TPA: TEST-TIME PROTOTYPE ADAPTATION FOR INDUSTRIAL INSPECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot industrial inspection methods adapt to new tasks from a small set of normal samples, but most hold their normal reference fixed after deployment, so valid batch-to-batch variation in material, finish, or illumination is scored as deviation. Methods that do update a reference online index it by feature proximity, which discards the positional correspondence that fixed-viewpoint acquisition already provides. Test-time prototype adaptation (TPA) keeps the feature encoder frozen and maintains one normal prototype token and one update counter per spatial location. Dual anomaly scoring (DS) combines normalized feature distance with spatial-context consistency and admits only the lowest-scoring patches; adaptive prototype update (AP) writes each admitted patch back to its own location under a counter-weighted mean; and dynamic attention weighting (DAW) reweights the anomaly map by the consistency each location has shown across the stream. Indexing by location is what makes contamination bounded per location, and it is also what ties the method to an approximate spatial alignment assumption. Across three anomaly detection benchmarks and one few-shot segmentation benchmark, adaptation adds 0.09 ms per image and 11 GFLOPs over the frozen reference while recovering 5.6 points of image-level AUROC over the static configuration, and image-level AUROC varies by at most 0.7 points across five test-stream orders. The counter-weighted mean converges as the update count grows, so it favors stability over long-horizon tracking; evidence for the regime beyond the evaluated stream lengths is not established here.

1 INTRODUCTION

Industrial visual inspection underpins defect detection and quality control in manufacturing (Pang et al., 2021), but learning-based inspection needs large labeled datasets that are difficult to collect because defects are rare and highly variable (Bergmann et al., 2019). Few-shot and zero-shot approaches lower this cost by adapting to a new task from a few normal samples, or with no target-domain training data (Feng et al., 2023; Liu et al., 2025; Li et al., 2024b). Fig. 1(a) groups three representative paradigms: distance-based methods that score deviation from normal prototypes, typically by ℓ_2 feature distance (Li et al., 2024b; Roth et al., 2022); generative methods that detect anomalies from reconstruction or likelihood deviations (Dai et al., 2025; Zhang et al., 2025b; Gudovskiy et al., 2022); and cross-modal methods that adapt vision-language models through prompt optimization or fine-tuning (Cao et al., 2025; Zhou et al., 2024).

Most approaches keep a static reference after deployment, leaving the encoder and the normal reference unchanged as test data arrive (Pang et al., 2021; Roth et al., 2022; Tao et al., 2022). The limitation addressed here is not the frozen encoder itself, but the fixed prototypes used with it: when material, surface finish, or illumination varies across production batches, a transferable encoder still compares new samples against an outdated reference and raises false alarms (Wan et al., 2022), and the effect grows under shifts in illumination, background, or viewpoint (Liang et al., 2024; Du et al., 2023; Sui et al., 2025; Zhang et al., 2024). Updating the encoder or a set of learnable prompts closes part of that gap but adds data, computation, and maintenance cost (Zhou et al., 2022). A cheaper strategy keeps the encoder frozen and updates compact prototypes as the stream evolves (Wang et al., 2021; Yuan et al., 2023); industrial methods that already do this (McIntosh & Albu, 2024; He et al., 2025; Wei et al., 2024) index their references by feature proximity, as unordered banks or nearest-node searches, so a patch is matched against whichever stored features resemble it. Under

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.

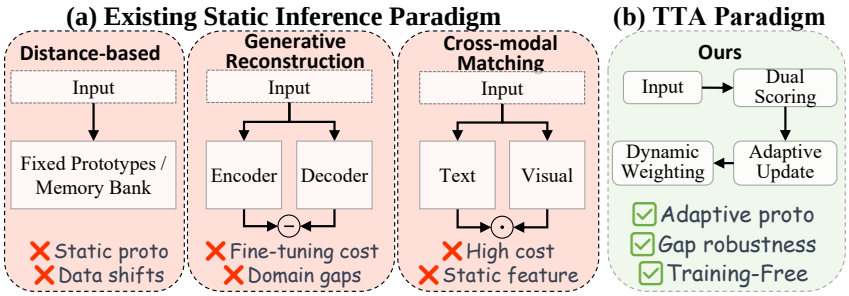


Figure 1: Comparison of representative static inspection paradigms with the proposed test-time adaptation (TTA) paradigm. The deployment limitations illustrated in (a) include (I) fixed normal references, (II) sensitivity to data drift, and (III) the cost of gradient-based fine-tuning. The proposed approach in (b) instead provides (I) online prototype updates, (II) adaptation to changes in the normal data distribution, and (III) gradient-free inference-time adaptation while retaining a frozen encoder.

fixed-viewpoint acquisition, where the same patch index denotes the same product region across images, that discards the positional correspondence the acquisition already provides.

Because a mistaken dense update affects its matched location, patch admission requires the anomaly-score gate developed in Section 2.2.

Test-time prototype adaptation (TPA) implements that gate as three stages sharing one score. Dual anomaly scoring (DS) combines normalized feature distance with spatial-context consistency to decide which patches are safe to admit; adaptive prototype update (AP) writes admitted patches into the matched prototype at their own location under a per-location counter; dynamic attention weighting (DAW) reweights the anomaly map by the consistency each location has shown across the stream. Only the prototypes and the attention state change, so the encoder stays frozen and no gradient-based optimization occurs at test time. The main contributions are summarized as follows:

- We formulate online prototype adaptation for industrial inspection. TPA updates normal reference prototypes at test time while keeping the feature encoder frozen.
- TPA combines DS for reliable patch selection, AP for spatial prototype refinement, and DAW for temporally informed anomaly localization.
- Under the evaluated settings (IAD: 2-shot; IFSS: standard 1/5-shot protocols), experiments on four benchmarks examine detection performance, computational efficiency, robustness to distribution shifts, and sensitivity to test-stream order.

