model. When the drift is small, a lower EMA range facilitates the incorporation of target-specific student updates. As the drift increases, stronger temporal averaging becomes necessary to stabilize the teacher and limit the propagation of accumulated pseudo-label errors. DESA-TTA therefore constrains the dynamic coefficient using a drift-dependent EMA range.

Let δ_t^T denote the normalized parameter drift of the teacher from the pretrained source model:

$$\delta_t^T = \frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\frac{\left\|\theta_{t,\ell}^T - \theta_{0,\ell}\right\|_1}{\max\left(\left\|\theta_{0,\ell}\right\|_1, \epsilon\right)}, \tag{5}$$

where $\mathcal{A}$ denotes the set of adapted parameters and $\epsilon > 0$ is a small constant introduced for numerical stability. Based on the teacher drift, DESA-TTA defines separate transition functions for the lower and upper EMA bounds:

$$g_{\min}(\delta_t^T) = \frac{(\delta_t^T)^p}{(\delta_t^T)^p + \lambda_{\min}^p}, \qquad g_{\max}(\delta_t^T) = \frac{(\delta_t^T)^p}{(\delta_t^T)^p + \lambda_{\max}^p},$$

where $\lambda_{\min}$ and $\lambda_{\max}$ determine the drift scales for the lower and upper EMA bounds, respectively, while p controls the sharpness of their transitions.

DESA-TTA defines the lower and upper bounds of the EMA range according to:

$$m_{\min}(\delta_t^T) = m_{\min} - \Delta_{\min}\left(1 - g_{\min}(\delta_t^T)\right),$$
$$m_{\max}(\delta_t^T) = m_{\max} - \Delta_{\max}\left(1 - g_{\max}(\delta_t^T)\right),$$

where $m_{\min}$ and $m_{\max}$ denote the asymptotic lower and upper EMA bounds, while $\Delta_{\min}$ and $\Delta_{\max}$ determine their offsets at low teacher drift. When the teacher remains close to the pretrained source model, both bounds take lower values, facilitating the incorporation of target-specific student updates. As δ_t^T increases, the transition functions approach one and the EMA range gradually shifts toward stronger temporal averaging.

The final dynamic EMA coefficient is selected within this drift-dependent range using the teacher-uncertainty score:

$$m_t = m_{\max}(\delta_t^T) - \left(m_{\max}(\delta_t^T) - m_{\min}(\delta_t^T)\right) r_t. \tag{6}$$

Since $r_t \in [0, 1]$, the resulting coefficient remains within $[m_{\min}(\delta_t^T), m_{\max}(\delta_t^T)]$. A larger teacher-uncertainty score therefore produces a smaller EMA coefficient, whereas a lower score results in stronger temporal averaging.

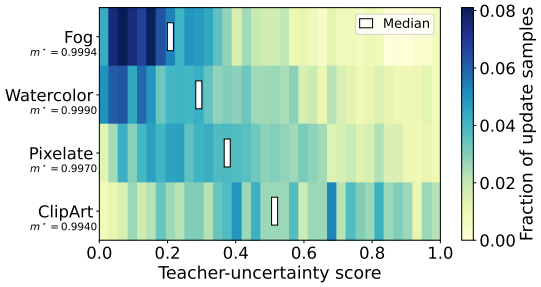


Figure 3: **Relationship between teacher uncertainty and the preferred EMA coefficient.** The heatmap shows the distribution of the teacher-uncertainty score over update samples. White markers indicate the median score, while $m^\star$ denotes the best fixed EMA coefficient for each domain.

As shown in Figure 3, the preferred fixed EMA coefficient decreases as the median teacher-uncertainty score increases. In particular, Fog has the lowest median uncertainty and favors $m^\star = 0.9994$, whereas ClipArt has the highest median uncertainty and favors $m^\star = 0.9940$. This observation empirically supports the inverse mapping between r_t and m_t. Appendix Section A.1 provides complementary expected-error analysis, and Appendix Section A.3 empirically validates r_t and the proposed uncertainty-to-EMA relationship.

Source Anchoring. The student model is updated sequentially using pseudo-labels generated from the target stream. Although these updates are necessary for adaptation, pseudo-label supervision may contain localization and classification errors. Repeated optimization using such pseudo-labels can bias the student trajectory toward accumulated prediction errors, particularly over long target streams. Dynamic temporal averaging regulates the propagation of student updates to the teacher, but it does not directly constrain drift in the student itself. Motivated by the strong ZS capability of the pretrained VLOD, DESA-TTA uses the source initialization as a reference for regularizing the student trajectory while preserving its ability to adapt to the target distribution.

Let $\tilde{\theta}^S_{t+1,\mathcal{A}}$ denote the values of the adapted parameters after the pseudo-label-based gradient update at step t. Source anchoring interpolates this adapted state with the corresponding source initialization:

$$\theta^S_{t+1,\mathcal{A}} = (1 - \alpha_t)\tilde{\theta}^S_{t+1,\mathcal{A}} + \alpha_t \theta_{0,\mathcal{A}}. \tag{7}$$

Because student drift is initially limited but may become more pronounced as pseudo-label errors accumulate over repeated updates, DESA-TTA determines the anchoring strength directly from the current student drift. We define the normalized drift of the gradient-updated student from the source initialization as

$$\delta_t^S = \frac{1}{|\mathcal{A}|} \sum_{\ell \in \mathcal{A}} \frac{\left\| \tilde{\theta}^S_{t+1,\ell} - \theta_{0,\ell} \right\|_1}{\max\left(\left\| \theta_{0,\ell} \right\|_1, \epsilon \right)}. \tag{8}$$

The anchoring coefficient is then defined as

$$\alpha_t = \alpha_{\max} \frac{(\delta_t^S)^p}{(\delta_t^S)^p + \lambda_\alpha^p}, \tag{9}$$

where $0 \leq \alpha_{\max} < 1$ denotes the maximum anchoring coefficient, λ_α determines the anchoring drift scale, and p controls the sharpness of the transition. Consequently, the source constraint remains weak when the student is close to its source initialization and progressively strengthens as student drift increases. Unlike penalty-based L_2 regularization, source anchoring constrains the student through post-update parameter interpolation rather than by adding a penalty to the adaptation objective. A proximal interpretation and empirical comparisons are provided in Appendix Section A.1 and Section A.3, respectively.

DESA-TTA Summary. Although DESA-TTA can, in principle, adapt any subset of model parameters, in our experiments we set $\mathcal{A}$ to the normalization parameters and keep all remaining parameters fixed. The teacher model is used to generate the final predictions. After source anchoring, the teacher is updated from the anchored student using the dynamic EMA coefficient:

$$\theta^T_{t+1} = m_t \theta^T_t + (1 - m_t)\theta^S_{t+1}. \tag{10}$$

4 RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate DESA-TTA under several distribution shifts, including style shifts, autonomous-driving scenarios, low-light conditions, and common corruptions. For style shifts, we use Watercolor, ClipArt, and Comic (Inoue et al., 2018). We use Foggy Cityscapes Sakaridis et al. (2018) and BDD100K Yu et al. (2020) to evaluate performance under adverse autonomous-driving conditions. ExDark Loh & Chan (2019) is used to assess performance under low-light conditions. VOC-C and COCO-C Michaelis et al. (2019) are used to evaluate robustness to 15 common corruption types. These benchmarks cover diverse domain shifts and allow us to assess adaptation behavior over both short and long target streams. Following prior work Gao et al. (2025); Belal et al. (2026), each dataset or corruption is treated as an independent single-domain stream. The evaluation therefore considers cumulative online adaptation within a fixed target domain rather than across changing domains.

