VLOD-TTA: Test-Time Adaptation of Vision-Language Object Detectors

Atif Belal, Heitor R. Medeiros, Marco Pedersoli, and Eric Granger

LIVIA, Dept. of Systems Engineering, ÉTS Montréal, Canada
International Laboratory on Learning Systems (ILLS)
{atif.belal.1, heitor.rapela-medeiros.1}@ens.etsmtl.ca,
{marco.pedersoli, eric.granger}@etsmtl.ca

Abstract. Vision-language object detectors (VLODs) such as YOLO-World and Grounding DINO exhibit strong zero-shot generalization, but their performance degrades under distribution shift. Test-time adaptation (TTA) offers a practical way to adapt models online using only unlabeled target data. However, despite substantial progress in TTA for vision-language classification, TTA for VLODs remains largely unexplored. The only prior method relies on a mean-teacher framework that introduces significant latency and memory overhead. To this end, we introduce VLOD-TTA, a TTA method that leverages dense proposal overlap and image-conditioned prompts to adapt VLODs with low additional overhead. VLOD-TTA combines (i) an IoU-weighted entropy objective that emphasizes spatially coherent proposal clusters and mitigates confirmation bias from isolated boxes, and (ii) image-conditioned prompt selection that ranks prompts by image-level compatibility and aggregates the most informative prompt scores for detection. Our experiments across diverse distribution shifts, including artistic domains, adverse driving conditions, low-light imagery, and common corruptions, indicate that VLOD-TTA consistently outperforms standard TTA baselines and the prior state-of-the-art method using YOLO-World and Grounding DINO. Code: https://github.com/imatif17/VLOD-TTA

Keywords: Test-time adaptation · Vision-language object detection

1 Introduction

Object detectors (ODs) localize and classify objects in images [43], with applications in surveillance [27], autonomous driving [11], and medical imaging [20]. Recently, vision-language ODs (VLODs) such as YOLO-World [4] and Grounding DINO [22] have demonstrated strong zero-shot (ZS) generalization by aligning region-level visual features with textual representations through large-scale image-text pretraining [14, 32].

Despite their strong ZS capability, VLODs remain sensitive to distribution shift at test time [33]. Although source-free domain adaptation [19, 34] can mitigate this issue, it typically assumes access to pre-collected target-domain data and offline adaptation. These assumptions are often impractical in deployment

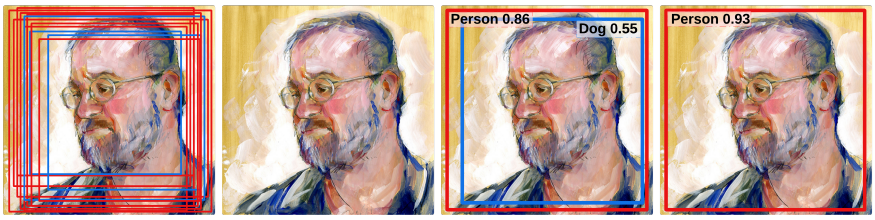


Fig. 1: Motivation for IWE. Left→right: (i) Raw predictions before thresholding for two classes — a true positive **Person (red)** (bbox cluster size = 167, max score = 0.14) and a false positive **Dog (blue)** (bbox cluster size = 45, max score = 0.15); (ii) In ZS, all scores remain below the detection threshold, causing a missed detection; (iii) Standard entropy minimization over-confidently sharpens scores, resulting in a dog false positive; and (iv) **IWE** focuses updates on dense bbox clusters and suppresses isolated boxes, producing only the correct person detection.

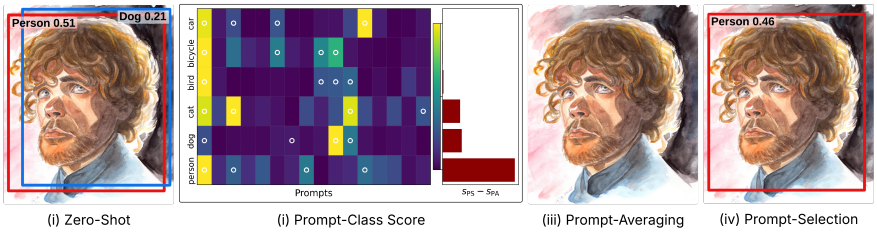


Fig. 2: Motivation for IPS. Left→right: (i) ZS produces a correct **Person (red)** detection but also a **Dog (blue)** false positive; (ii) prompt–class score heatmap, where circles denote prompts selected by **IPS** and the right-margin bars show $S_{\mathrm{PS}} - S_{\mathrm{PA}}$, indicating that prompt selection increases the correct person-class score; (iii) prompt averaging (PA) reduces detection confidence, producing no detections; and (iv) prompt selection (PS) suppresses the dog false positive while preserving the person detection.

settings, where the test environment changes over time and adaptation must be performed online. This motivates test-time adaptation (TTA) for VLODs, which adapts the model online during deployment using only unlabeled test data.

The only prior TTA method specialized for VLODs, TTAOD-F [9], adopts a multimodal prompt-based mean-teacher framework with joint text and visual prompt tuning, together with memory-enhanced pseudo-labeling. While effective, its teacher-based design introduces substantial computational overhead at test time due to extra model components and auxiliary feature processing. Moreover, it is tailored to transformer-based detectors and does not readily extend to CNN-based VLODs such as YOLO-World. This highlights the need for an efficient TTA framework for VLODs that avoids teacher-based adaptation and extends beyond transformer-based detectors.

In parallel, TTA methods for vision-language classification often rely on entropy minimization over augmented views [24, 33]. While entropy minimization offers a lower-overhead alternative to teacher-based approaches, it has two key limitations for ODs. First, it amplifies confirmation bias by sharpening the highest class score, which can make mislocalized proposals overly confident [7]. Sec-

ond, it ignores proposal structure and assigns the same weight to isolated or cross-instance boxes as to spatially consistent overlapping proposal clusters. As illustrated in Figure 1, standard entropy minimization can increase the scores of both *person* and *dog* predictions without considering spatial coherence, leading to a false-positive dog detection.

Beyond model adaptation, prompt ensembling is a common strategy for improving the robustness of vision-language models (VLMs), typically by averaging multiple prompt templates per class [29]. However, for VLODs, prompt averaging provides only marginal gains and can even degrade performance, as shown in Section 4.3. As shown in Figure 2, prompt averaging reduces the *person* score below the detection threshold, resulting in a missed detection.