Related work is summarized in Section S2 of the supplementary material, which positions TPA against online and continual industrial anomaly detection, test-time adaptation for anomaly detection under distribution shift, few-shot anomaly detection, and generic test-time adaptation.

2 Methodology

For clarity, the main symbols used in this section are summarized in Supplementary Table S5.

2.1 Overview

Given an input batch $\mathbf{X}$, TPA produces pixel-level anomaly maps and image-level predictions through $\hat{\mathcal{A}}, \hat{\mathrm{E}} = f_{\mathrm{TTA}}(\mathbf{X})$. A pretrained and frozen feature extractor $\mathcal{E}$ maps B images $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times 3}$ to D-dimensional patch embeddings

$$\boldsymbol{x} = \mathcal{E}(\mathbf{X}) \in \mathbb{R}^{BN\times D}, \tag{1}$$

with N patches per image, and TPA maintains C spatially aligned prototype maps $\mathbf{P} = \{\boldsymbol{p}^1, \ldots, \boldsymbol{p}^C\}$, each $\boldsymbol{p}^c \in \mathbb{R}^{N\times D}$, whose token $\boldsymbol{p}_i^c$ is the normal feature at location i. From a normal support set $\mathcal{S} = \{(\mathbf{X}_{g,k}, g)\}$, where g indexes the object category and k the support shot,

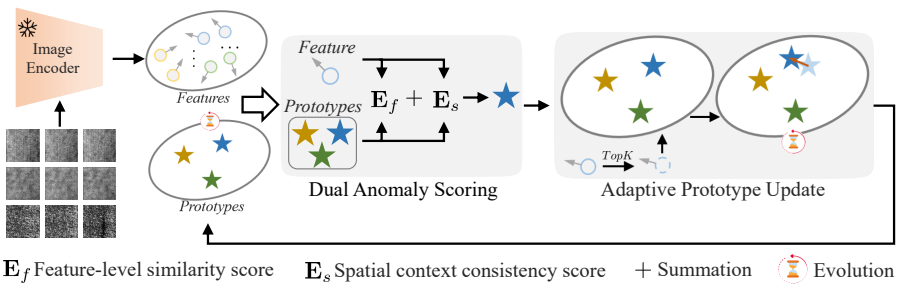


Figure 2: Prototype-oriented online adaptation loop of TPA. A frozen encoder maps each incoming image to patch features that are compared with the current prototype maps; dual anomaly scoring (DS) selects reliable low-anomaly patches, and adaptive prototype update (AP) refines the matched prototype within the same inference loop.

the frozen encoder extracts dense features from each support image and each resulting map is retained as one spatial prototype, $\boldsymbol{p}_i^{g,k,0} = \mathcal{E}(\mathbf{X}_{g,k})_i$, so the initial maps preserve the spatial layout and normal appearance of every support image and $C = \sum_g K_g$ when categories and shot counts are known. These maps are then updated online as the test stream is processed: as summarized in Figs. 2 and 3, DS identifies high-confidence normal patches, AP refines the matched prototype through a counter-weighted running mean, and DAW reweights the anomaly map.

2.2 Design Rationale

Feature-only scores can mistake local variation for defects, static references lag valid batch variation, and single-frame anomaly maps can be noisy. Accordingly, DS combines feature and spatial-context evidence to select low-scoring patches, AP updates their matched spatial tokens, and DAW reweights the map using historical consistency. Patch-level gating is particularly important for a dense reference: a token at location i pools only patches admitted at that location, so an erroneous admission remains local but can carry substantial weight while its update counter is small. This motivates the anomaly-score selection rule and its explicit early-contamination limitation; the full counter argument is in Section S3.2 of the supplementary material.

2.3 Dual Anomaly Scoring (DS)

As shown in Fig. 2, DS combines the feature-level distance $\mathbf{E}_f$ with the spatial-context consistency score $\mathbf{E}_s$. $\mathbf{E}_f$ measures the deviation of each input patch from each prototype map:

$$\{\mathbf{E}_f\}^c = \| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{p}^c}{\|\boldsymbol{p}^c\|_2} \|_2 \in \mathbb{R}^{BN}. \tag{2}$$

Because $\mathbf{E}_f$ ignores neighborhood information, we complement it with $\mathbf{E}_s$. For a Vision Transformer backbone (Dosovitskiy et al., 2021), the N patch tokens are rearranged into an $L \times L$ grid with $L = \sqrt{N}$, and

$$\{\mathbf{E}_s\}^c = \frac{1}{|\mathcal{N}|} \sum_{\Delta i \in \mathcal{N}} \left\| \frac{\boldsymbol{x}_{i+\Delta i}}{\|\boldsymbol{x}_{i+\Delta i}\|_2} - \frac{\boldsymbol{p}^c_{i+\Delta i}}{\|\boldsymbol{p}^c_{i+\Delta i}\|_2} \right\|_2, \tag{3}$$

where $\mathcal{N} = \{\pm 1, \pm L, \pm(L-1), \pm(L+1)\}$ holds the eight adjacent offsets. Averaging over $|\mathcal{N}|$ puts $\mathbf{E}_s$ on the same scale as $\mathbf{E}_f$, so α weights two comparable quantities. Spatially clustered anomalous patches give a large $\mathbf{E}_s$, whereas an isolated response surrounded by normal patches gives a smaller one.