Baselines. Following Belal et al. (2026), we compare DESA-TTA with ZS, TPT, VPT, and DPE. The ZS baseline directly evaluates the pretrained VLOD. TPT Shu et al. (2022) and VPT Medeiros et al. (2025) adapt textual and visual prompts, respectively, at test time using target images. DPE Zhang et al. (2024) adapts dual prototypes at test time to accumulate target-domain knowledge from both visual and textual modalities. We also include a standard mean-teacher (MT) baseline with a fixed EMA coefficient and without source anchoring. Finally, we compare with VLOD-TTA Belal et al. (2026), a recent TTA method for VLODs.

Table 1: **Detection performance on benchmark datasets.** We report mAP and AP_{50} on six benchmark datasets: Watercolor, ClipArt, Comic, Foggy Cityscapes, BDD100K, and ExDark. Best results are in bold.

Method	Watercolor		ClipArt		Comic		Foggy Cityscapes		BDD100K		ExDark	
	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}
ZS	26.9	47.9	24.4	40.1	17.8	29.4	13.2	21.3	13.3	22.0	35.2	64.7
TPT	27.3	48.5	24.9	41.3	18.1	29.9	13.3	21.5	13.4	22.2	35.8	65.1
VPT	26.9	49.1	25.0	41.4	18.3	30.9	13.4	21.7	13.7	22.4	35.7	65.3
DPE	27.2	48.9	24.9	41.5	18.9	31.7	13.2	21.6	13.5	22.3	35.9	65.8
MT	29.1	50.8	27.2	45.1	21.1	36.5	13.7	22.0	12.5	20.3	35.5	65.4
VLOD-TTA	29.6	53.1	28.1	45.4	21.4	36.1	13.9	22.2	14.6	24.3	36.4	67.4
DESA-TTA	**31.2**	**55.0**	**29.5**	**48.3**	**23.0**	**38.6**	**15.2**	**24.4**	**16.8**	**27.8**	**37.3**	**69.2**

Table 2: **Detection performance on VOC-C.** AP_{50} is reported under 15 common corruptions.

Method	Noise			Blur				Weather			Digital					
	Gau.	Shot	Imp.	Def.	Glass	Mot.	Zoom	Snow	Frost	Fog	Bright.	Cont.	Elast.	Pix.	JPEG	Avg.
ZS	16.9	17.2	16.2	32.4	11.3	26.9	30.1	46.4	50.6	70.4	73.6	41.3	42.0	8.9	26.1	34.0
TPT	17.4	17.9	16.4	33.1	11.7	27.1	30.7	47.2	51.9	68.9	71.1	42.8	42.7	9.5	27.2	34.4
VPT	17.8	18.2	16.7	33.0	12.3	27.7	30.5	47.7	51.3	69.9	72.5	42.5	43.5	10.9	28.4	34.9
DPE	17.7	18.7	17.5	32.8	12.9	28.5	30.7	48.4	52.1	70.9	73.4	42.9	43.2	11.5	30.7	35.5
MT	27.1	28.3	28.0	33.7	14.8	28.3	30.3	51.5	54.2	69.0	70.0	55.8	51.5	10.9	39.4	39.5
VLOD-TTA	22.9	24.4	24.2	37.2	16.6	29.3	30.9	50.9	53.8	73.2	74.1	48.7	49.4	17.8	40.2	39.6
DESA-TTA	**34.4**	**35.8**	**37.8**	**39.6**	**32.7**	**37.8**	**36.1**	**54.5**	**57.8**	**73.6**	**74.2**	**63.3**	**59.8**	**37.9**	**52.6**	**48.5**

Implementation Details. We implement DESA-TTA using the pretrained YOLO-World-S checkpoint Cheng et al. (2024). The student is optimized with AdamW at a learning rate of 6×10^{-3}. Following VLOD-TTA, we use a batch size of 1. Teacher pseudo-labels are filtered at a confidence threshold of 0.4. For dynamic temporal averaging, we set $M_{\max} = 300$, $m_{\min} = 0.9990$, $\Delta_{\min} = 0.0060$, $\lambda_{\min} = 0.036$, $m_{\max} = 0.9996$, $\Delta_{\max} = 0.0016$, and $\lambda_{\max} = 0.021$. For source anchoring, we use $\alpha_{\max} = 0.035$ and $\lambda_{\alpha} = 0.07$. We use $p = 2$ as the exponent. These hyperparameters are selected using six corruption types applied to the VOC training split, which contains no images from the evaluation sets, and are then kept fixed across all target domains without target-specific tuning. A detailed hyperparameter sensitivity analysis is provided in the Appendix Section A.4. We report AP metrics using the COCO API. Detection results are averaged over three independent runs, with standard deviations provided in the Appendix Section A.3. All experiments are run on an RTX 3090 GPU.

Table 3: **Performance and efficiency comparison using Grounding DINO on COCO-C.** Average mAP for each corruption group is reported along with trainable parameters, GPU memory usage, and latency.

Method	Corruption Type Avg. mAP ↑					Efficiency ↓		
	Noise	Blur	Weather	Digital	Avg.	Trainable Params. (M)	Memory (GB)	Latency (ms/img)
ZS	14.9	11.2	34.7	23.0	20.6	0.00	1.36	210.6
TTAOD-F	21.2	14.3	37.0	31.7	26.0	**0.08**	11.37	701.9
VLOD-TTA	21.1	15.6	38.5	30.4	26.2	0.89	3.76	531.6
DESA-TTA	**22.6**	**15.9**	**38.9**	**33.1**	**27.6**	0.09	**3.37**	**492.5**

Table 4: **Accuracy and efficiency comparison on VOC-C.** FPS, trainable parameters, and average AP_{50} are reported.

Method	FPS ↑	Params (M) ↓	AP_{50} ↑
ZS	89	0	34.0
TPT	9	1.12	34.4
VPT	18	3.93	34.9
DPE	15	0.31	35.5
VLOD-TTA	20	1.61	39.6
DESA-TTA	**31**	**0.021**	**48.5**

4.2 Comparison with SOTA methods

Style Shift. Table 1 shows that DESA-TTA achieves the strongest performance across all three style-shift benchmarks. Relative to the ZS baseline, DESA-TTA improves AP_{50} by $+7.1$, $+8.2$, and $+9.2$ on Watercolor, ClipArt, and Comic, respectively. It also outperforms VLOD-TTA by $+1.9$, $+2.9$, and $+2.5$ AP_{50}, demonstrating stronger adaptation to appearance changes induced by artistic styles. Compared with standard MT, DESA-TTA further improves AP_{50} by $+4.2$, $+3.2$, and $+2.1$, indicating that a fixed EMA coefficient and unconstrained student adaptation are less effective for cumulative TTA.

Adverse Driving Scenes. For adverse driving scenes, DESA-TTA consistently improves performance on both Foggy Cityscapes and BDD100K as shown in Table 1. It increases AP_{50} over the ZS baseline by $+3.1$ on Foggy Cityscapes and $+5.8$ on BDD100K. Compared with VLOD-TTA, DESA-TTA obtains additional gains of $+2.2$ and $+3.5$ AP_{50}, respectively. Notably, standard MT degrades performance on BDD100K, which contains a longer target stream and therefore increases the risk of cumulative student drift. In contrast, DESA-TTA remains effective, indicating that source anchoring is important for stable online adaptation in adverse driving scenes.

Illumination Shift. For the low-light benchmark, the ExDark results in Table 1 show that DESA-TTA obtains the best performance. It improves AP_{50} over the ZS baseline by $+4.5$, showing that online adaptation is beneficial when low-light conditions degrade the reliability of the pretrained detector. DESA-TTA also outperforms VLOD-TTA and standard MT by $+1.8$ and $+3.8$ AP_{50}, respectively. These results indicate that dynamic temporal averaging and source anchoring provide more stable adaptation when pseudo-label quality is affected by poor visibility.