To address these limitations, we propose **VLOD-TTA**, a TTA framework for VLODs with two key components: **IoU-weighted entropy minimization (IWE)** and **image-conditioned prompt selection (IPS)**. Modern ODs generate dense, overlapping proposals that provide partially redundant views of the same instance. IWE leverages this spatial redundancy by assigning each proposal a weight based on its local IoU affinity with overlapping, class-consistent proposals. This biases adaptation toward spatially coherent proposals and reduces confirmation bias from isolated or mislocalized ones. In Figure 1, IWE increases scores within the dominant *person* cluster instead of uniformly amplifying all proposals, thereby suppressing the false-positive *dog* prediction produced by standard entropy minimization.

We further introduce IPS, an OD-specific prompt selection mechanism. Rather than averaging all prompts, IPS computes an image-conditioned selection score for each prompt and retains only the top-ρ fraction per class. The retained prompt logits are then fused with the base OD logits. By conditioning on test-image features, IPS retains the most informative, context-relevant prompts while suppressing irrelevant ones. In Figure 2, IPS selects prompts aligned with the input image and raises the *person* score relative to prompt averaging, resulting in a correct detection. Together, IWE and IPS provide an efficient TTA framework that improves VLOD robustness under distribution shift with low overhead.

Our main contributions are summarized as follows. (1) We introduce IWE, a detection-specific entropy minimization objective for VLOD TTA that exploits local proposal overlap and mitigates confirmation bias. **(2)** An IPS mechanism is introduced that replaces prompt averaging with image-relevant prompt fusion, improving VLOD robustness under distribution shift. **(3)** Comprehensive benchmarking on six standard detection datasets and 15 common corruptions, covering 96 distinct test scenarios, shows that VLOD-TTA consistently outperforms TTA baselines on state-of-the-art CNN- and transformer-based VLODs.

2 Related Work

Test-Time Adaptation. TTA mitigates domain shift by updating a subset of model parameters using only unlabeled test data. Tent [35] minimizes predictive

entropy by updating batch normalization parameters over batches of images. Memo [40] avoids batches by minimizing marginal entropy across augmented views of a single test image. For VLMs, TTA methods mainly fall into two families: prompt-tuning [8,33] and cache-based methods [17,39]. Prompt-tuning methods optimize continuous prompt tokens in the text encoder for each test image. Cache-based methods maintain a memory of high-confidence target features and pseudo-labels to calibrate predictions online. These methods operate at the image level for classification and do not involve region proposals, leaving localization unaddressed. In contrast, VLOD-TTA performs proposal-level adaptation for open-vocabulary detection, jointly addressing localization and classification.

Vision-Language Object Detectors. VLODs localize and classify categories specified by text, thereby relaxing the closed-set constraint of conventional ODs. Early methods leverage VLMs [29] to transfer language-aligned semantics into detector classifier heads [10, 38, 41]. Vocabulary scaling further decouples localization and classification by training large-vocabulary classifiers with image-level labels while keeping proposals class-agnostic [42]. Grounded pretraining unifies detection and phrase grounding to learn language-aware object representations [18]. Grounding DINO [22] fuses language into a transformer detector via language-conditioned queries and cross-modal decoding. For efficiency, YOLO-World [4] introduces reparameterizable vision–language fusion for real-time open-vocabulary detection. Despite strong ZS performance, VLODs remain sensitive to domain shift, motivating the study of TTA for open-vocabulary ODs.

Test-Time Adaptation for Object Detectors. For ODs, TTA aims to improve robustness under domain shift without requiring source data during deployment. Prior TTA methods for ODs mainly rely on self-training and pseudo-label refinement. These include mean-teacher adaptation with feature alignment [3], stability-aware adapter updates [36], object-level contrastive alignment with selective restoration [1], and single-image adaptation with IoU-guided pseudo-label filtering [30]. Although effective, many TTA methods for ODs rely on heavy augmentations and multi-step updates, which can limit deployment practicality. Moreover, they are designed for conventional closed-set ODs with a fixed vision-only label space and assume source-pretrained detectors. In comparison, VLOD-TTA adapts VLODs in an open-vocabulary setting by leveraging both visual and textual cues. TTAOD-F [9] is the only prior study on TTA for VLODs. It uses a mean-teacher framework with text and visual prompt tuning in a transformer architecture. This design incurs substantial adaptation-time overhead, as it maintains a second detector as teacher, requires an additional forward pass for pseudo-label generation, and uses DINOv2 [28] features for memory-based prediction refinement. In contrast, VLOD-TTA uses an efficient entropy-based objective that does not require a teacher forward pass and generalizes to both CNN- and transformer-based VLODs.

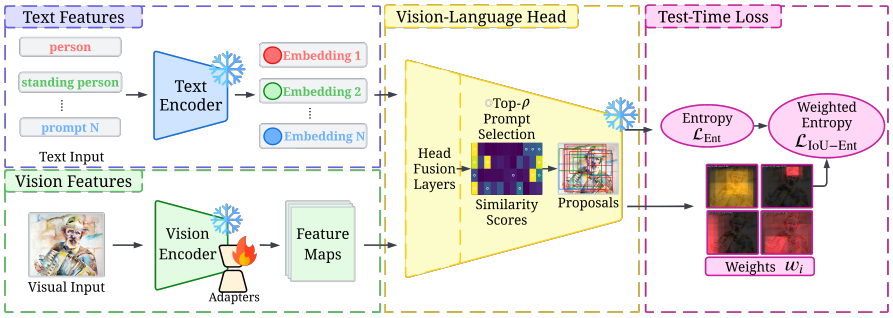


Fig. 3: Overview of VLOD-TTA. Given an input image and a pool of class prompts, the text encoder produces embeddings that interact with region proposals through the vision–language head to compute similarity scores. IPS selects the top-ρ prompts and averages their scores to obtain per-proposal class scores. IWE then combines per-proposal entropy with IoU-based weights to form the adaptation objective.

3 Proposed VLOD-TTA Method

An overview of VLOD-TTA is shown in Figure 3. Our framework consists of two components: IoU-weighted entropy minimization (IWE) and image-conditioned prompt selection (IPS). IWE reweights proposal entropies using local IoU consistency to emphasize reliable regions during adaptation. IPS computes image-conditioned prompt relevance and retains only the most informative prompts. We first introduce the preliminaries and then describe each component in detail.