Only the prototype map with the lowest combined score is retained for anomaly evaluation:

$$\boldsymbol{c} = \arg\min_{c \in \mathcal{C}} \left\{ \mathbf{E}_f^c + \alpha \mathbf{E}_s^c \right\}, \tag{4}$$

$$\mathbf{E}^{\boldsymbol{c}} = \mathbf{E}_f^{\boldsymbol{c}} + \alpha \mathbf{E}_s^{\boldsymbol{c}} \in \mathbb{R}^{BN}. \tag{5}$$

The feature term preserves sensitivity to an isolated deviant patch, while the spatial term moderates its score when neighboring patches still match the prototype. Both terms are evaluated for each

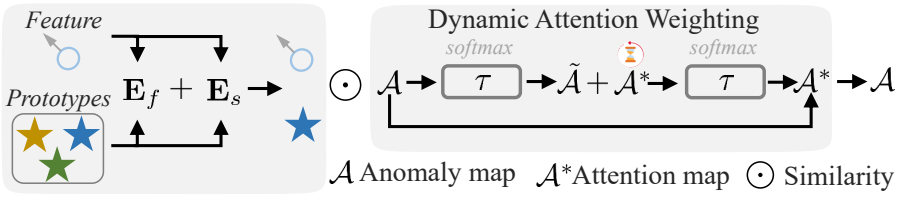


Figure 3: Attention-guided result refinement in TPA. The base anomaly map $\mathcal{A}$ is computed from the distance between the patch features and the matched prototype map. Dynamic attention weighting (DAW) derives instantaneous attention from feature-prototype similarity, combines it with historical attention through softmax normalization, and reweights $\mathcal{A}$ to obtain the final anomaly map.

prototype; the default $\alpha = 1.0$ weights them equally. Although normalized ℓ_2 and cosine distance are monotonically related for an individual feature pair, changing one term's scale in the combined score can change prototype and admitted-patch rankings; Section S3.1 of the supplementary material gives the derivation, and Table 5 reports the ablation.

2.4 Adaptive Prototype Update

As shown in Fig. 2, adaptive prototype update (AP) modifies the normal reference online by matching each image to one prototype map and updating only that map. Because anomalous regions are sparse, most patches remain normal even in a defective image, so selection targets the patches least likely to carry defect evidence. The binary mask $\mathbf{M}$ marks patch $\boldsymbol{x}_i$ as normal when its score falls below the $\lfloor BN\delta \rfloor$-th smallest value of $\mathbf{E}^{\boldsymbol{c}}$ in the batch:

$$\mathbf{M}_i = \mathbf{1}\{\mathbf{E}^{\boldsymbol{c}}_i < \mathbf{E}^{\boldsymbol{c}}_{\text{top}\lfloor BN\delta \rfloor}\}. \tag{6}$$

Selection uses the matched-prototype DS score and admits only a small fraction of patches; early contamination remains possible (Section 2.2).

Selected locations are updated through a counter-weighted running mean, where each map keeps a patch-wise counter $n^{\boldsymbol{c}}_i$, initialized to 1:

$$\boldsymbol{p}^{\boldsymbol{c}}_i \leftarrow \frac{n^{\boldsymbol{c}}_i}{n^{\boldsymbol{c}}_i + 1}\boldsymbol{p}^{\boldsymbol{c}}_i + \frac{1}{n^{\boldsymbol{c}}_i + 1}\boldsymbol{x}_i \quad \forall \mathbf{M}_i = 1, \tag{7}$$

$$n^{\boldsymbol{c}}_i \leftarrow n^{\boldsymbol{c}}_i + 1 \quad \forall \mathbf{M}_i = 1. \tag{8}$$

Only admitted patches at location i increment $n^{\boldsymbol{c}}_i$, as motivated in Section 2.2. The resulting step size $1/(n^{\boldsymbol{c}}_i + 1)$ decreases over time, favoring stability over long-horizon tracking; Section S3.2 of the supplementary material discusses this trade-off.

2.5 Dynamic Attention Weighting

Dynamic attention weighting (DAW) refines the base anomaly map by the historical stability of each spatial location, assigning greater weight to locations that have remained consistent with their matched prototype. As illustrated in Fig. 3, the base map is the ℓ_1 distance between the normalized input and prototype features,

$$\mathcal{A}_i = \| \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2} - \frac{\boldsymbol{p}^{\boldsymbol{c}}_i}{\|\boldsymbol{p}^{\boldsymbol{c}}_i\|_2} \|_1, \tag{9}$$

which, unlike the normalized ℓ_2 distance used in DS, is more sensitive to component-wise deviations and retains responses to small defects.