Common Corruptions. The VOC-C results in Table 2 show that DESA-TTA provides strong robustness gains across diverse corruption types. DESA-TTA achieves the highest average AP_{50} across all 15 corruption types, improving from 34.0 to 48.5 over the ZS baseline. This corresponds to a gain of $+14.5$ AP_{50}. DESA-TTA also outperforms VLOD-TTA and standard MT by $+8.9$ and $+9.0$ AP_{50}, respectively. The improvements are particularly pronounced under noise and digital corruptions, where pseudo-label supervision is more error-prone and standard MT can drift. These results show that combining dynamic temporal averaging with source anchoring improves adaptation across diverse corruption types.

4.3 Ablation Studies

Extension to Grounding DINO. To evaluate the generalizability of DESA-TTA across VLOD architectures, we extend it to Grounding DINO Liu et al. (2024a) and evaluate it on COCO-C. This also enables a direct comparison with TTAOD-F (Gao et al., 2025), another mean-teacher-based method whose applicability is limited to transformer-based architectures. As shown in Table 3, DESA-TTA consistently outperforms the ZS model and both TTA baselines across all corruption groups. It improves the average mAP from 20.6 to 27.6, surpassing TTAOD-F and VLOD-TTA by 1.6 and 1.4 points, respectively. DESA-TTA also provides a favorable accuracy–efficiency trade-off. It adapts only 0.09M parameters, compared with 0.89M for VLOD-TTA. Moreover, DESA-TTA requires 3.37 GB of GPU memory and 492.5 ms per image, reducing memory consumption and latency by 10.4% and 7.4%, respectively, compared with VLOD-TTA. Compared with TTAOD-F, it reduces memory consumption by 70.4% and latency by 29.8%. These results demonstrate that DESA-TTA remains effective across both CNN- and transformer-based VLOD architectures.

Contribution of Individual Components. Table 5 shows the contribution of each DESA-TTA component. Starting from MT, the teacher-drift-dependent EMA range provides limited gains, especially on Style, indicating that adjusting the EMA range based on teacher drift is insufficient by itself. Using the teacher-uncertainty score without the drift-dependent range achieves larger gains on Style and comparable gains on VOC-C, reaching 45.6 and 41.4 AP_{50}, respectively. This shows that the sample-wise uncertainty signal is informative but remains insufficient to achieve the best performance. Among the two uncertainty cues, confidence is more effective than box density alone, while combining both cues yields the strongest dynamic EMA (D-EMA) result. Full D-EMA improves AP_{50} to 46.6 on Style and 43.1 on VOC-C, confirming that the teacher-drift-dependent range and uncertainty cues are complementary. For source anchoring, adding a static source anchor to MT improves performance over standard MT, while student-drift-dependent anchoring further improves both benchmarks. Therefore, the anchoring coefficient should adapt to current student drift rather than remain fixed. Finally, combining D-EMA with student-drift-dependent source anchoring yields the best results, reaching 47.3 AP_{50} on Style and 48.5 AP_{50} on VOC-C. These results indicate that the two mechanisms address complementary failure modes: source anchoring constrains the adapted student state before the teacher update, while dynamic temporal averaging regulates its incorporation into the teacher.

Table 5: **Component ablation of DESA-TTA.** Average AP_{50} is reported over the three style-shift datasets and the 15 VOC-C corruptions. Here, δ_t^T indicates the use of teacher drift to determine the EMA range, r_t denotes the teacher-uncertainty score, and $\bar{s}_t$ and d_t denote the confidence and density cues used to compute r_t. SA denotes static source anchoring, D-SA denotes source anchoring controlled by α_t, and α_t is the student-drift-dependent anchoring coefficient.

Method	D-EMA				SA		AP_{50}	
	δ_t^T	r_t	$\bar{s}_t$	d_t	SA	α_t	Style	VOC-C
ZS	–	–	–	–	–	–	39.1	34.0
MT	–	–	–	–	–	–	44.1	39.5
δ_t^T only	✓	–	–	–	–	–	44.2	41.2
r_t only	–	✓	✓	✓	–	–	45.6	41.4
$\bar{s}_t$ only	✓	✓	✓	–	–	–	46.1	42.7
d_t only	✓	✓	–	✓	–	–	43.8	41.3
D-EMA	✓	✓	✓	✓	–	–	46.6	43.1
MT+SA	–	–	–	–	✓	–	45.7	40.6
MT+D-SA	–	–	–	–	✓	✓	46.3	42.2
D-EMA+SA	✓	✓	✓	✓	✓	–	46.4	45.8
DESA-TTA	✓	✓	✓	✓	✓	✓	**47.3**	**48.5**

Runtime Cost. TTA methods adapt the model during inference, introducing additional latency compared with ZS evaluation. Table 4 compares the efficiency of DESA-TTA with existing TTA baselines on YOLO-World. Compared with TPT, VPT, DPE, and VLOD-TTA, DESA-TTA adapts substantially fewer parameters and achieves higher inference speed than all adaptation baselines. At the same time, it obtains the highest AP_{50} of 48.5 on VOC-C, demonstrating a favorable efficiency–performance trade-off among TTA methods. Appendix Section A.3 further shows that, for a stationary target domain, DESA-TTA can be adapted once on an unlabeled target stream and subsequently deployed at full inference speed.

5 Conclusion

We studied cumulative mean-teacher TTA for VLODs and identified two key limitations: the domain sensitivity of a fixed EMA coefficient and cumulative student drift caused by repeated pseudo-label optimization. To address these limitations, we introduced DESA-TTA, combining dynamic temporal averaging with source anchoring. Dynamic temporal averaging regulates teacher updates according to teacher uncertainty and teacher drift, while source anchoring constrains the adapted student toward its pretrained initialization according to student drift. Experiments under style shifts, adverse driving conditions, low-light conditions, and common corruptions show consistent improvements across YOLO-World and Grounding DINO. Ablation studies further show that the two components address complementary failure modes and achieve the strongest performance when combined. Overall, DESA-TTA provides a cost-effective and stable method for adapting VLODs to unlabeled target streams without access to source data.

Appendix. It provides theoretical analysis, implementation details, additional results on backbone generalization, random-seed variation, and held-out target data, design ablations, comparisons with source-anchoring methods, hyperparameter analysis, and qualitative results.

AI USE STATEMENT

We used ChatGPT for feedback on experimental design, assistance with implementation and analysis code, identifying and summarizing related work, and editing manuscript text for clarity. The authors reviewed and tested AI-assisted code, checked suggested references against the original papers, and verified the reported results. We take full responsibility for the methods, results, and claims in this work.

REPRODUCIBILITY STATEMENT

Section 3 specifies the DESA-TTA update rules, with pseudocode in Algorithm 1 of Appendix A.2. Section 4.1 describes the datasets, evaluation metrics, and hyperparameter selection. Appendix A.2 details the online adaptation protocol and implementation, while Appendices A.3–A.4 provide additional results, variation across random seeds, and hyperparameter analysis. Code is included in the supplementary material.

REFERENCES

Atif Belal, Heitor R. Medeiros, Marco Pedersoli, and Eric Granger. Vlod-tta: Test-time adaptation of vision-language object detectors. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2510.00458.

Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2023.

Shilei Cao, Juepeng Zheng, Yan Liu, Baoquan Zhao, Ziqi Yuan, Weijia Li, Runmin Dong, and Haohuan Fu. Exploring test-time adaptation for object detection in continually changing environments, 2025. URL https://arxiv.org/abs/2406.16439.

Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. Santa: Source anchoring network and target alignment for continual test time adaptation. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=V7guVYzvE4.

Yijin Chen, Xun Xu, Yongyi Su, and Kui Jia. Stfar: Improving object detection robustness at test-time by self-training with feature alignment regularization, 2023. URL https://arxiv.org/abs/2303.17937.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7704–7714, 2023. doi: 10.1109/CVPR52729.2023.00744.