3.1 Preliminary Definitions

Vision–Language Object Detection. A VLOD couples a visual detector with a text encoder in a shared embedding space. Given an image $X \in \mathbb{R}^{C\times H\times W}$, the visual detector outputs N candidate boxes $B = \{b_i\}_{i=1}^{N}$ and corresponding region features $\{\mathbf{v}_i\}_{i=1}^{N}$, where $\mathbf{v}_i \in \mathbb{R}^d$. For each category name y_k in a label set Y with $|Y| = K$, the text encoder produces an embedding $\mathbf{t}_k \in \mathbb{R}^d$. The similarity score between proposal i and class k is

$$s_{i,k} = \hat{\mathbf{v}}_i^{\top}\hat{\mathbf{t}}_k, \qquad \hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}, \quad \hat{\mathbf{t}}_k = \frac{\mathbf{t}_k}{\|\mathbf{t}_k\|_2}. \tag{1}$$

Final detections are obtained using detector-specific post-processing, such as score thresholding and non-maximum suppression.

Entropy Minimization for ODs. Given class scores $s_{i,k}$ for proposal i and class k, the categorical posterior for proposal b_i is $p_{i,k} = \big[\text{softmax}(s_{i,1}, \ldots, s_{i,K})\big]_k$. The Shannon entropy [31] of proposal i is $\mathcal{H}(\mathbf{p}_i) = -\sum_{k=1}^{K} p_{i,k} \log p_{i,k}$. A standard TTA objective minimizes the average entropy over all proposals:

$$\mathcal{L}_{\text{Ent}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(\mathbf{p}_i). \tag{2}$$

This objective sharpens proposal-level class posteriors and reduces predictive uncertainty.

3.2 IoU-weighted Entropy Minimization (IWE)

VLODs produce a large number of candidate boxes per image. In standard configurations, YOLO-World (YW) produces approximately 8,400 proposals, and Grounding DINO (GD) produces approximately 900. Although post-processing removes most proposals, their spatial structure remains informative. Regions containing many mutually overlapping proposals with the same predicted class are more likely to correspond to true objects, whereas sparse or dispersed proposals are typically less reliable. Standard entropy minimization, such as Tent [35], treats proposals independently and ignores their overlap structure, which can sharpen predictions in unreliable regions and amplify confirmation bias.

To address this limitation, VLOD-TTA introduces IWE, which assigns larger weights to groups of mutually overlapping proposals and smaller weights to isolated proposals. Let $\hat{c}_i = \arg\max_k p_{i,k}$ denote the predicted class of proposal b_i. For each class c, we construct a class-specific IoU graph $G_c = (V_c, E_c)$, whose vertices are proposals with predicted label c. Two vertices u and v are connected by an edge if $\mathrm{IoU}(b_u, b_v) \geq \theta$, where $\theta \in [0, 1]$ is a fixed threshold. Let $\mathcal{C}(i)$ denote the connected component of G_c containing proposal i. We define the weight as $w_i = |\mathcal{C}(i)|^\gamma$, where $|\mathcal{C}(i)|$ is the size of the corresponding connected component and $\gamma \geq 0$ controls the influence of component size. The IWE objective is

$$\mathcal{L}_{\mathrm{IoU-Ent}} = \frac{\sum_{i=1}^{N} w_i \, \mathcal{H}(\mathbf{p}_i)}{\sum_{i=1}^{N} w_i}. \tag{3}$$

The weights depend only on the IoU graph and are treated as constants during backpropagation, so gradients flow only through $\mathcal{H}(\mathbf{p}_i)$. This objective emphasizes spatially coherent proposal groups while reducing the influence of isolated or inconsistent proposals.

3.3 Image-Conditioned Prompt Selection (IPS)

The performance of VLMs is sensitive to prompt wording. A common strategy is to use multiple prompt templates per class and average their text embeddings. Although CLIP reports improved zero-shot accuracy with this approach [29], we find that uniform prompt averaging is often ineffective for VLODs and can even degrade performance. Instead of treating all prompts equally, IPS selects informative prompts for each image and adapts the text representation at test time.

For each class $k \in \{1, \ldots, K\}$, let $\{t_{k,1}, \ldots, t_{k,T}\}$ denote a pool of T prompts with text embeddings $\{\mathbf{e}_{k,t} \in \mathbb{R}^d\}$. Given N region proposals with features $\{\mathbf{v}_i\}_{i=1}^{N}$, we compute prompt-conditioned similarities as $z_{i,k,t} = \hat{\mathbf{v}}_i^\top \hat{\mathbf{e}}_{k,t}$, where $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{e}}_{k,t}$ are ℓ_2-normalized. For each class k and prompt t, we define an image-conditioned compatibility score as $r_{k,t} = \frac{1}{N} \sum_{i=1}^{N} z_{i,k,t}$. This score measures the

average compatibility of prompt t with the proposal features for class k in the current image (see Supp. for theoretical justification). To suppress irrelevant prompts, IPS selects, for each class k, the top-ρ fraction of prompts with the highest compatibility scores $r_{k,t}$. Let $\mathcal{S}_k$ denote the selected prompt indices. The class score for proposal i is then computed by averaging only over the selected prompts, $\tilde{z}_{i,k} = \frac{1}{|\mathcal{S}_k|} \sum_{t \in \mathcal{S}_k} z_{i,k,t}$.

To further align text and region embeddings, we introduce a lightweight residual vector in the text embedding space [39]. Let $\Delta \in \mathbb{R}^d$ be a learnable residual added to each prompt embedding:

$$\tilde{\mathbf{e}}_{k,t} = \frac{\mathbf{e}_{k,t} + \Delta}{\|\mathbf{e}_{k,t} + \Delta\|_2}. \tag{4}$$

Let $s_{i,k}$ denote the base VLOD class score from Equation (1). The final fused score is $g_{i,k} = \lambda \tilde{z}_{i,k} + (1 - \lambda) s_{i,k}$, where $\lambda \in (0, 1)$.