DAW combines instantaneous similarity with cumulative historical consistency. The instantaneous term is the cosine similarity between each patch and its matched prototype, normalized by a softmax over all patch indices:

$$\tilde{\mathcal{A}}^{*,t}_i = \text{softmax}_i \left(\left\langle \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2}, \frac{\boldsymbol{p}^{\boldsymbol{c}}_i}{\|\boldsymbol{p}^{\boldsymbol{c}}_i\|_2} \right\rangle \right). \tag{10}$$

Here t denotes an online test step after prototype initialization, one step per incoming test image, with a separate history for each matched prototype. The state combines the instantaneous attention with the previous state and is renormalized:

$$\mathcal{A}_i^{*,t} \leftarrow \mathrm{softmax}_i \left(\tilde{\mathcal{A}}_i^{*,t} + \mathcal{A}_i^{*,t-1} \right). \tag{11}$$

The accumulated quantity is agreement with the normal prototype, not anomaly evidence: locations consistent with their prototype contribute repeatedly to the sum, whereas transient defects, noise, and blur do not. Relative to overwriting the history at every step, the recurrence retains information from earlier observations and reduces the influence of a single activation. Eq. 12 reweights the base map, and Eq. 13 gives the image-level decision:

$$\mathcal{A}_i \leftarrow \mathcal{A}_i^* \times \mathcal{A}_i, \tag{12}$$

$$\mathrm{E} = \begin{cases} 1, & \text{if } \sum_{i=1}^{BN} \mathbf{E}_i^{\boldsymbol{c}} > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The decision E is used only for IAD; for IFSS the dense map $\mathcal{A}$ is directly treated as the segmentation score map and Eq. 13 is not used. Every image-level metric reported here is AUROC, which ranks scores and requires no threshold, so τ is specified for deployments that need a binary decision but is not exercised by any reported result. Location-wise accumulation assumes industrial acquisition from a fixed viewpoint, under which a given ViT patch index approximately denotes the same normalized product region across images; Section S3.3 of the supplementary material states what fails when this assumption breaks. The complete test-time procedure is given in Algorithm S1 of Section S3.0 of the supplementary material, alongside the information-flow diagram in Figure S1.

3 Experiments

We evaluate TPA on three industrial anomaly detection benchmarks, MVTec AD, VisA and BTAD, scored by image-level and pixel-level AUROC, and on one industrial few-shot segmentation benchmark, FSSD-12, scored by mIoU. The comparison covers five vision-language detectors for IAD—AA-CLIP (Ma et al., 2025), AdaCLIP (Cao et al., 2025), Bayes-PFL (Qu et al., 2025), MVFA-AD (Huang et al., 2024) and WinCLIP (Jeong et al., 2023)—plus the test-time training method TTT4AS (Costanzino et al., 2024), and five IFSS baselines: CANet (Zhang et al., 2019), QGNet (Liu et al., 2021), CPANet (Feng et al., 2023), MAPTNet (Huang et al., 2025) and PTNet (Liu et al., 2025); the distribution-shift comparison additionally includes ATTA (Gao et al., 2023). Their data regimes, including the caveats that apply to reporting TTT4AS outside its designed regime, are described in Section S4 of the supplementary material, together with datasets, splits, implementation details, the evaluation protocol, the statistics over seeds and the additional analyses; the default configuration is given in Supplementary Table S6. Supplementary Table S8 compares three method properties: gradient-free inference, class-agnostic operation, and test-time adaptation.

3.1 Performance

We report results for both IAD and IFSS. Unless stated otherwise, TPA uses two IAD support samples per class and the IFSS 1/5-shot protocol; the baseline data settings are listed in Table 1.

Industrial anomaly detection. Table 1 shows that TPA is competitive under the 2-shot setting, including against several baselines that use full normal training data; the Shot row states each method's data regime, which the comparison depends on. On MVTec AD, TPA reports 92.4% image-level and 95.2% pixel-level AUROC, the highest values in the table and ahead of the full-data baselines Bayes-PFL (92.3) and AdaCLIP (89.2). On VisA it attains the highest reported image-level AUROC and matches the highest pixel-level AUROC (95.6, tied with Bayes-PFL), and on BTAD it reaches 94.5/93.9, the highest image-level value and within 0.1 of the best pixel-level result. Averaged over the three datasets the method obtains 91.0/94.6, ahead of every baseline on both metrics, with standard deviations of at most 0.03 over three seeds. These results suggest that online prototype refinement can retain competitive performance when only two normal support samples per class are available. Online zero-shot methods built on CLIP have reported higher absolute numbers, with RareCLIP (He et al., 2025) reaching 98.2 image-level AUROC on MVTec AD and 94.4 on VisA; those results are not comparable with this table, since they use a different encoder, require no support images, and follow that method's own protocol instead of the 2-shot setting evaluated here.

Table 1: Comparison across industrial anomaly detection (IAD) and industrial few-shot segmentation (IFSS). All metrics are higher-is-better. The best result per row is shown in bold and the second-best is underlined. Parentheses give standard deviations over three seeds. The Shot row reports each method's data setting; WinCLIP uses zero-shot inference ($K = 0$), and TTT4AS inherits its regime from an upstream detector. Cells marked † were run by us; all other baseline cells are quoted. TTT4AS reports neither mIoU nor any of VisA, BTAD and FSSD-12, and presumes anomalous test images, so its marked cells lie outside its designed regime.