Yingjie Gao, Yanan Zhang, Zhi Cai, and Di Huang. Test-time adaptive object detection with foundation model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MO4U4mg0oT.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://api.semanticscholar.org/CorpusID:238744187.

Qi He, Xiao Wu, Jun-Yan He, and Shuai Li. Dual-rate dynamic teacher for source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2067–2076, 2025.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation, 2018. URL https://arxiv.org/abs/1803.11365.

Byungjai Kim, Chanho Ahn, Wissam J. Baddar, Kikyung Kim, Huijin Lee, Saehyun Ahn, Seungju Han, Sungjoo Suh, and Eunho Yang. Test-time ensemble via linear mode connectivity: A path to better adaptation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c242f2b7f60d8c685b6481939330e241-Abstract-Conference.html.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024a.

Yabo Liu, Jinghua Wang, Chao Huang, Yiling Wu, Yong Xu, and Xiaochun Cao. Mlfa: Toward realistic test time adaptive object detection by multi-level feature alignment. *IEEE Transactions on Image Processing*, 33:5837–5848, 2024b. doi: 10.1109/TIP.2024.3473532.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: https://doi.org/10.1016/j.cviu.2018.10.010.

Heitor Medeiros, Atif Belal, Srikanth Muralidharan, Eric Granger, and Marco Pedersoli. Visual modality prompt for adapting vision-language object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2172–2182, 10 2025. doi: 10.1109/ICCV51701.2025.00210.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pp. 728–755, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20079-3. doi: 10.1007/978-3-031-20080-9_42. URL https://doi.org/10.1007/978-3-031-20080-9_42.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The International Conference on Machine Learning*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Xiaoqian Ruan and Wei Tang. Fully test-time adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1038–1047, June 2024.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, March 2018. ISSN 1573-1405. doi: 10.1007/s11263-018-1072-8. URL http://dx.doi.org/10.1007/s11263-018-1072-8.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jiaxu Tian and Fan Lyu. Parameter-selective continual test-time adaptation, 2024. URL https://doi.org/10.1007/978-981-96-0966-6_19.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201, 2022. doi: 10.1109/CVPR52688.2022.00706.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15922–15932, 2023. doi: 10.1109/CVPR52729.2023.01528.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

A Appendix

A.1 Additional Theoretical Analysis

Expected-Error Analysis of Dynamic Temporal Averaging.

Proposition 1. Let $\theta_t^\star$ denote an unknown target-optimal parameter state, and let

$$e_t^T = \theta_t^T - \theta_t^\star, \qquad e_t^S = \theta_{t+1}^S - \theta_t^\star$$

denote the errors of the current teacher and the student state used in the EMA update, respectively. Conditioned on a teacher-uncertainty level r, define

$$V_T(r) = \mathbb{E}\left[\|e_t^T\|_2^2 \mid r_t = r\right],$$

$$V_S(r) = \mathbb{E}\left[\|e_t^S\|_2^2 \mid r_t = r\right],$$

and

$$C(r) = \mathbb{E}\left[(e_t^T)^\top e_t^S \mid r_t = r\right].$$

For an EMA update

$$\theta_{t+1}^T = m\theta_t^T + (1-m)\theta_{t+1}^S,$$

when the teacher and student errors are not identical almost surely, the unique unconstrained coefficient minimizing the expected squared parameter error is

$$m^\star(r) = \frac{V_S(r) - C(r)}{V_T(r) + V_S(r) - 2C(r)}.$$

Furthermore, for fixed $V_S(r)$ and $C(r)$, if $V_S(r) > C(r)$, then the optimal unconstrained EMA coefficient decreases with $V_T(r)$:

$$\frac{\partial m^\star}{\partial V_T} < 0.$$

Proof. The error of the EMA-updated teacher relative to $\theta_t^\star$ is

$$e_{t+1}^T = \theta_{t+1}^T - \theta_t^\star = m e_t^T + (1-m) e_t^S.$$

Therefore, conditioned on $r_t = r$, its expected squared error is

$$\begin{aligned}\mathcal{R}(m \mid r) &= \mathbb{E}\left[\left\| m e_t^T + (1-m) e_t^S \right\|_2^2 \mid r_t = r\right] \\ &= m^2 V_T(r) + (1-m)^2 V_S(r) + 2m(1-m)C(r).\end{aligned}$$

Differentiating with respect to m gives

$$\frac{\partial \mathcal{R}(m \mid r)}{\partial m} = 2m\left(V_T(r) + V_S(r) - 2C(r)\right) - 2\left(V_S(r) - C(r)\right).$$

Setting the derivative to zero yields

$$m^\star(r) = \frac{V_S(r) - C(r)}{V_T(r) + V_S(r) - 2C(r)}.$$

Moreover,

$$\frac{\partial^2 \mathcal{R}(m \mid r)}{\partial m^2} = 2\left(V_T(r) + V_S(r) - 2C(r)\right).$$

Since

$$V_T(r) + V_S(r) - 2C(r) = \mathbb{E}\left[\|e_t^T - e_t^S\|_2^2 \mid r_t = r\right] > 0,$$

the objective is strictly convex in m, and the stationary point is its unique unconstrained minimizer. If the EMA coefficient is constrained to $[0, 1]$, the solution is the projection of the unconstrained minimizer onto this interval.

For fixed $V_S(r)$ and $C(r)$, differentiating $m^\star(r)$ with respect to $V_T(r)$ gives

$$\frac{\partial m^\star}{\partial V_T} = -\frac{V_S(r) - C(r)}{\left(V_T(r) + V_S(r) - 2C(r)\right)^2}.$$

Thus, when $V_S(r) > C(r)$,

$$\frac{\partial m^\star}{\partial V_T} < 0.$$

Therefore, when a larger teacher-uncertainty score reflects greater teacher error, assigning a smaller EMA coefficient is consistent with reducing the expected error of the EMA update. This result establishes the decreasing direction of the uncertainty–EMA relationship; it does not claim that the specific bounded linear mapping used by DESA-TTA is the exact optimal mapping.

Proximal Interpretation and Contraction of Source Anchoring.

Proposition 2. Let $\tilde{\theta}_{t+1,\mathcal{A}}^S$ denote the adapted parameter subset after the pseudo-label-based gradient update, and let $\theta_{0,\mathcal{A}}$ denote its source initialization. For $\lambda_t \geq 0$, consider

$$\theta_{t+1,\mathcal{A}}^S = \arg\min_{\theta_\mathcal{A}} \frac{1}{2}\left\|\theta_\mathcal{A} - \tilde{\theta}_{t+1,\mathcal{A}}^S\right\|_2^2 + \frac{\lambda_t}{2}\left\|\theta_\mathcal{A} - \theta_{0,\mathcal{A}}\right\|_2^2.$$

Its unique solution is

$$\theta_{t+1,\mathcal{A}}^S = (1-\alpha_t)\tilde{\theta}_{t+1,\mathcal{A}}^S + \alpha_t \theta_{0,\mathcal{A}}, \qquad \alpha_t = \frac{\lambda_t}{1+\lambda_t}.$$

Furthermore, this update scales the post-gradient distance from the source initialization by exactly the factor $1 - \alpha_t$:

$$\left\|\theta^S_{t+1,\mathcal{A}} - \theta_{0,\mathcal{A}}\right\|_1 = (1-\alpha_t)\left\|\tilde{\theta}^S_{t+1,\mathcal{A}} - \theta_{0,\mathcal{A}}\right\|_1.$$

Proof. Define

$$\mathcal{J}(\theta_{\mathcal{A}}) = \frac{1}{2}\left\|\theta_{\mathcal{A}} - \tilde{\theta}^S_{t+1,\mathcal{A}}\right\|_2^2 + \frac{\lambda_t}{2}\left\|\theta_{\mathcal{A}} - \theta_{0,\mathcal{A}}\right\|_2^2.$$

