

Comparative Study of Test-Time Adaptation for Object Detection in Closed-Set vs. Open-Vocabulary Scenarios: From SHOT to Grounding DINO

Yitong Qin*, Zhichen Lu, Lihong Wu

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

**Corresponding Author. Email: qytong123@gmail.com*

Abstract. Test-time adaptation of object detection models is a key technique for dealing with data distribution drift in dynamic scenarios such as security checks. This paper systematically compares test-time adaptation (TTA) for closed sets represented by SHOT and TTA for open vocabulary with GroundingDINO as the core for two detection paradigms: closed sets and open vocabulary. The study shows that the closed set TTA has stable performance when dealing with domain offsets of known categories, but has the fundamental defect of category rigidity and inability to recognize new categories. In contrast, the open vocabulary TTA, through dynamic text prompts and cross-modal attention mechanisms, achieves minut-level rapid adaptation to unknown categories and strong generalization ability. Facing challenges in computational efficiency and real-time performance, ultimately, the paper proposes a hybrid hierarchical detection system that combines the advantages of both, providing strong support for a feasible path for the intelligent upgrade of security checks.

Keywords: Adaptation during testing, Open Vocabulary Target detection, Domain Shift, Security inspection, Cross-modal learning

1. Introduction

Object detection is a core task in the field of computer vision and plays a crucial role in security checks, autonomous driving, and other scenarios with extremely high requirements for safety and reliability. Traditional closed-set object detection models rely on predefined categories, independent and identially distributed data assumptions, in the face of emerging new types of contraband in the real world, such as new liquid explosive containers, In response to the challenge of complex and variable imaging environments, such as X-ray machine parameter differences and background interference, the generalization ability is seriously inadequate and there are safety blind spots. To address this challenge, test-time adaptation techniques have emerged, aiming to rapidly fine-tuning the model using only online test data streams to mitigate performance degradation caused by domain shift, while existing research is mostly limited to the closed set category. It fails to fundamentally solve the problem of category expansion.

2. Theoretical divergence between closed-set and open-vocabulary target detection and adaptation challenges during testing

2.1. Differences in detection paradigms between closed sets and open vocabulary scenarios

2.1.1. Static characteristics and generalization bottlenecks of closed set detection

Conventional closed-set object detection models are strictly confined to predefined sets of categories, and their training and testing processes assume that the data follow an independent and homogeneous distribution. This kind of paradigm is highly dependent on large-scale labeled data, and the model parameters are fixed and unrecognizable

Undefined categories during training, in dynamic scenarios such as security checks, where the types of contraband are complex and updated rapidly, closed set models have significantly reduced detection performance when facing new types of contraband, such as new liquid explosive containers, due to the lack of semantic extension capabilities, and there are serious safety blind spots. In addition, when the test environment has domain shifts, such as changes in X-ray machine imaging parameters, When the luggage background becomes complex, the feature extractor of the closed set model is difficult to adapt, resulting in insufficient generalization ability [1].

2.1.2. Dynamic semantic understanding mechanisms for open vocabulary detection

Open vocabulary object detection breaks through the hard limit of the number of categories and uses visual-language cross-modal pre-trained models, such as CLIP, to obtain rich semantic priors. Models represented by GroundingDINO reconstruct the detection task as a text-image matching problem and parse any category description input by the user through a text encoder. For example, "Metal lighter", the visual encoder extracts multi-scale features, and then the cross-modal decoder calculates the image-text similarity and generates the detection box. This mechanism enables the model to recognize unseen categories without retraining, and only update the text prompt library to adapt to new prohibited items, with extremely strong scene scalability.

2.1.3. Different demands of test adaptation (TTA) in both paradigms

Test-time adaptation aims to fine-tune a trained model online using only the test data stream to address domain offsets. In a closed set scenario, the core of TTA is feature distribution alignment, which mitigated domain offsets by minimizing feature statistical differences between the source domain and the target domain, such as MMD and CORAL, while its optimization range is strictly limited to predefined categories. The technical challenges are more complex in the open-vocabulary scenario, where TTA not only needs to handle the distribution variations of visual features but also maintain consistency across the cross-modal semantic space to prevent semantic confusion or modal alignment failures caused by the introduction of new concepts.