Domain	Metric	Dataset	AA-CLIP	AdaCLIP	Bayes-PFL	MVFA-AD	WinCLIP	TTT4AS	Ours
		Shot	2	full	full	2	0	wrapper	2
IAD	AUROC-image, AUROC-pixel	MVTec AD (Bergmann et al., 2019)	85.9, 91.0	89.2, 88.7	92.3, 91.8	86.6, 84.9	91.8, 85.1	90.1, 93.2	**92.4**(±0.01), **95.2**(±0.02)
		VisA (Zou et al., 2022)	78.4, 93.4	85.8, 95.5	87.0, **95.6**	76.5, 93.4	78.1, 79.6	83.7†, 94.0†	**87.2**(±0.01), **95.6**(±0.00)
		BTAD (Mishra et al., 2021)	88.0, 92.8	88.6, 92.1	93.2, **93.9**	94.3, 90.1	68.2, 72.6	91.2†, 91.8†	**94.5**(±0.02), **93.9**(±0.01)
		Average	84.1, 92.4	87.9, 92.1	90.8, 93.8	85.8, 89.5	79.4, 79.1	88.3, 93.0	**91.0**(±0.00), **94.6**(±0.03)
Domain	Metric	Dataset	CANet	QGNet	CPANet	MAPTNet	PTNet	TTT4AS	Ours
		Shot	1, 5	1, 5	1, 5	1, 5	1, 5	wrapper	1, 5
IFSS	mIoU	FSSD-12 Fold-0	38.6, 42.4	47.7, 47.7	44.0, 45.4	49.2, 51.2	48.5, 48.6	47.8†, 49.0†	**51.1**(±0.03), **51.8**(±0.02)
		FSSD-12 Fold-1	37.3, 39.1	44.5, 45.7	42.6, 42.7	38.6, 43.9	47.0, 49.3	46.6†, 49.4†	**49.6**(±0.01), **50.4**(±0.03)
		FSSD-12 Fold-2	37.8, 38.5	40.8, 42.6	32.8, 34.4	38.3, 38.8	42.6, 43.1	42.2†, 42.5†	**43.1**(±0.03), **43.5**(±0.01)
		FSSD-12 Average	37.9, 40.0	44.3, 45.3	39.8, 40.8	42.0, 44.7	46.0, 47.0	44.4†, 45.0†	**47.9**(±0.02), **48.6**(±0.02)

(a) Initialization ($t = 0$). (b) Test time ($t = 800$).

Figure 4: t-SNE (Van der Maaten & Hinton, 2008) projection of patch prototypes for four classes: bottle (green), cable (red), grid (blue), and carpet (purple). The step index t counts test images.

Industrial few-shot segmentation. In the IFSS comparison of Table 1, TPA reports the highest mIoU in every fold under both the 1-shot and 5-shot protocols, with fold averages of 47.9% and 48.6% (per fold: 51.1/49.6/43.1 and 51.8/50.4/43.5 on FSSD-12 Fold-0/1/2). The closest competitors are MAPTNet on Fold-0 (49.2/51.2) and PTNet on Fold-1 (49.3 at 5-shot), and the margin is widest on Fold-2, where the next-best result is 42.6/43.1; 5-shot adds only 0.7 mIoU over 1-shot on average. Together with the IAD results, this indicates that dense prototype adaptation applies to both evaluated low-data settings. Within the scope of Supplementary Table S8, TPA is the only method marked as satisfying all three listed properties—gradient-free inference, class-agnostic operation, and test-time adaptation—whereas AA-CLIP, SeaS, AdaCLIP and PromptAD are not class-agnostic under the definitions of the caption, MuSC and UniVAD do not update their representations during inference, and WinCLIP and LogSAD are static at test time. Table 2 places the accuracy-speed trade-off against WinCLIP and AdaCLIP: at 721 GFLOPs per image TPA uses 7.7 times fewer operations than AdaCLIP and 36.6-fold fewer than WinCLIP, and reaches 21.3 FPS at the default batch size of 12. This throughput is not the fastest available for industrial inspection, since lightweight retrieval pipelines reach higher frame rates with smaller backbones or lower-resolution inputs, and the cost is dominated by the 768×768 input, the DINOv3 encoder, dense patch-prototype matching, online updates and DAW. Fig. 4 shows t-SNE projections of patch prototypes at $t = 0$ and $t = 800$: the displayed classes overlap substantially at initialization and appear more compact and separated later, consistent with progressive refinement, although t-SNE alone does not establish improved generalization. Qualitative examples are in Fig. 5, and the full protocol in Sections S4 and S5 of the supplementary material.

Table 2: Comparison of model size, computation, throughput, and IAD performance.

Method	GFLOPs ↓	Params (M)	FPS ↑	I/P AUROC ↑
WinCLIP (Jeong et al., 2023)	26359	**78**	0.6	79.4, 79.1
AdaCLIP(Cao et al., 2025)	5570	528	5.5	87.9, 92.1
Ours	**721**	300	**21.3**	**91.0**, **94.6**

Table 3: Inference overhead comparison at 768×768, batch=12.

Mode	ms/img ↓	FPS ↑	Peak Mem (MB) ↓	GFLOPs ↓
w/o TTA	46.86	21.34	1328.02	710.0
w/TPA	46.95	21.30	1328.04	721.0

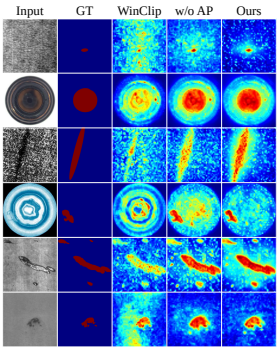
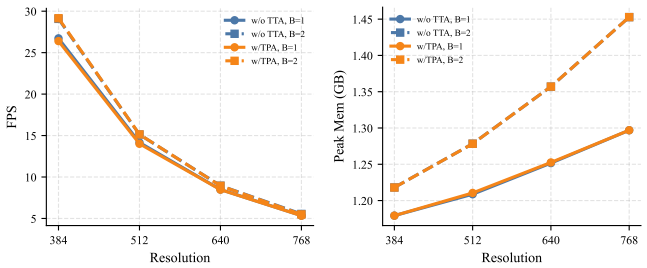


Figure 5: Qualitative comparison. Columns from left to right show the input image, ground truth (GT), WinCLIP, the ablated variant without AP, and TPA



(a) Wall-clock FPS across resolutions and batch sizes.