Its gradient is

$$\nabla_{\theta_{\mathcal{A}}}\mathcal{J} = \theta_{\mathcal{A}} - \tilde{\theta}^S_{t+1,\mathcal{A}} + \lambda_t\left(\theta_{\mathcal{A}} - \theta_{0,\mathcal{A}}\right).$$

Setting the gradient to zero gives

$$(1+\lambda_t)\theta_{\mathcal{A}} = \tilde{\theta}^S_{t+1,\mathcal{A}} + \lambda_t\theta_{0,\mathcal{A}}.$$

Therefore,

$$\theta^S_{t+1,\mathcal{A}} = \frac{1}{1+\lambda_t}\tilde{\theta}^S_{t+1,\mathcal{A}} + \frac{\lambda_t}{1+\lambda_t}\theta_{0,\mathcal{A}}.$$

Because

$$\alpha_t = \frac{\lambda_t}{1+\lambda_t}, \qquad 1-\alpha_t = \frac{1}{1+\lambda_t},$$

this is exactly the source-anchoring update

$$\theta^S_{t+1,\mathcal{A}} = (1-\alpha_t)\tilde{\theta}^S_{t+1,\mathcal{A}} + \alpha_t\theta_{0,\mathcal{A}}.$$

The Hessian of $\mathcal{J}$ is

$$\nabla^2_{\theta_{\mathcal{A}}}\mathcal{J} = (1+\lambda_t)I,$$

which is positive definite for $\lambda_t \geq 0$. Hence, the solution is unique.

Subtracting $\theta_{0,\mathcal{A}}$ from the anchoring update gives

$$\begin{aligned}\theta^S_{t+1,\mathcal{A}} - \theta_{0,\mathcal{A}} &= (1-\alpha_t)\tilde{\theta}^S_{t+1,\mathcal{A}} + \alpha_t\theta_{0,\mathcal{A}} - \theta_{0,\mathcal{A}} \\ &= (1-\alpha_t)\left(\tilde{\theta}^S_{t+1,\mathcal{A}} - \theta_{0,\mathcal{A}}\right).\end{aligned}$$

Since $0 \leq \alpha_t < 1$, taking the ℓ_1 norm yields

$$\left\|\theta^S_{t+1,\mathcal{A}} - \theta_{0,\mathcal{A}}\right\|_1 = (1-\alpha_t)\left\|\tilde{\theta}^S_{t+1,\mathcal{A}} - \theta_{0,\mathcal{A}}\right\|_1.$$

Thus, source anchoring retains a $1-\alpha_t$ fraction of the post-gradient parameter distance from the source initialization, equivalently removing an α_t fraction of that distance.

A.2 Implementation and Evaluation Protocol

Online Adaptation Procedure. Algorithm 1 summarizes the online adaptation procedure of DESA-TTA. For each incoming target image, the current teacher first produces the detection output used for evaluation. The same teacher detections are then filtered and used as pseudo-labels to adapt the student. After the student update, source anchoring partially restores the adapted parameters toward the pretrained model, and the teacher is updated with the dynamic EMA coefficient.

Additional Implementation Details. For each target stream, the student and teacher are initialized from the same pretrained VLOD checkpoint and target images are processed sequentially with batch size one. Each image is used once, and no source images or target annotations are used during adaptation. The teacher branch uses a weak view of the target image for detection and pseudo-label generation. This weak view applies YOLOv5 keep-ratio resizing to 640×640, letterbox resizing with padding value 114, text loading, and input packing.

The student branch uses a strong view of the same image for pseudo-label supervision. This strong view applies YOLOv5 keep-ratio resizing to 640×640, random horizontal flipping with probability 0.5, photometric distortion, random erasing with probability 0.7, letterbox resizing with padding value 114, text loading, and input packing. Teacher detections are filtered by confidence thresholding and NMS before being used as pseudo-labels. If no pseudo-label is retained, the optimizer step, source anchoring step, and teacher EMA update are skipped for that image.

Algorithm 1: DESA-TTA Online Adaptation

input: pretrained VLOD θ_0, target stream $\{x_t\}_{t=1}^{T}$, class names $\mathcal{C}$
output: predictions for the target stream

Initialize student model $\theta^S \leftarrow \theta_0$
Initialize teacher model $\theta^T \leftarrow \theta_0$
for $t \leftarrow 1$ **to** T **do**
  Predict with the current teacher: $\mathcal{P}_t \leftarrow D_{\theta^T}(x_t, \mathcal{C})$
  Use $\mathcal{P}_t$ as the output prediction for image x_t
  Filter $\mathcal{P}_t$ to obtain pseudo-labels $\hat{\mathcal{Y}}_t$
  if $\hat{\mathcal{Y}}_t$ *is empty* **then**
    Keep the models unchanged
  else
    Compute teacher uncertainty from pseudo-label confidence and box density
    Compute dynamic EMA coefficient m_t
    Update student θ^S using pseudo-labels $\hat{\mathcal{Y}}_t$
    Compute source anchoring coefficient α_t
    Partially restore θ^S toward θ_0
    Update teacher for the next target image: $\theta^T \leftarrow m_t\theta^T + (1 - m_t)\theta^S$

Table 6: **Summary of target datasets used in the experiments.**

Dataset	Shift Type	# Target Images	# Classes
Watercolor	Artistic style shift	1,000	6
ClipArt	Artistic style shift	500	20
Comic	Artistic style shift	1,000	6
Foggy Cityscapes	Synthetic fog	500	8
BDD100K	Real-world driving scenes	10,000	10
ExDark	Low-light scenes	7,363	12
VOC-C	Common corruptions	4,952 per corruption	20
COCO-C	Common corruptions	5,000 per corruption	80

For YOLO-World experiments, only the batch-normalization parameters are adapted, while all other parameters are kept fixed. The student is optimized with AdamW using a learning rate of 6×10^{-3}, and one gradient update is performed per target image. For Grounding DINO, we adapt the layer-normalization and group-normalization parameters while keeping the remaining parameters fixed. After each student update, source anchoring is applied to the adapted parameters, and the teacher is updated using the dynamic EMA coefficient. Each dataset or corruption type is adapted independently from the pretrained initialization.

Dataset Details. Table 6 summarizes the target datasets used in our experiments. For VOC-C and COCO-C, each corruption type is processed as an independent target stream.

A.3 Additional Experimental Results

Teacher-Uncertainty Score and Pseudo-Label Quality. We examine how the teacher-uncertainty score r_t relates to pseudo-label quality at the image level. Using ground-truth annotations only for this analysis, we measure pseudo-label (PL) precision and IoU and compute their Spearman correlations with r_t. We also report partial correlations controlling for ground-truth object count and summarize pseudo-label quality across quintiles of r_t.

Table 7 shows negative correlations with PL precision and IoU for every style dataset and VOC-C corruption. On average, the correlations become stronger after controlling for ground-truth object count. In Table 8, both quality measures decrease across the Style quintiles. The VOC-C aver-

Table 7: Image-level Spearman correlations between r_t and pseudo-label quality. Partial correlations control for ground-truth object count. Each Avg. row is the unweighted mean over its datasets or corruptions.