2.2. Core challenges of target detection adaptation during testing

2.2.1. Closed set scenarios: domain offset and error accumulation risk

In test-time adaptation of passive domain labels, closed set models typically rely on self-training strategies to generate pseudo-labels, and in complex domain offsets such as haze, low light, or low-dose X-ray imaging, the predictions of the initial model are prone to generating noise. If fine-tuning

is directly done with low-confidence pseudo-labels, the regression results of the error box will accumulate over the iteration process. Ultimately leading to ModelCollapse, especially in high-precision scenarios such as security checks, where the accumulation of errors may cause the missed detection of deadly contraband.

2.2.2. Open vocabulary scenarios: semantic drift with modal alignment

When the open vocabulary model introduces dynamic text prompts, the model needs to adjust the visual-language mapping relationship online. When the text description is ambiguous, for example, only "suspicious device" is input, or there is semantic ambiguity, the model's attention mechanism may be dispersed to irrelevant background areas, resulting in false detections. Additionally, during testing, if the image style of the target domain changes significantly, such as the imaging differences of X-ray machines at different airports, It may disrupt the modal alignment established during the pre-training phase, resulting in the problem of "accurate language description but visual positioning bias".

2.2.3. Algorithmic design trade-offs under real-time constraints

Security scenarios require millisecond-level responses, and complex TTA algorithms, such as online feature distribution alignment and cross-modal attention recalculation, introduce significant computational overhead. For example, the Transformer decoder of GroundingDINO has a large computational load and is difficult to meet both high precision and high frame rate simultaneously on resource-constrained edge devices. Therefore, The algorithm design must strike a balance between adaptive speed, detection accuracy and computational cost to avoid system latency due to over-optimization [2].

3. Adaptive methods for testing closed sets based on SHOT

3.1. Adaptation of the SHOT framework in object detection

3.1.1. Information maximization and pseudo-label generation mechanism

The core idea of the SHOT framework is to freeze the classifier trained in the source domain and adjust only the feature extractor in the target domain. In the task of object detection, this mechanism evolves to optimize the features of the target domain by maximizing the loss of information, which is defined as follows:

$$\mathcal{L}_{IM} = - \sum_{c=1}^C \hat{p}_c \log \hat{p}_c + \sum_{c=1}^C \hat{p}_c \log \hat{p}_c \quad (1)$$

Here, $\hat{p}_c$ represents the average probability of the model predicting that the samples in the target domain belong to category c , where C is the total number of categories. This loss forces the model to make high-confidence predictions by minimizing the conditional entropy, and then maintains the diversity of predictions by maximizing the marginal entropy.

Combined with the self-supervised pseudo-label strategy, the model can iteratively generate pseudo-labels, screen out high-confidence samples, fine-tune them, and experiments on corrupted datasets such as COcoco -C show that by setting dynamic classification thresholds, regression stability verification can effectively filter out noisy pseudo-labels and improve detection reliability.

3.1.2. Multi-level feature alignment strategy

To address X-ray image material texture differences and background interference, the closed set TTA needs to adopt multi-level, different-level functional alignment. For example, domain alignment constraints are applied at different levels of the backbone network, such as the shallow edge layer and the deep semantic layer, and multi-layer adversarial learning or statistical moment matching is used to align the feature distribution between the source domain and the target domain. For high-density areas such as metal in security inspection images, a channel attention mechanism should be designed to be embedded in the feature alignment process to enhance the focus on key areas and suppress background noise interference. This strategy provides strong support for reliable technical paths for stable detection of security inspection X-ray images.

3.2. Limitations and improvement paths of the closed set TTA

3.2.1. The problem of category rigidity

The SHOT class method can only optimize the detection performance of predefined classes and cannot break through the class space during training. In the security check scenario, when faced with undefined contraband, such as new ceramic knives, the model completely fails. This class rigidity problem makes the closed set TTA inherently flawed in open-world applications and unable to meet the rapidly evolving demands of contraband [3].

3.2.2. Error suppression based on distribution alignment

To reduce pseudo-label bias, subsequent studies, such as the STFAR framework, introduced distribution alignment regularization terms like CoreAlign, a technique that models feature distributions through multivariate Gaussian distributions, uses exponential moving averages to estimate the distribution of key features in the target domain online, and then matches the distribution of homologous domains. Experiments on the PascalVOC-C dataset showed that with distribution alignment added, the mAP of the model under strong domain offsets improved by approximately 8.8%, effectively suppressing model degradation.