(b) Peak memory across resolutions and batch sizes.

Figure 6: Computational-overhead diagnostics. The curves use batch sizes of 1 and 2 to show scaling trends, whereas Table 7 reports the default batch size of 12.

Table 4: Robustness comparison under distribution shifts. Results are macro-averaged over the five shift families and severity levels of Table S4 in the supplementary. Corruptions are synthesized here, so every row was run by us (†).

Method	Image-AUROC ↑	Pixel-AUROC ↑	mIoU ↑
WinCLIP†	80.6	83.2	29.4
AdaCLIP†	85.4	88.5	36.8
MuSC†	87.2	90.4	39.6
ATTA†	84.9	89.1	37.2
TPA (ours)†	**89.3**	**91.7**	**41.5**

Table 5: Distance-metric ablation for DS and DAW on the IAD average, defined as the mean of the image-level and pixel-level AUROC averages of Table 1. Higher is better and the best value is shown in bold.

DS metric	DAW metric	IAD Avg.
ℓ_2 on normalized features	ℓ_1 (ours)	**92.8**
ℓ_2 on normalized features	ℓ_2	92.4
Cosine distance	ℓ_1	92.1
Cosine distance	ℓ_2	91.7

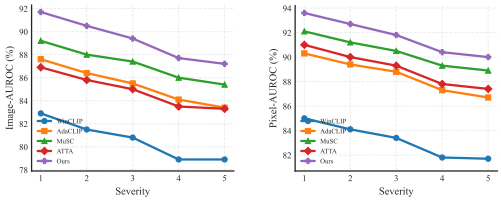
3.2 Extended Ablation Study

3.2.1 Deployment diagnostics

We first examine inference overhead, robustness to distribution shifts, and sensitivity to test-stream order. At 768×768 with a batch size of 12, Table 3 reports similar latency and peak memory for the frozen baseline and TPA, with GFLOPs rising from 710 to 721 when adaptation is enabled, and Fig. 6 shows that input resolution rather than batch size dominates the scaling of throughput and memory.

Supplementary Table S4 defines five shift families—*Light*, *Noise*, *Blur*, *Color*, and *Cross-batch*—with their parameter schedules. Table 4 reports macro-averaged Image-AUROC, Pixel-AUROC and mIoU over these families and severity settings, with Fig. 7 showing the corresponding severity-wise averages. Under this protocol, TPA has the highest values in Table 4 on all three metrics, followed by MuSC. Because a single mean over 25 heterogeneous conditions can conceal a family in which the method does not help, and it averages away *Cross-batch*, the aggregate evidences average behavior under this protocol rather than uniform improvement across shift types; Section S4 of the supplementary material gives the protocol in full.

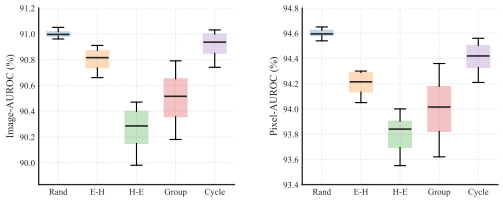
Table 6 and Fig. 8 examine five test-stream organizations and summarize their mean, variance, and worst-case degradation. This analysis is relevant because prototype refinement and historical reweighting are path-dependent, so different orders can produce different adaptation states even when the test set is unchanged. With the default warmup $W = 0$ the first image is already eligible for adaptation; the selection rule of Eq. 6 and the counter of Eq. 8 apply unchanged from the first step onward. The differences are modest, with random shuffling and cyclic mixing highest and hard→easy lowest; for that order a 25% warmup performs best among the tested lengths, and we retain $W = 0$ as the strict online default while treating warmup as a safeguard when hard samples are expected at the beginning of a stream. Section S5 of the supplementary material reports the warmup variants in full.



(a) Severity-wise Image-AUROC trend.

(b) Severity-wise Pixel-AUROC trend.

Figure 7: Robustness curves across shift severity levels. Each severity point is averaged over the five shift families in Supp. Table S4.



(a) Image-AUROC distribution across stream orders.

(b) Pixel-AUROC distribution across stream orders.

Figure 8: Order-sensitivity boxplots across stream orders.

Table 6: Order-sensitivity and warmup analysis across test-stream organizations. Metrics are higher-is-better; the summary block aggregates the five $W{=}0$ orders, reporting their population variance and best-to-worst range (worst-case drop).

Order type	Warmup W	Image-AUROC ↑	Pixel-AUROC ↑	mIoU ↑
Random-shuffle	0	91.0	94.6	47.9
Easy→Hard	0	90.8	94.2	47.4
Hard→Easy	0	90.3	93.8	46.8
Hard→Easy	10%	90.6	94.1	47.2
Hard→Easy	25%	90.7	94.3	47.4
Hard→Easy	50%	90.5	94.0	47.1
Class-grouped	0	90.5	94.0	47.1
Cyclic mixed	0	90.9	94.4	47.6
Mean (orders, $W = 0$)	–	90.7	94.2	47.4
Population variance (orders, $W = 0$)	–	0.07	0.08	0.15
Worst-case drop (orders, $W = 0$)	–	0.7	0.8	1.1