Dataset / corruption	ρ PL Prec.	ρ PL IoU	Partial ρ PL Prec.	Partial ρ PL IoU
Watercolor	-0.642	-0.595	-0.656	-0.592
ClipArt	-0.433	-0.411	-0.561	-0.516
Comic	-0.458	-0.428	-0.541	-0.476
Style Avg.	-0.511	-0.478	-0.586	-0.528
Gaussian noise	-0.287	-0.403	-0.361	-0.448
Shot noise	-0.316	-0.435	-0.396	-0.471
Impulse noise	-0.300	-0.430	-0.371	-0.469
Defocus blur	-0.357	-0.462	-0.412	-0.484
Glass blur	-0.318	-0.396	-0.375	-0.426
Motion blur	-0.346	-0.452	-0.398	-0.467
Zoom blur	-0.361	-0.336	-0.389	-0.353
Snow	-0.403	-0.510	-0.471	-0.528
Frost	-0.418	-0.530	-0.495	-0.557
Fog	-0.485	-0.598	-0.563	-0.606
Brightness	-0.479	-0.595	-0.562	-0.605
Contrast	-0.429	-0.549	-0.497	-0.566
Elastic transformation	-0.394	-0.512	-0.495	-0.546
Pixelate	-0.259	-0.364	-0.334	-0.405
JPEG compression	-0.306	-0.449	-0.404	-0.495
VOC-C Avg.	-0.364	-0.468	-0.435	-0.495

Table 8: Pseudo-label quality across quintiles of r_t, from lowest (Q_1) to highest (Q_5). Quintiles are formed separately for each dataset or corruption. Values are then averaged across the three style datasets or 15 VOC-C corruptions.

Group	Metric	Q_1	Q_2	Q_3	Q_4	Q_5
Style	PL precision	0.927	0.817	0.676	0.591	0.530
	PL IoU	0.793	0.693	0.590	0.542	0.496
VOC-C	PL precision	0.954	0.859	0.746	0.657	0.711
	PL IoU	0.877	0.765	0.651	0.575	0.634

ages decrease from quintiles Q_1 through Q_4, then rise in Q_5. These results indicate that r_t is an informative image-level signal of pseudo-label quality.

Effect of the EMA Mapping Direction. We evaluate the uncertainty-to-EMA relationship by comparing three update rules: fixed EMA, the reversed uncertainty-to-EMA mapping, and the proposed mapping. All three variants retain dynamic source anchoring. The proposed and reversed mappings use the same drift-dependent bound functions, but differ in their response to uncertainty: higher uncertainty decreases the EMA coefficient under DESA-TTA and increases it under the reversed mapping. All other adaptation settings are unchanged.

As shown in Table 9, DESA-TTA improves over fixed EMA with dynamic source anchoring by 1.0 AP_{50} on Style and 6.3 AP_{50} on VOC-C. It also outperforms the reversed mapping by 2.6 and 5.2 AP_{50}, respectively. Compared with the reversed mapping, improvements occur on all three style datasets and all 15 corruptions, with the largest gains under pixelation (+11.0), glass blur (+8.4), and JPEG compression (+7.0). The preceding analysis shows that higher r_t generally indicates lower pseudo-label quality. This ablation shows that lowering the EMA coefficient when r_t increases gives higher AP_{50} than fixed EMA or the reversed mapping. Together, these findings support using r_t to make the teacher more responsive when its pseudo-labels are less reliable.

Detailed Grounding DINO Results on COCO-C. Table 10 reports the corruption-wise results corresponding to the grouped comparison presented in the main paper. DESA-TTA improves over

Table 9: Comparison of EMA update rules with dynamic source anchoring. AP_{50} is reported for each dataset and corruption. Style Avg. averages the three style datasets, and VOC-C Avg. averages the 15 corruptions.

Dataset / corruption	Fixed EMA	Reversed EMA	DESA-TTA
Watercolor	53.1	53.0	55.0
ClipArt	47.5	44.8	48.3
Comic	38.3	36.2	38.6
Style Avg.	46.3	44.7	47.3
Gaussian noise	29.1	28.7	34.4
Shot noise	30.5	30.2	35.8
Impulse noise	30.5	31.0	37.8
Defocus blur	34.7	36.5	39.6
Glass blur	19.8	24.3	32.7
Motion blur	30.1	32.5	37.8
Zoom blur	31.5	33.3	36.1
Snow	52.5	50.5	54.5
Frost	55.1	54.9	57.8
Fog	69.6	71.7	73.6
Brightness	70.8	72.6	74.2
Contrast	59.2	57.0	63.3
Elastic transformation	55.3	53.7	59.8
Pixelate	18.9	26.9	37.9
JPEG compression	45.4	45.6	52.6
VOC-C Avg.	42.2	43.3	48.5

Table 10: **Detailed Grounding DINO results on COCO-C.** mAP is reported for each corruption type, with the best result in each column highlighted in bold.

Method	**Noise**			**Blur**				**Weather**			**Digital**					
	Gau.	Shot	Imp.	Def.	Glass	Mot.	Zoom	Snow	Frost	Fog	Bright.	Cont.	Elast.	Pix.	JPEG	Avg.
ZS	13.7	16.0	15.0	16.8	7.5	13.6	6.7	27.5	32.5	44.2	44.1	21.9	22.5	5.3	21.1	20.6
TTAOD-F	20.2	22.0	21.4	17.8	14.5	16.9	7.9	31.1	34.7	45.1	44.9	30.6	29.9	23.6	29.2	26.0
VLOD-TTA	20.5	21.8	20.9	**20.6**	13.1	**18.4**	**10.1**	30.8	**36.6**	**48.2**	**46.2**	28.8	28.7	19.7	28.6	26.2
DESA-TTA	**21.9**	**22.8**	**23.2**	20.4	**16.9**	17.6	8.5	**32.7**	36.1	47.9	45.4	**35.7**	**30.1**	**24.5**	**29.9**	**27.6**

both ZS inference and TTAOD-F across all 15 corruption types and achieves the best performance on 9 corruptions. Relative to the strongest baseline for each corruption, the largest improvements are obtained on contrast ($+5.1$ mAP), glass blur ($+2.4$), impulse noise ($+1.8$), and snow ($+1.6$). Overall, DESA-TTA achieves 27.6 mAP, outperforming ZS inference, TTAOD-F, and VLOD-TTA by 7.0, 1.6, and 1.4 points, respectively.

Generalization Across Backbones. To assess whether DESA-TTA remains effective for a larger VLOD, we evaluate it using YOLO-World-Large on the style-shift benchmarks. As shown in Table 11, the same trend observed with YOLO-World-Small is preserved at a larger model scale. Standard mean-teacher adaptation improves over zero-shot inference, while DESA-TTA further improves performance across all three datasets. On average, DESA-TTA improves AP_{50} by $+8.4$ points over zero-shot inference, $+4.6$ points over VLOD-TTA, and $+3.2$ points over standard mean-teacher adaptation. These results indicate that DESA-TTA remains effective beyond the YOLO-World-S setting.

Comparison with CoTTA. We further compare DESA-TTA with CoTTA (Wang et al., 2022), a continual TTA method that mitigates forgetting by randomly restoring a subset of model parameters to their source-pretrained values. As shown in Table 12, CoTTA improves over standard MT on VOC-C, indicating that source restoration can help under long corruption streams. However, random restoration remains less effective on the shorter style-shift streams. Adding dynamic EMA to CoTTA improves both Style and VOC-C, but DESA-TTA achieves the best performance on both benchmarks. These results show that random source restoration can improve robustness, while student-drift-dependent source anchoring provides more effective control of cumulative student drift.

Table 11: **Performance with YOLO-World-L on style-shift benchmarks.** AP_{50} is reported.

Method	Watercolor	Comic	ClipArt	Avg.
ZS	55.3	37.9	50.6	47.9
VLOD-TTA	58.3	42.8	53.9	51.7
MT	58.9	45.1	55.3	53.1
DESA-TTA	**61.8**	**48.3**	**58.7**	**56.3**

Table 12: **Comparison with source-restoration baselines.** Average AP_{50} is reported on Style and VOC-C benchmarks.