3.4 Model Update

Although VLOD-TTA can in principle adapt any subset of parameters, we keep the base network fixed and update only a small set of adapter parameters [2, 13]. For YW, we insert adapters into the backbone and neck, whereas for GD, we insert adapters only into the text encoder. This distinction reflects their different architectures and pretraining paradigms (see Section 4.3 for details). Let Θ denote all network parameters, with $\Theta = (\Theta_{\text{frozen}}, \Phi, \Delta)$, where Φ represents the adapter parameters and Δ is the residual parameter defined in Equation (4). At test time, we zero-initialize (Φ, Δ) as (Φ_0, Δ_0). Given the final fused score $g_{i,k}$, we retain the top-M proposals ranked by $\max_k g_{i,k}$ to accelerate IoU graph construction. We then optimize (Φ, Δ) for a single adaptation step using the IWE objective. Since adaptation is performed on a single test image, the updated parameters are not expected to generalize across images. We therefore reset (Φ, Δ) to (Φ_0, Δ_0) after each prediction.

4 Results and Discussion

We evaluate VLOD-TTA on a diverse benchmark of domain shifts. We first describe the benchmark setup and TTA baselines. We then analyze the main results across different shift types, followed by ablations on the key design choices of the proposed method.

4.1 Benchmarking VLOD-TTA

We compare ZS inference, four TTA baselines adapted to VLODs, and our VLOD-TTA. The benchmark covers four types of domain shift: texture and style, weather, illumination, and common corruptions. It includes six domain-shift datasets and two corruption benchmarks. We report results for YOLO-World (YW) and Grounding DINO (GD) using mean average precision (mAP) [21].

Datasets. Watercolor, ClipArt, Comic: Watercolor, ClipArt, and Comic [15] are stylized artistic datasets used to evaluate robustness to style shifts. **Cityscapes:** Cityscapes contains urban street scenes from multiple European cities [5]. We use it to study domain shift in driving scenes across geography, weather, and time of day. **BDD100K:** BDD100K is a large-scale driving dataset spanning day and night, multiple cities, and diverse weather conditions [37]. We use it to evaluate robustness under real-world distribution shifts. **ExDark:** ExDark is a low-light OD dataset [23] used to measure robustness under poor illumination. **PASCAL-C and COCO-C:** PASCAL-C and COCO-C [26] corrupt PASCAL-VOC [6] and COCO [21] using 15 common corruption types at five severity levels, following [12], to evaluate OD robustness under common corruptions.

Baselines. Zero-shot: Pretrained VLODs are used directly for inference without any adaptation [4, 22]. **Test-Time Prompt Tuning (TPT):** We adapt TPT [33] from classification to OD by optimizing only text prompt vectors at test time. Candidate boxes are selected based on entropy, and a marginal-entropy objective is minimized over those proposals. **Visual Prompt Tuning (VPT):** Following the TPT pipeline, we optimize only visual prompts [16]. Visual prompting has been shown to be effective for modality adaptation [25]. **DPE:** We adapt DPE [39] to OD by maintaining per-class text and visual caches constructed from high-confidence proposals. At test time, only the residual cache parameters are updated using a marginal-entropy objective combined with a cache-contrastive loss. **Adapter Tuning:** We adapt lightweight bottleneck adapters using the Tent objective [35]. This removes dependence on model-specific prompt parameters and provides a fairer comparison with our method.

Implementation Details. We report AP using the COCO API [21]. Unless stated otherwise, we use YW-Small and GD-Tiny. Each experiment uses a batch size of 1 and a single adaptation step to target real-time deployment settings. We set $\gamma = 1.1$, $\rho = 0.25$, $M = 600$, and $\lambda = 0.3$ for YW, and $\lambda = 0.1$ for GD (see Supp. for hyperparameter sensitivity analysis). Due to architectural differences, we use Conv-Adapters [2] in YW with a reduction factor of 4 and a kernel size of 3, and an MLP Adapter [13] in GD with a reduction factor of $r = 16$ (see Supp. for details). We use $T = 16$ GPT-generated prompts per class. We use ChatGPT-5 to generate textual prompts with the instruction: "Generate 16 prompts for each object category: $<$ *category list* $>$" (see Supp. for examples of prompts).

4.2 Main Results

Texture and style shifts (*Watercolor, ClipArt, Comic*). Results in Table 1 show that VLOD-TTA achieves the best performance on all three stylized datasets for both YW and GD. Among prompt-based baselines, VPT is slightly more effective than TPT on YW, with an average gain of +0.6 AP_{50}, whereas TPT is slightly more effective than VPT on GD by +0.4 AP_{50} on average. This trend is consistent with the adapter placement analysis in Section 4.3. Adapter

Table 1: Detection performance on benchmark datasets. We report mAP, AP_{50}, and AP_{75} for both **YW** and **GD** ODs on six benchmark datasets – Watercolor, ClipArt, Comic, Cityscapes, BDD100K, and ExDark. Best results are in bold.

YOLO-World																		
	Watercolor			ClipArt			Comic			Cityscapes			BDD100K			ExDark		
Method	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
ZS [4]	26.9	47.9	25.9	24.4	40.1	26.2	17.8	29.4	18.8	18.8	31.0	17.9	13.3	22.0	13.4	35.2	64.7	34.6
TPT [33]	27.3	48.5	26.1	24.9	41.3	26.8	18.1	29.9	19.1	18.8	31.1	18.0	13.4	22.2	13.5	35.8	65.1	34.7
VPT [16]	26.9	49.1	25.1	25.0	41.4	26.9	18.3	30.9	19.3	18.9	31.2	18.0	13.5	22.3	13.2	35.8	65.8	34.9
DPE [39]	27.2	48.9	26.3	24.9	41.5	27.1	18.9	31.7	19.8	19.0	31.3	18.0	13.5	22.3	13.3	35.9	66.4	35.1
Adapter [35]	28.3	51.5	26.7	26.9	44.1	27.8	20.8	34.7	21.7	19.1	31.3	18.3	13.7	21.7	13.1	35.8	66.4	35.1
VLOD-TTA	**29.6**	**53.1**	**28.7**	**28.1**	**45.4**	**29.9**	**21.4**	**36.1**	**22.1**	**19.4**	**31.8**	**18.6**	**14.6**	**24.3**	**14.8**	**36.4**	**67.4**	**35.6**
Grounding DINO																		
	Watercolor			ClipArt			Comic			Cityscapes			BDD100K			ExDark		
Method	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
ZS [22]	37.4	62.9	37.6	38.4	58.8	41.8	31.2	52.9	31.5	24.1	38.2	24.5	16.6	28.3	16.2	35.4	66.2	34.5
TPT [33]	37.4	63.1	37.6	38.6	59.1	42.3	31.5	53.6	31.8	24.6	38.6	24.8	16.6	28.4	16.2	35.6	66.5	34.9
VPT [16]	37.2	63.0	37.8	38.6	59.0	41.9	31.1	52.6	31.2	24.0	38.2	24.3	16.7	28.5	16.2	35.1	66.4	34.4
DPE [39]	37.6	63.2	38.1	38.2	59.3	42.0	31.8	53.3	31.7	24.4	38.3	24.4	16.7	28.6	16.3	35.2	66.6	34.3
Adapter [35]	38.4	63.6	39.0	38.6	58.9	41.8	31.7	54.1	32.0	24.6	39.1	24.6	16.8	28.7	16.6	35.7	66.8	34.5
VLOD-TTA	**38.9**	**64.7**	**39.5**	**41.2**	**62.1**	**43.3**	**34.2**	**57.8**	**35.3**	**25.8**	**40.8**	**25.9**	**18.1**	**31.1**	**18.5**	**37.3**	**68.9**	**36.8**