3.2.2 Module Ablations

Table 7 examines AP and DAW with DINOv3 (Siméoni et al., 2025), DINOv2 (Oquab et al., 2023), and ViT-B/16 (Dosovitskiy et al., 2021) backbones, all under the default configuration. To separate confidence-based patch selection from online updating itself, we include a naive-update baseline that applies the same prototype update to all test patches without the δ-based mask; the variant without AP or DAW is the static reference against which every gain is measured. With DINOv3, the naive update moves the IAD average from 85.4/89.2 to 88.4/91.9, and adding δ-based selection yields 89.9/93.1; FSSD-12 shows the same ordering. Added without AP, DAW still improves on the static reference in every configuration—with DINOv3 the IAD average image-level AUROC rises from 85.4 to 87.5—and the full configuration is highest in every column, supporting complementary contributions from both modules. The same ordering holds across the three backbones. Naive updating beats the static reference but remains below δ-selection, consistent with the motivation for confidence-based filtering: using every patch can incorporate defects, sensor noise or blurred

Table 7: Ablation of AP and DAW across DINOv3, DINOv2, and ViT-B/16 backbones. Cells give image/pixel AUROC (IAD average of Table 1) and 1-shot/5-shot mIoU (FSSD-12 three-fold average); best per column in bold.

AP setting	DAW	IAD average			FSSD-12 average	
		DINOv3	DINOv2	ViT-B/16	DINOv3	ViT-B/16
w/o AP	✗	85.4,89.2	85.7,89.5	84.8,89.5	41.5,42.4	41.0,41.9
AP w/ naive update	✗	88.4,91.9	88.3,91.8	88.2,91.6	44.8,45.7	44.2,44.9
AP w/ δ selection	✗	89.9,93.1	89.8,93.1	90.0,92.8	46.3,46.9	46.0,46.5
w/o AP	✓	87.5,91.7	87.4,91.5	86.7,90.9	43.1,45.0	42.5,43.7
AP w/ δ selection	✓	**91.0**,**94.6**	90.9,94.2	91.0,94.0	**47.9**,**48.6**	47.7,48.5

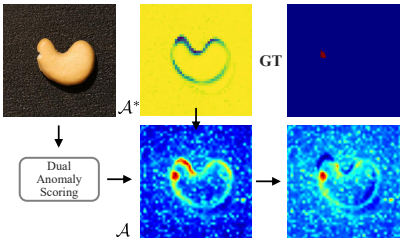


Figure 9: Qualitative comparison of DAW. The visualization contrasts the base anomaly map $\mathcal{A}$ with the DAW-refined anomaly map and GT.

Table 8: Spatial consistency assumption ablation on the IAD average, defined as in Table 5.

Spatial assumption	IAD Avg.
Fixed-position neighborhood (ours)	**92.8**
Local-window alignment	92.4
No spatial term	91.5

Table 9: Historical update policy ablation for DAW on the IAD average, defined as in Table 5.

Update policy	Description	IAD Avg.
Overwrite	$\mathcal{A}^{*,t} \leftarrow \tilde{\mathcal{A}}^{*,t}$	90.4
EMA-decay	$\mathcal{A}^{*,t} \leftarrow \lambda \mathcal{A}^{*,t-1} + (1-\lambda)\tilde{\mathcal{A}}^{*,t}$	92.2
Confidence-weighted	$\mathcal{A}^{*,t} \leftarrow \omega^t \tilde{\mathcal{A}}^{*,t} + (1-\omega^t)\mathcal{A}^{*,t-1}$	92.5
Softmax cumulative (ours)	$\mathcal{A}^{*,t} \leftarrow \mathrm{softmax}(\tilde{\mathcal{A}}^{*,t} + \mathcal{A}^{*,t-1})$	**92.8**

regions into the prototype, whereas δ-controlled selection limits updates to low-scoring patches. Fig. 9 shows the intended effect of cumulative reweighting: the base map of Eq. 9 responds to the defect but also carries diffuse background activation, which the DAW-weighted map of Eq. 12 suppresses while concentrating the response near the ground truth. Per-backbone and per-dataset breakdowns are in Section S5 of the supplementary material.

3.2.3 Design-choice comparisons

We next compare three design choices: the distance metrics used by DS and DAW, the DAW history-update rule, and the spatial assumption used by DS. These ablations use the aggregate IAD average and characterize system-level trends; per-dataset behavior is not resolved at this granularity.

Table 5 compares combinations of DS and DAW distance metrics. Normalized ℓ_2 outperforms cosine distance for DS under the combined-score ranking discussed in Section 2.3; ℓ_1 outperforms ℓ_2 for DAW in both paired comparisons.

Table 9 compares four history-update rules for DAW. For the confidence-weighted baseline, we define an adaptive fusion weight

$$\omega_i^t = \frac{\tilde{\mathcal{A}}_i^{*,t}}{\tilde{\mathcal{A}}_i^{*,t} + \mathcal{A}_i^{*,t-1} + \epsilon}, \tag{14}$$

and update the history as

$$\mathcal{A}_i^{*,t} \leftarrow \omega_i^t \tilde{\mathcal{A}}_i^{*,t} + (1-\omega_i^t)\mathcal{A}_i^{*,t-1}, \tag{15}$$

where ϵ is a small constant for numerical stability. The overwrite rule produces the lowest aggregate value in this comparison. EMA decay improves on overwrite, and confidence-weighted fusion provides a further increase. The softmax-cumulative rule reports the highest value, which is consistent with a benefit from temporal accumulation under the evaluated protocol.

Table 8 compares three spatial assumptions. They are the fixed-position eight-neighbor model of Eq. 3; a relaxed variant that, instead of comparing each patch with the co-located prototype token, searches a small window around that location and matches against the best-aligned token; and a variant without $\mathbf{E}_s$. The fixed-position neighborhood has the highest aggregate value, which supports treating spatial alignment as a useful prior after the evaluated resizing and feature-extraction pipeline. Local-window alignment is slightly lower, while removing the spatial term produces the largest decrease in this comparison.