Method	Style	VOC-C
ZS	39.1	34.0
MT	44.1	39.5
CoTTA	44.6	44.4
D-EMA + CoTTA	46.0	47.1
DESA-TTA	**47.3**	**48.5**

Comparison with L_2 Regularization. To determine whether conventional regularization can effectively control student drift, we compare source anchoring with a source-centered L_2 baseline that augments the detection loss with $\lambda_{\text{reg}} \sum_{\ell \in \mathcal{A}} \|\theta_\ell^S - \theta_{0,\ell}\|_2^2$. Unlike source anchoring, which interpolates the gradient-updated student parameters with their pretrained values, this baseline imposes the source constraint directly through the optimization objective. All variants use dynamic temporal averaging, thereby isolating the effect of the student constraint. As shown in Table 13, L_2 regularization improves VOC-C performance from 43.1 to 44.2 AP_{50} but reduces Style performance from 46.6 to 45.1. Static source anchoring provides a larger improvement on VOC-C, reaching 45.8 AP_{50}, but also remains below D-EMA on Style. Thus, neither penalty-based regularization nor static source anchoring provides consistent improvements across both benchmarks. In contrast, drift-dependent source anchoring achieves the best performance on both Style and VOC-C, outperforming L_2 regularization by 2.2 and 4.3 AP_{50}, respectively.

Variation Across Random Seeds. All results reported in the main paper are averaged over three independent runs with different random seeds for each dataset. Table 14 reports the corresponding standard deviations. Across both benchmark groups, the standard deviation does not exceed 0.24 AP_{50} for any method, while DESA-TTA retains the highest mean performance, indicating consistent performance across runs.

Generalization to Held-Out Target Data. DESA-TTA performs cumulative adaptation by continuously updating the model over the target stream. To assess whether the adapted model generalizes beyond the samples used for adaptation, we hold out VOC-C target images from the adaptation stream and evaluate model snapshots saved at different adaptation iterations on this held-out subset. As shown in Figure 4, AP_{50} consistently increases with the number of adaptation iterations, yielding an average gain of $+16.6$ AP_{50} over the ZS model. The adaptation process takes approximately 3 minutes on a single RTX 3090 GPU. These results indicate a practical deployment setting in which the detector can be adapted to a specific target environment using an unlabeled target stream and then deployed at full inference speed.

A.4 Hyperparameter Analysis

The hyperparameter values used in all experiments were selected using six corruption types from the VOC-C training set, as described in the main paper. Figure 5 provides a sensitivity analysis of five hyperparameters: $\alpha_{\max}$ and λ_α for source anchoring, $\lambda_{\min}$ and $\lambda_{\max}$ for dynamic temporal averaging, and the exponent p, which is used in both components. These parameters control the maximum anchoring strength and the responses of both components to parameter drift. We do not treat $M_{\max}$ as a tunable hyperparameter because it is fixed to the detector's maximum number of retained predictions. We fix $m_{\min}$, $m_{\max}$, $\Delta_{\min}$, and $\Delta_{\max}$ because they jointly define the low- and high-drift EMA intervals, chosen based on the effective coefficient range observed in the fixed-EMA

Table 13: **Comparison of source anchoring and L_2 regularization.** Average AP_{50} is reported on Style and VOC-C benchmarks.

Method	Style	VOC-C
D-EMA	46.6	43.1
D-EMA + L_2 regularization	45.1	44.2
D-EMA + SA	46.4	45.8
DESA-TTA	**47.3**	**48.5**

Table 14: **Mean AP_{50} $\pm$ standard deviation over three seeds.** Style is averaged over Watercolor, ClipArt, and Comic, and VOC-C over all 15 corruptions.

Method	Style	VOC-C
ZS	39.1	34.0
MT	44.1 ± 0.21	39.5 ± 0.16
D-EMA	46.6 ± 0.13	43.1 ± 0.21
DESA-TTA	$\mathbf{47.3 \pm 0.24}$	$\mathbf{48.5 \pm 0.15}$

analysis. Each analyzed hyperparameter is varied over five values while the others are fixed at their selected values.

Maximum Anchoring Coefficient. Figure 5(a) shows different responses to the anchoring strength across the two benchmark groups. VOC-C performance improves as $\alpha_{\max}$ increases up to 0.045, suggesting that stronger restoration toward the pretrained model is beneficial over long corruption streams, where repeated pseudo-label updates create a greater risk of cumulative student drift. In contrast, Style performance peaks at $\alpha_{\max} = 0.025$ and subsequently decreases, indicating that excessive anchoring can restrict the parameter changes required to adapt to substantial appearance shifts. Performance on both benchmarks decreases at the largest value, showing that overly strong anchoring eventually suppresses useful target adaptation. Our setting, $\alpha_{\max} = 0.035$, lies between the two benchmark-specific optima and achieves 47.3 AP_{50} on Style and 48.5 AP_{50} on VOC-C.

Anchoring Drift Scale. Figure 5(b) examines the anchoring drift scale λ_α, which controls how rapidly the anchoring coefficient increases with student drift. When λ_α is small, the anchoring coefficient becomes large at relatively low drift, pulling the updated student parameters strongly toward their pretrained values before sufficient adaptation has occurred. When λ_α is too large, anchoring remains weak even as student drift increases, reducing its ability to limit accumulated pseudo-label errors. The selected value $\lambda_\alpha = 0.07$ lies in this intermediate regime, achieving the highest VOC-C performance of 48.5 AP_{50} and tying the highest Style performance of 47.3 AP_{50}.

Lower EMA Drift Scale. Figure 5(c) evaluates the lower EMA drift scale $\lambda_{\min}$, which controls how quickly the lower EMA bound increases from its reduced initial value as teacher drift grows. Increasing $\lambda_{\min}$ keeps the lower bound reduced over a wider drift range, allowing the teacher to incorporate student updates more rapidly for longer during adaptation. This improves VOC-C performance up to $\lambda_{\min} = 0.036$, whereas Style performance begins to decrease beyond 0.030, suggesting that maintaining a lower EMA bound for too long can make the teacher overly responsive to recent student updates on shorter target streams. We select $\lambda_{\min} = 0.036$, which attains the highest VOC-C performance of 48.5 AP_{50} while remaining within 0.2 points of the best Style result.

Upper EMA Drift Scale. Figure 5(d) examines the upper EMA drift scale $\lambda_{\max}$, which controls how quickly the upper EMA bound increases from its reduced initial value as teacher drift grows. When $\lambda_{\max}$ is small, the upper bound increases rapidly, imposing strong temporal averaging before the teacher has sufficiently adapted to the target distribution. Conversely, a large $\lambda_{\max}$ keeps the upper bound reduced over a wider drift range, weakening temporal smoothing and increasing the teacher's sensitivity to recent student updates. At $\lambda_{\max} = 0.021$, both benchmark groups attain their highest performance, reaching 47.3 AP_{50} on Style and 48.5 AP_{50} on VOC-C.

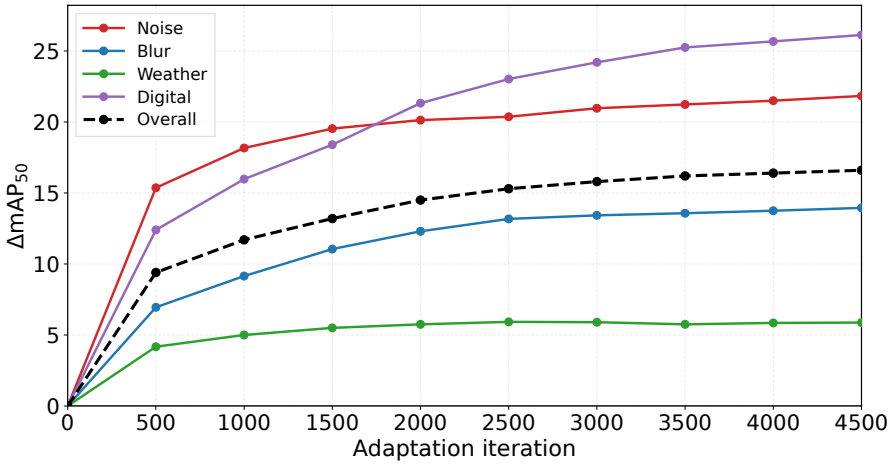
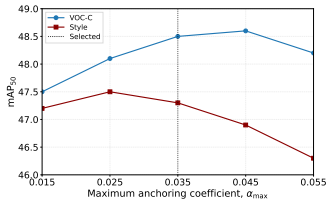
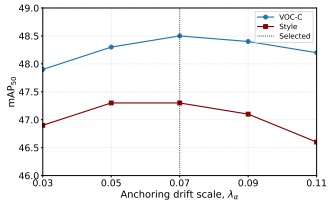


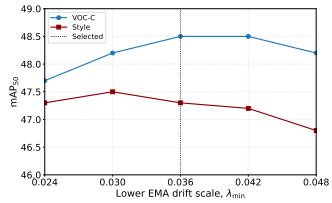
Figure 4: **Generalization to a held-out subset of VOC-C.** DESA-TTA snapshots are evaluated at different adaptation iterations on held-out target data. We report AP_{50} gains over the ZS checkpoint for each corruption group and the overall average.