Table 1. Comparison of performance of the SHOT closed set TTA method on COcoco -C versus Pascal VOC-C datasets (mAP@0.5)

Method	Dataset	Clean Data (mAP)	Corrupted data (no TTA)	Corrupted data (+SHOT)	Corrupted data (+STFAR)
Faster R-CNN	COCO-C	38.9	22.3	26.5	29.1
Faster R-CNN	VOC-C	53.2	35.8	41.7	46.1
YOLOv5	COCO-C	45.1	24.7	28.9	32.8

Note: STFAR

4. Adaptation for groundingdino-driven open vocabulary testing

4.1. A new TTA paradigm in open vocabulary scenarios

4.1.1. Dynamic text prompts are the same as query selection

DINO, through the text-guided query selection module, converts security check rules into text embeddings in real time, calculates similarity to image features, filters out the most relevant regions as the initial query vector of the decoder, and during the adaptation phase in testing, the model can dynamically load new text prompts, such as adding a list of prohibited items without modifying the network weights. The "prompt as detection" paradigm can expand the detection categories, significantly reducing the cost of model updates and shifting towards cost optimization.

4.1.2. Data augmentation of image-text pair relationships

In response to the complex situation often seen in X-ray security images where "one image contains multiple prohibited items of the same kind" or "there is no corresponding target in the image", drawing on the improved strategy in the document, simulate "one-to-many correspondence" and "one-to-space correspondence" relationships in the TTA data stream, and enhance the robustness of the model to complex luggage combinations by randomly combining text cues and image regions, Avoid overfitting the model to a single positive sample pattern, thereby providing strong support for effective methods for stable detection in complex security inspection scenarios.

4.2. Online optimization of cross-modal attention mechanisms

4.2.1. Visual-language feature reweighting

At the TTA stage, computationally heavy image encoders are typically frozen, and only text encoders are fine-tuned, or cross-attention layers in cross-modal decoders are used to narrow the distance between positive sample pairs by contrast loss, which is defined by the formula:

$$\mathcal{L}_{cls} \left(q, k \right) = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^K \exp(q \cdot k_i / \tau)} \quad (2)$$

Where q is the query vector, k_+ is the positive sample text feature, k_i is the negative sample feature, τ is the temperature hyperparameter, push the negative sample pair distance, such partial fine-tuning strategies can both quickly adapt to the target domain distribution and significantly reduce the computational overhead, meeting the real-time requirements of the security check scenario.

4.2.2. Zero-shot and few-shot adaptability

Experimental data showed that in the lighter detection task, by introducing attribute-level text enhancements, such as refining "lighter" to "plastic lighter", the model's recall rate increased significantly from 47.08% to 94.39%, and the open vocabulary model could achieve effective detection based on semantic description alone with very few or even zero target domain labeled samples. It demonstrated a shift from strong zero-shot generalization ability to strong ability.

Table 2. Grounding DINO ablation experiment on X-ray security inspection dataset (lighter detection)

Model configuration	Recall rate (Recall@0.5)	False positive rate (FPR)	mAP@0.5
Standard Model (one-to-one correspondence)	47.08%	0.15	42.5
+ one-to-many correspondence	68.42%	0.12	51.3
+ one-to-many correspondence	75.61%	0.09	58.7
+ Text Attribute Enhancement (Final)	94.39%	0.07	65.2

5. Experimental comparison with security check scenario application verification

5.1. Construction of the benchmark dataset and evaluation metrics

5.1.1. Dataset construction

For a fair comparison of the two paradigms, a mixed benchmark needs to be constructed. The closed set test uses standard damage datasets such as COcoco -C, Foggy Cityscapes, etc. The open vocabulary test requires a self-built X-ray contraband dataset, including basic categories such as metal/plastic lighters, knives, liquid containers, etc., while reserving new types of contraband, such as e-cigarettes, Specific flammable liquids, as the new class, are strictly classified as the basic class and the new class according to the documentation method.

5.1.2. Evaluation dimensions

In addition to the regular mAP@0.5, a new class recall rate and false alarm rate are added. The new class recall rate mainly assesses the model's ability to detect unknown threats, while the false alarm rate measures the model's stability in complex backgrounds, which is crucial for the actual deployment of security checks.

5.2. Comparison of performance between closed set and open vocabulary TTA

5.2.1. Domain offset robustness comparison

In a strong domain shift scenario simulating smog and low-dose X-ray noise, the SHOT based closed set TTA method was stable in base class detection, with mAP improvement of approximately 5.4%, while the new class detection was completely ineffective. GroundingDINO, through semantic generalization of text prompts, still maintained a recall rate of over 85% for the new class. It shows that the open vocabulary paradigm is shifting from superiority in dynamic environments to environmental superiority [4].