is the strongest prior baseline, improving AP_{50} over ZS by +4.3 on YW averaged across the three datasets, but only by +0.7 on GD. In contrast, VLOD-TTA delivers larger and more consistent gains. Relative to ZS, VLOD-TTA improves YW by an average of +3.3 mAP, +5.8 AP_{50}, and +3.2 AP_{75} across the three datasets. For GD, the corresponding gains are +2.4 mAP, +3.3 AP_{50}, and +2.4 AP_{75}. These results indicate that combining IWE and IPS is particularly effective under appearance shifts that alter texture and visual style while preserving object semantics.

Autonomous driving under varying conditions (*Cityscapes, BDD100K*). Table 1 shows that adaptation on driving scenes is more challenging than on stylized domains, likely because these datasets contain many small objects with weak or sparse proposal overlap (see Supp. for a detailed analysis and ways to mitigate this issue). As a result, standard entropy-based baselines provide only marginal improvements over ZS. On YW, Adapter even underperforms ZS on BDD100K in both AP_{50} and AP_{75}, suggesting that standard entropy can overfit unreliable proposals in crowded scenes. In contrast, VLOD-TTA achieves the best results on both datasets and both detectors. Averaged over Cityscapes and BDD100K, it improves YW over ZS by +1.0 mAP, +1.6 AP_{50}, and +1.1 AP_{75}. For GD, the corresponding average gains are +1.6 mAP, +2.7 AP_{50}, and +1.9 AP_{75}. Although the absolute gains are smaller than on stylized datasets, they remain consistent, showing that VLOD-TTA is effective even in more structured and challenging driving scenarios.

Illumination shift (*ExDark*). Under low-light conditions, VLOD-TTA again achieves the best performance on both YW and GD, as shown in Table 1. On YW, DPE and Adapter are the strongest prior baselines, both reaching 66.4 AP_{50}, which indicates that low-light adaptation benefits from strong prior information. VLOD-TTA further improves over ZS by +1.2 mAP, +2.7 AP_{50}, and +1.0 AP_{75}. On GD, Adapter is the strongest prior baseline, while the gains

Table 2: Detection performance on PASCAL-C. AP_{50} is reported for the **YW** detector on 15 different data corruptions.

	Noise			Blur				Weather			Digital					
Method	**Gauss**	**Shot**	**Impul**	**Defoc**	**Glass**	**Motn**	**Zoom**	**Snow**	**Frost**	**Fog**	**Brit**	**Contr**	**Elast**	**Pixel**	**JPEG**	**Avg**
ZS [4]	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 [33]	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 [16]	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 [39]	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
Adapter [35]	20.5	22.8	23.1	35.9	15.1	29.0	30.2	48.1	52.5	71.1	72.1	45.8	46.5	16.1	36.8	37.7
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**

of the other baselines remain limited. In comparison, VLOD-TTA improves over ZS by +1.9 mAP, +2.7 AP_{50}, and +2.3 AP_{75}. These results indicate that VLOD-TTA improves both detection confidence and localization quality under severe low-light conditions.

Common corruptions. Table 2 reports AP_{50} on PASCAL-C across 15 corruption types using YW. VLOD-TTA achieves the best performance on every corruption and the highest overall average of 39.6 AP_{50}. It outperforms the strongest baseline, Adapter, by 1.9 AP_{50} and improves over ZS by 5.6 AP_{50}. The largest gains over ZS are observed on JPEG Compression (+14.1), Pixelate (+8.9), Impulse Noise (+8.0), Contrast (+7.4), Elastic Transform (+7.4), and Shot Noise (+7.2). These improvements are particularly strong under noise and digital corruptions. Overall, the results show that VLOD-TTA provides broad and consistent robustness across common corruption types.

4.3 Ablation Studies

Contribution of individual components. We ablate the two components of VLOD-TTA on YW across Watercolor, ClipArt, and Comic in Table 3. Adapter serves as the standard entropy-minimization baseline. IWE replaces standard entropy with IoU-weighted entropy, and VLOD-TTA combines IWE with IPS. Compared with Adapter, IWE consistently improves both mAP and AP_{50} across all three datasets. Adding IPS on top of IWE yields further consistent gains. These results show that both components contribute positively, and their combination achieves the best performance.

Adapters versus batch normalization parameters. In the main experiments, we optimize adapter parameters. To show that VLOD-TTA is not restricted to a particular parameter subset, we compare adapter tuning with batch normalization updates under both standard entropy and VLOD-TTA in Table 4. Under standard entropy, the two parameter choices yield similar improvements over ZS on Watercolor, ClipArt, and Comic. Applying VLOD-TTA further improves both and consistently outperforms the corresponding entropy-only baseline. Batch norm with VLOD-TTA nearly matches the adapter-based version, showing that the proposed objective is effective beyond adapters. However, batch norm requires dataset-specific learning-rate tuning to achieve its best performance, for example $1e{-}2$ on Watercolor and $3e{-}2$ on ClipArt, whereas

Table 3: Ablation study on the components of VLOD-TTA. Detection performance on three style-shift datasets.

	Watercolor		ClipArt		Comic	
Method	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}
ZS	26.9	47.9	24.4	40.1	17.8	29.4
Adapter	28.3	51.5	26.9	44.1	20.8	34.7
IWE	29.3	52.6	27.5	44.7	21.2	35.6
VLOD-TTA	**29.6**	**53.1**	**28.1**	**45.4**	**21.4**	**36.1**

Table 4: Adapters vs. batch norm. as adaptation parameters. Detection performance on three style-shift datasets.