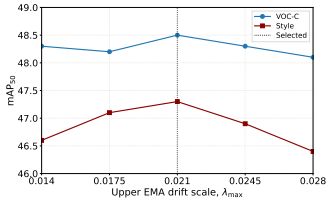
(a) Maximum anchoring coefficient.



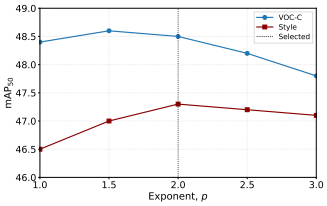
(b) Anchoring drift scale.



(c) Lower EMA drift scale.



(d) Upper EMA drift scale.



(e) Transition exponent.

Figure 5: **Hyperparameter sensitivity of DESA-TTA on VOC-C and the style-shift datasets.** Average AP_{50} is reported, with Style averaged over Watercolor, ClipArt, and Comic and VOC-C averaged over the 15 corruption types. Each hyperparameter is varied independently while the remaining settings are fixed at their selected values. The vertical dotted line indicates the selected value.

Transition Exponent. Finally, Figure 5(e) examines the exponent p, which controls the sharpness of the drift-dependent transition in the source-anchoring coefficient and both EMA bounds. A small value produces gradual transitions that weakly differentiate between low- and high-drift states, whereas a large value makes the transitions increasingly abrupt and therefore more sensitive to small changes around their respective drift scales. The results favor an intermediate transition: Style performance peaks at $p = 2$, while VOC-C peaks at $p = 1.5$ and differs by only 0.1 AP_{50} at $p = 2$. At the selected value $p = 2$, DESA-TTA achieves 47.3 AP_{50} on Style and 48.5 AP_{50} on VOC-C. The resulting configuration uses $\alpha_{\max} = 0.035$, $\lambda_\alpha = 0.07$, $\lambda_{\min} = 0.036$, $\lambda_{\max} = 0.021$, and $p = 2$.

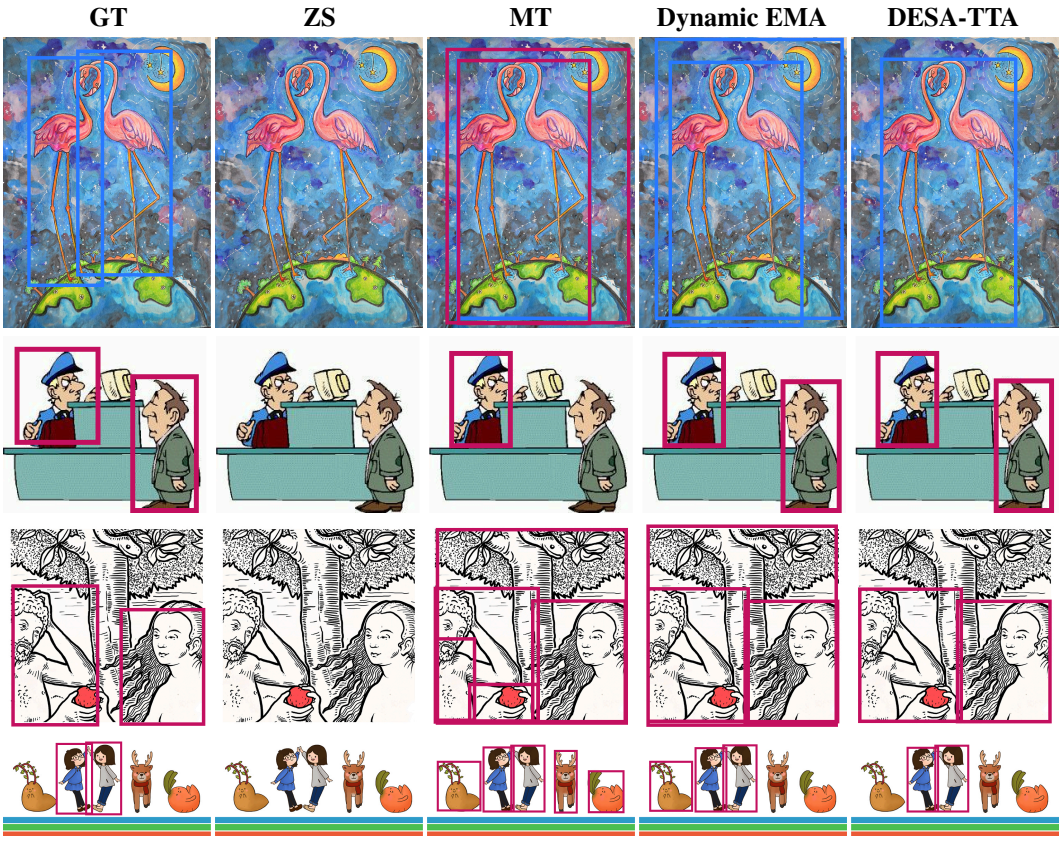


Figure 6: **Qualitative comparison under style shifts.** The first two rows show examples from Watercolor and ClipArt, respectively, while the last two rows show Comic examples. Columns compare ground truth (GT), zero-shot inference (ZS), standard mean-teacher adaptation (MT), dynamic temporal averaging without source anchoring (Dynamic EMA), and DESA-TTA.

A.5 Qualitative Analysis

Style Shift. Figure 6 presents qualitative comparisons on Watercolor, ClipArt, and Comic. In the Watercolor example, ZS produces no detections for the two annotated birds, whereas DESA-TTA identifies both instances. In the ClipArt example, ZS misses both annotated persons and MT recovers only one, while Dynamic EMA and DESA-TTA detect both. In the Comic examples, MT introduces several detections that do not correspond to annotated objects. Dynamic EMA reduces these false positives, while DESA-TTA further suppresses the remaining spurious detections without removing the correctly detected persons. These examples indicate that DESA-TTA can improve object recovery while limiting false-positive predictions under style shifts.

Common Corruptions. Figure 7 presents qualitative comparisons under impulse noise, glass blur, snow, and elastic transformation. Under impulse noise, ZS fails to detect the annotated objects, while MT and Dynamic EMA identify the two prominent riders but miss the smaller instance on the right. DESA-TTA additionally identifies this instance while retaining the other detections. Under glass blur, DESA-TTA recovers a more complete set of annotated persons than the compared variants. Under snow, it preserves detections for several objects missed by ZS, and under elastic transformation, it identifies more of the densely arranged motorcycles. These observations are consistent with the quantitative results, indicating that DESA-TTA improves instance recovery across diverse corruption types.



Figure 7: **Qualitative comparison under common corruptions.** Rows show examples under impulse noise, glass blur, snow, and elastic transformation, from top to bottom. Columns compare ground truth (GT), zero-shot inference (ZS), standard mean-teacher adaptation (MT), dynamic temporal averaging without source anchoring (Dynamic EMA), and DESA-TTA.