Table 3. Comparison of the combined performance of closed set TTA (SHOT) vs. open vocabulary TTA (Grounding DINO)

Evaluation metrics	Closed set TTA (SHOT)	Open word TTA (Grounding DINO)
Base class mAP (known class)	+5.4% (relative boost)	+3.1% (relatively stable)
Novel Recall	0% (undetectable)	85.4%
Semantic extension takes time	Hours (fine-tuning required)	<1 minute (Change Prompt)
Strong noise robustness (mAP reduction)	12.1% (decrease)	8.7% (down)

5.2.2. Comparison of semantic expansion efficiency

The closed set model requires re-collecting data to add new categories, which takes hours to days like full fine-tuning. The open vocabulary model only needs to update the text prompt library, such as modifying the configuration file, to achieve minute-level policy iteration. Such efficiency advantages enable security inspection systems to respond quickly to policy changes, such as adding prohibited items, and have significant engineering application value to shift to application value. This provides a strong support for the dynamic update of the security inspection system with flexible means.

5.3. Actual deployment challenges and optimization recommendations

5.3.1. Computational overhead analysis

The Transformer decoder of GroundingDINO has a large computational load and faces latency challenges when deployed on edge devices. Dynamic resolution scaling is recommended, with short edges controlled at 480-800px and query vector pruning techniques to reduce computational complexity while maintaining accuracy. Additionally, an asynchronous TTA mechanism can be designed. Model updates are carried out during system idle periods to avoid affecting real-time detection throughput, which cannot solve the full computational burden [5].

5.3.2. Hybrid architecture outlook

A single closed set or open vocabulary TTA has its limitations. A feasible future direction is to build hierarchical detection systems. Base classes, such as common knives and lighters, are processed by lightweight closed set TTA models, such as SHOT+, to ensure high speed and stability. Difficult new classes or low-confidence regions trigger open vocabulary modules, such as GroundingDINO. For verification, this hybrid architecture, which combines the stability of closed set TTA with the flexibility of open vocabulary, is a feasible path for upgrading security inspection to an upgraded path [6].

6. Conclusion

This paper systematically compares and analyzes the technical paths of object detection adaptation during testing in closed sets and open vocabulary scenarios. The study shows that closed set TTA (such as SHOT) shifts to feature alignment and pseudo-label optimization and performs stably when dealing with domain offsets of known categories, while its inherent category rigidity problem gives the model prospects for application in the open world. This cannot be fundamentally solved by the open vocabulary TTA (represented by Grounding DINO), which endows the model with powerful zero-shot semantic extension capabilities through dynamic text prompts and cross-modal alignment, achieving rapid response to unknown threats, while its computational overhead and real-time constraints, It is an urgent challenge to be addressed in actual deployment. Overall, the future development direction is not to choose one or two, but to build a hierarchical hybrid detection architecture that combines the efficient stability of closed set TTA with the flexible generalization of open vocabulary TTA, thereby achieving the unity of high precision, high robustness, and high adaptability in critical mission scenarios such as security checks.

References

- [1] V. Singh, D. Ganguly, W. Chen, S. Sahoo, S. Sankar, B. Zhang, M. Harir, S. Wang, O. Zafar, C. Gagné, and V. Chaudhary, "reliability-gated source anchoring for continual test-time adaptation, " arXiv preprint arXiv: 2605.14063, May 2026.
- [2] W. Huang, H. Qiao, K. Zhang, Z. Ye, Y.-M. Shang, and X. Fu, "When normality shifts: risk-aware test-time adaptation for unsupervised tabular anomaly detection, " arXiv preprint arXiv: 2605.10242, May 2026.
- [3] Z. Li, J. Xue, W. Wang, A. Liu, X. Gao, Z. Sun, and Q. Li, "AGC: adaptive geodesic correction for adversarial robustness on vision-language models, " arXiv preprint arXiv: 2605.15584, May 2026.
- [4] Y. Wang and T. He, "Warp-as-history: generalizable camera-controlled video generation from one training video, " arXiv preprint arXiv: 2605.15182, May 2026.
- [5] L. Li, W. Wei, Y. Fu, W. Zhang, Y. Zhang, and W. Zhao, "BSONet: deep learning solution for optimizing image quality of portable backscatter imaging systems, " IEEE Trans. Comput. Imaging, to be published, 2026. (Submitted Feb. 2026; accepted May 2026)
- [6] J. Cani, C. Diou, S. Evangelatos, V. Argyriou, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, and G. Th. Papadopoulos, "Illicit object detection in X-ray imaging using deep learning techniques: a comparative evaluation, " arXiv preprint arXiv: 2507.17508, Jul. 2025.