	Watercolor		ClipArt		Comic	
Method	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}
ZS	26.9	47.9	24.4	40.1	17.8	29.4
BN	28.4	51.3	26.7	44.1	20.6	34.5
Adapters	28.3	51.5	26.9	44.1	20.8	34.7
VLOD-TTA BN	29.4	52.9	**28.3**	45.3	21.3	**36.1**
VLOD-TTA	**29.6**	**53.1**	28.1	**45.4**	**21.4**	**36.1**

adapters perform well with a single learning rate of $5e{-}3$ across datasets. In the TTA setting, the target domain is unknown at test time, which makes per-dataset learning-rate tuning impractical for real-time detection. We therefore use adapters in the main experiments.

Table 5: Performance and inference cost comparison on COCO-C. We report corruption-group mAP along with total and tuned parameter counts, GPU memory usage, and latency. All experiments are run on an RTX3090 GPU. VLOD-TTA* denotes our method with TTAOD-F-based initialization.

	Corruption Type Avg. mAP ↑				Overall	Inference Cost ↓			
Method	Noise	Blur	Weather	Digital	Avg	Total Params (Mil)	Tuned Params (Mil)	Memory (GB)	Latency (ms/img)
ZS [22]	14.9	11.2	34.7	23.0	20.6	172.9	0.00	1.36	210.6
TTAOD-F [9][1]	21.2	14.3	37.0	31.7	26.0	629.9	**0.08**	11.37	701.9
VLOD-TTA	21.1	15.6	38.5	30.4	26.2	**173.9**	0.89	3.76	**531.6**
VLOD-TTA*	**21.5**	**15.9**	**38.6**	**32.8**	**27.3**	**173.9**	0.95	3.92	567.3

Comparison with TTAOD-F. TTAOD-F [9] is the only prior TTA method designed specifically for VLODs. In Table 5, we compare it with VLOD-TTA on COCO-C using GD, reporting corruption-group mAP together with parameter count and latency (see Supp. for full results). VLOD-TTA improves over ZS on all corruption groups and achieves slightly higher average mAP than TTAOD-F (26.2 vs. 26.0), while using far fewer total parameters (173.9M vs. 629.9M) and lower latency (531.6 vs. 701.9 ms/img), with substantially lower GPU memory usage (3.76 vs. 11.37 GB). TTAOD-F also uses a test-time warm-start strategy that initializes visual prompts by average pooling image tokens from the first test sample. For fair comparison, we additionally evaluate a warm-start variant, denoted VLOD-TTA*, using the same initialization strategy as TTAOD-F. This variant further improves the average mAP to 27.3 and outperforms both TTAOD-F and the default VLOD-TTA across all corruption groups. In addition, VLOD-TTA keeps the tuned parameter budget small (0.89M, or 0.95M for VLOD-TTA*). These results show that VLOD-TTA offers a substantially better accuracy–efficiency trade-off at test time. The latency gap is consistent with the higher adaptation-time cost of TTAOD-F, which requires multiple forward passes and one backward pass, whereas VLOD-TTA uses a single adaptation

[1] TTAOD-F uses batch size 4 (original setting), whereas VLOD-TTA uses batch size 1. Although batch size 1 reduces TTAOD-F memory to 4.43 GB, we observe a deterioration in its performance under this setting.

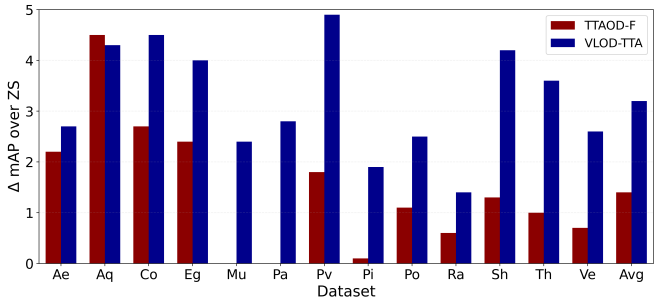


Fig. 4: Improvement over ZS on ODinW-13 using Grounding DINO. VLOD-TTA achieves larger mAP gains than TTAOD-F on 12 of the 13 datasets.

step with one forward pass and one backward pass.

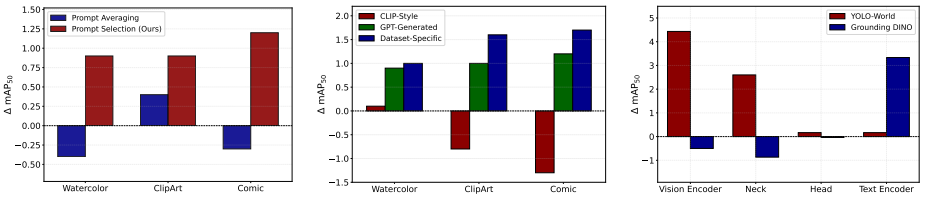
Performance on ODinW-13. ODinW-13 consists of 13 diverse object detection datasets spanning heterogeneous visual domains. As shown in Figure 4, VLOD-TTA improves the ZS baseline from 52.8 to 56.0 mAP, corresponding to a gain of +3.2 mAP, compared with +1.4 mAP for TTAOD-F. Moreover, VLOD-TTA outperforms TTAOD-F on 12 of the 13 datasets. These results demonstrate that VLOD-TTA generalizes effectively across diverse object detection domains.

Scalability to larger VLODs. We further evaluate VLOD-TTA on the larger YOLO-World-Large and Grounding DINO-Big variants. As shown in Table 6, VLOD-TTA consistently improves AP_{50} across all three style-shift datasets. For YOLO-World-Large, the gains range from +3.0 to +4.9 AP_{50}, while Grounding DINO-Big improves by +2.3 to +3.3 AP_{50}. These results indicate that VLOD-TTA remains effective as model capacity increases and is not limited to the smaller detector variants used in our main experiments.

Table 6: Scalability to larger VLODs. We report AP_{50} on three style-shift datasets using YOLO-World-Large and Grounding DINO-Big.