Sensitivity to α, δ, and the number of support shots is reported in Figure S2 and Section S5 of the supplementary material.

4 Conclusion

Test-time prototype adaptation keeps the encoder frozen while dual anomaly scoring gates patch admission, adaptive prototype update refines the matched spatial tokens with per-location counters, and dynamic attention weighting uses historical consistency to refine anomaly maps. Across MVTec AD, VisA and BTAD, it reaches 91.0% average image-level and 94.6% average pixel-level AUROC with two normal support images per class, supporting online refinement of a dense normal reference under the evaluated protocol. Its applicability is bounded by three conditions: approximate spatial alignment between test images and prototypes, the possibility that anomalous patches admitted early contaminate sparsely updated tokens, and the diminishing step size of the counter-weighted mean, which limits tracking under persistent drift.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, to modify scientific figures, and to propose keywords.

ETHICS STATEMENT

All authors declare that this work was conducted in accordance with relevant ethical guidelines and standards.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we will release the complete source code upon acceptance of this paper. All datasets used in the experiments are publicly available, and the detailed data preprocessing steps are described in the supplementary materials. Additional implementation details and experimental configurations are provided in the appendix.

REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2025.

Zixuan Chen, Xiaohua Xie, Lingxiao Yang, and Jian-Huang Lai. Hard-normal example-aware template mutual matching for industrial anomaly detection. *International Journal of Computer Vision*, 133(5):2927–2949, 2025.

Alex Costanzino, Pierluigi Zama Ramirez, Mirko Del Moro, Agostino Aiezzo, Giuseppe Lisanti, Samuele Salti, and Luigi Di Stefano. Test time training for industrial anomaly segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3910–3920, June 2024.

Zhewei Dai, Shilei Zeng, Haotian Liu, Xurui Li, Feng Xue, and Yu Zhou. Seas: few-shot industrial anomaly image generation with separation and sharing fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23135–23144, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pp. 475–489. Springer, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Yongjie Du, Deyun Zhou, Yu Xie, Yu Lei, and Jiao Shi. Prototype-guided feature learning for unsupervised domain adaptation. *Pattern Recognition*, 135:109154, 2023.

Hu Feng, Kechen Song, Wenqi Cui, Yiming Zhang, and Yunhui Yan. Cross position aggregation network for few-shot strip steel surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–10, 2023.

Zhitong Gao, Shipeng Yan, and Xuming He. Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. *Advances in Neural Information Processing Systems*, 36: 45150–45171, 2023.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15194–15203, 2025.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 98–107, 2022.

Jianfang He, Min Cao, Silong Peng, and Qiong Xie. Rareclip: Rarity-aware online zero-shot industrial anomaly detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24478–24487. IEEE, 2025.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11375–11385, June 2024.

Jiacheng Huang, Yong Wu, Xiaofei Zhou, Jia Lin, Zhangping Chen, Guodao Zhang, Lei Xia, and Jiyong Zhang. Multi-scale adaptive prototype transformer network for few-shot strip steel surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19606–19616, 2023.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16838–16848, 2024a.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *The Twelfth International Conference on Learning Representations*, 2024b.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pp. 1–34, 2024.

Weide Liu, Zhonghua Wu, Henghui Ding, Fayao Liu, Jie Lin, and Guosheng Lin. Few-shot segmentation with global and local contrastive learning. *arXiv preprint arXiv:2108.05293*, 2021.

Xingyue Liu, Qian Wu, Yahui Cheng, and Guojun Wen. A novel prior knowledge-based and texture-enhanced network for few-shot industrial defect segmentation. *Engineering Applications of Artificial Intelligence*, 162:112408, 2025.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S.Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4744–4754, June 2025.

Declan McIntosh and Alexandra Branzan Albu. Unsupervised, online and on-the-fly anomaly detection for non-stationary image distributions. In *European Conference on Computer Vision*, pp. 428–445. Springer, 2024.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06. IEEE, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30398–30408, 2025.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 825–835. IEEE, 2025.

Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022. doi: 10.1109/TIM.2022.3196436.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Qian Wan, Liang Gao, Xinyu Li, and Long Wen. Unsupervised image anomaly detection and segmentation based on pretrained feature mapping. *IEEE Transactions on Industrial Informatics*, 19 (3):2330–2339, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Shenxing Wei, Xing Wei, Zhiheng Ma, Songlin Dong, Shaochen Zhang, and Yihong Gong. Few-shot online anomaly detection and segmentation. *Knowledge-Based Systems*, 300:112168, 2024.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Jiayi Lyu, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 54(5):2720–2733, 2024. doi: 10.1109/TCYB.2024.3357213.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O Arik, and Tomas Pfister. Spade: Semi-supervised anomaly detection under distribution mismatch. *arXiv preprint arXiv:2212.00173*, 2022.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

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.

Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5217–5226, 2019.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15204–15213, 2025a.

Zeyu Zhang, Danqing Kang, Biaohua Ye, and Jianhuang Lai. Symmflow: Unsupervised anomaly detection via symmetric normalizing flow. *Computers in Industry*, 173:104393, 2025b. ISSN 0166-3615. doi: https://doi.org/10.1016/j.compind.2025.104393.

Jichun Zhao, Haiyong Jiang, Haoxuan Song, Jun Xiao, and Dong Gong. Dˆ 3ctta: Domain-dependent decorrelation for continual test-time adaption of 3d lidar segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11864–11874, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.