YOLO-World-Large			
Method	**Watercolor**	**ClipArt**	**Comic**
ZS [4]	55.3	50.6	37.9
VLOD-TTA	58.3	53.9	42.8
Grounding DINO-Big			
Method	**Watercolor**	**ClipArt**	**Comic**
ZS [22]	70.5	77.9	64.7
VLOD-TTA	72.8	81.2	67.1

Prompt averaging vs. prompt selection. In Figure 5a, we compare two ways of using language prompts in YW: (i) prompt averaging, which averages multiple templates per class into a single text embedding as in CLIP [29], and (ii) prompt selection, which retains only the most informative prompts for each



(a) Prompt averaging vs. prompt selection.

(b) Prompt generation strategies.

(c) Adapter placement across modules.

Fig. 5: Ablation studies on prompt design and adapter placement. (a) Comparison between prompt averaging and prompt selection. (b) Comparison of prompt generation strategies. (c) Effect of adapter placement across modules. All values report ΔmAP$_{50}$ over three style-shift datasets relative to zero-shot (ZS).

image. Prompt averaging provides only a limited benefit, slightly improving ClipArt while reducing AP$_{50}$ on Watercolor (-0.35) and Comic (-0.30) relative to ZS. In contrast, prompt selection improves AP$_{50}$ on all three datasets, with an average gain of $+1.0$ over ZS. These results show that uniform prompt aggregation offers limited benefit for VLODs, whereas prompt selection provides a more reliable mechanism for adapting text representations at test time.

Variation with different prompt-generation strategies. Our default setting uses dataset-agnostic GPT prompts. We compare this setting with dataset-specific GPT prompts and CLIP-style prompts. The AP$_{50}$ improvement over ZS on Watercolor, ClipArt, and Comic is summarized in Figure 5b. CLIP-style prompts provide a small gain on Watercolor but reduce performance on ClipArt (-0.8) and Comic (-1.3), which we attribute to a pretraining bias toward label-only text prompts. Dataset-agnostic GPT prompts improve over ZS on all three datasets, with gains of $+0.9$ on Watercolor, $+1.0$ on ClipArt, and $+1.2$ on Comic. Dataset-specific prompts yield the largest gains, reaching $+1.0$ on Watercolor, $+1.6$ on ClipArt, and $+1.7$ on Comic, which suggests that domain-specific cues can further improve adaptation. Since such information is unrealistic in the TTA setting, we do not use dataset-specific prompts in the main experiments.

Effect of adapter placement across modules. In Figure 5c, we insert adapters into one module at a time and report the change in AP$_{50}$ relative to ZS, averaged over Watercolor, ClipArt, and Comic. For YW, adapting the vision backbone yields the largest gain ($+4.4$ AP$_{50}$), followed by the neck ($+2.6$), whereas adapting the head or text encoder provides little benefit. This behavior reflects YW's architecture, where the detector and text encoder are largely decoupled and interact only at the final scoring stage. As a result, updating the text encoder mainly affects classification scores and has a limited impact on detection performance. This observation is also consistent with YW pretraining results, where fine-tuning the text encoder can even degrade performance [4]. We therefore omit head adapters in the main experiments to reduce computation.

For GD, the trend is reversed. Adapting the text encoder yields the largest gain ($+3.3$ AP$_{50}$), whereas adapting the vision encoder or neck slightly degrades

Table 7: IoU-weighted pseudo-labeling (IWPL). Integrating IoU-weighting into a mean-teacher pseudo-labeling TTA scheme improves AP on three style-shift datasets.

	Watercolor		ClipArt		Comic	
Method	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**
ZS	26.9	47.9	24.4	40.1	17.8	29.4
Pseudo-label	28.3	50.1	25.9	42.3	19.2	31.9
IWPL	**29.1**	**51.6**	**27.3**	**43.7**	**20.5**	**33.2**

Table 8: Efficiency and performance on YW. We report FPS, trainable parameters (M), and AP$_{50}$.

Method	**FPS↑**	**Params. (M)↓**	**AP$_{50}$↑**
ZS	89	0.00	47.9
TPT	9	1.12	48.5
VPT	18	3.93	49.1
DPE	15	**0.31**	48.9
Adapter	**22**	1.52	51.5
VLOD-TTA	20	1.61	**53.1**

performance. This difference is consistent with GD's early cross-modal fusion design, in which text features influence both localization and classification throughout the detector. Consequently, adapting the text encoder is more effective in GD than in YW. This finding also aligns with GD pretraining, where jointly optimizing the text and vision encoders improves performance [22]. We do not adapt GD's vision encoder in the main experiments because it tends to overfit to a single test image, likely due to its much larger capacity (172M parameters for GD vs. 13M for YW).

IoU-weighted pseudo-labeling. To test whether IoU-based weighting is useful beyond entropy minimization, we integrate it into a standard pseudo-label-based TTA scheme for ODs. Specifically, teacher predictions are used as supervision for student updates, and overlapping teacher boxes are clustered using IoU, with each pseudo-label weighted by its normalized cluster size. As shown in Table 7, IoU-weighted pseudo-labeling consistently improves both mAP and AP$_{50}$ over the standard pseudo-label baseline across Watercolor, ClipArt, and Comic. This suggests that IoU-based proposal weighting is a general principle that can also strengthen pseudo-label-driven adaptation.

Runtime and parameter cost. TTA introduces additional computation beyond ZS inference, so practical deployment requires a favorable accuracy–efficiency trade-off. Table 8 compares throughput, trainable parameters, and AP$_{50}$ on YW. VLOD-TTA is faster than TPT, VPT, and DPE, and only slightly slower than Adapter because of IoU-graph construction. Despite this small overhead, VLOD-TTA uses only 1.61M trainable parameters, achieves the best AP$_{50}$, and provides a favorable efficiency–performance trade-off among the TTA baselines.

4.4 Qualitative Analysis

Figure 6 compares detections from ZS, the Adapter baseline, and VLOD-TTA. The Adapter baseline often sharpens incorrect predictions, for example, the person in the bottom row, which reflects proposal-level confirmation bias under standard entropy minimization. In contrast, VLOD-TTA suppresses isolated proposals and yields more consistent detections with fewer false positives by emphasizing spatially coherent clusters of overlapping proposals. We also observe

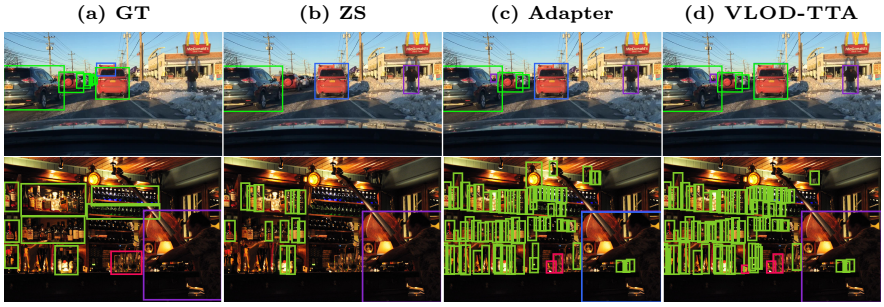


Fig. 6: YW detections across different approaches: Each column corresponds to a different method: (a) GT (Ground Truth), (b) ZS (Zero-Shot), (c) Adapter, and (d) VLOD-TTA. Each color represents a different object category.

cleaner localization, including fewer duplicate boxes. Moreover, VLOD-TTA can detect plausible objects that are missing from the ground truth or refine loosely annotated instances. Although such cases are counted as errors under standard evaluation, they qualitatively suggest improved localization and semantic grounding.

5 Conclusion

TTA has emerged as a practical strategy for improving robustness under domain shift, yet it remains largely unexplored for VLODs. In this paper, we introduce VLOD-TTA, a TTA framework for VLODs that combines IoU-weighted entropy minimization with image-conditioned prompt selection while updating only lightweight adapter parameters. We validate VLOD-TTA across diverse distribution shifts on two representative VLODs, YOLO-World and Grounding DINO. Across these settings, VLOD-TTA consistently outperforms standard TTA baselines and the prior state-of-the-art VLOD TTA method while maintaining low adaptation overhead.

Despite these strong and consistent gains, IoU-weighted entropy can be less effective in scenes dominated by many small objects with limited proposal overlap, such as Cityscapes. Although VLOD-TTA substantially reduces adaptation latency relative to the prior method, it remains slower than zero-shot due to backpropagation at test time. Inspired by recent TTA strategies explored for VLMs, future work will investigate gradient-free adaptation for VLODs to further reduce latency.

Supplementary material. It provides additional theoretical analysis, implementation details, Cityscapes failure case analysis, hyperparameter and design ablations, robustness studies, extended COCO-C and PASCAL-C results, and additional qualitative visualizations.

Acknowledgments. This work was supported in part by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and MITACS.

References

1. Cao, S., Zheng, J., Liu, Y., Zhao, B., Yuan, Z., Li, W., Dong, R., Fu, H.: Exploring test-time adaptation for object detection in continually changing environments (2025), https://arxiv.org/abs/2406.16439
2. Chen, H., Tao, R., Zhang, H., Wang, Y., Li, X., Ye, W., Wang, J., Hu, G., Savvides, M.: Conv-adapter: Exploring parameter efficient transfer learning for convnets (2024)
3. Chen, Y., Xu, X., Su, Y., Jia, K.: Stfar: Improving object detection robustness at test-time by self-training with feature alignment regularization (2023), https://arxiv.org/abs/2303.17937
4. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2024)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (Jun 2010). https://doi.org/10.1007/s11263-009-0275-4, https://doi.org/10.1007/s11263-009-0275-4
7. Farina, M., Franchi, G., Iacca, G., Mancini, M., Ricci, E.: Frustratingly easy test-time adaptation of vision-language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=eQ6VjBhevn
8. Feng, C.M., Yu, K., Liu, Y., Khan, S., Zuo, W.: Diverse Data Augmentation with Diffusions for Effective Test-time Prompt Tuning . In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2704–2714. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00255, https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00255
9. Gao, Y., Zhang, Y., Cai, Z., Huang, D.: Test-time adaptive object detection with foundation model. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=MO4U4mgOoT
10. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022), https://api.semanticscholar.org/CorpusID:238744187
11. Gupta, A., Anpalagan, A., Guan, L., Khwaja, A.S.: Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array **10**, 100057 (2021)
12. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
13. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (2019)
14. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering (2019)

15. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation (2018), https://arxiv.org/abs/1803.11365
16. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision (ECCV) (2022)
17. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
18. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022)
19. Li, S., Ye, M., Zhu, X., Zhou, L., Xiong, L.: Source-free object detection by learning to overlook domain style. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8014–8023 (June 2022)
20. Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P., Zeng, Z.: Clu-cnns: Object detection for medical images. Neurocomputing **350**, 53–59 (2019)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2024)
23. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding **178**, 30–42 (2019). https://doi.org/https://doi.org/10.1016/j.cviu.2018.10.010
24. MA, X., ZHANG, J., Guo, S., Xu, W.: Swapprompt: Test-time prompt adaptation for vision-language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 65252–65264. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/cdd0640218a27e9e2c0e52e324e25db0-Paper-Conference.pdf
25. Medeiros, H.R., Belal, A., Muralidharan, S., Granger, E., Pedersoli, M.: Visual modality prompt for adapting vision-language object detectors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2172–2182 (2025)
26. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019)
27. Mishra, P.K., Saroha, G.P.: A study on video surveillance system for object detection and tracking. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). pp. 221–226 (2016)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning trans-

ferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445
30. Ruan, X., Tang, W.: 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)
31. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948)
32. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8429–8438 (2019). https://doi.org/10.1109/ICCV.2019.00852
33. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: NeurIPS (2022)
34. Varailhon, S., Aminbeidokhti, M., Pedersoli, M., Granger, E.: Source-free domain adaptation for yolo object detection. In: Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVIII. p. 218–235. Springer-Verlag, Berlin, Heidelberg (2025). https://doi.org/10.1007/978-3-031-91672-4_14, https://doi.org/10.1007/978-3-031-91672-4_14
35. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
36. Yoo, J., Lee, D., Chung, I., Kim, D., Kwak, N.: What how and when should object detectors update in continually changing test domains? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23354–23363 (June 2024)
37. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
38. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14388–14397 (2021). https://doi.org/10.1109/CVPR46437.2021.01416
39. Zhang, C., Stepputtis, S., Sycara, K., Xie, Y.: Dual prototype evolving for test-time generalization of vision-language models. Advances in Neural Information Processing Systems **37**, 32111–32136 (2024)
40. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation (2022), https://arxiv.org/abs/2110.09506
41. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., 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)
42. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022)
43. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE **111**(3), 257–276 (2023)