DeMuS: Learning Decoupled Matching and Scoring for Batch Zero-Shot Industrial Anomaly Detection

Zhengyang Zhao[1,2], Hailong Sun[1,2]*, Binhang Qi[1,2], Hongrui Yu[1,2], Zhongchi Wang[1], and Hang Xu[1]

[1] Beihang University, Beijing, China
{zhaozhengyang,sunhl,binhangqi,yuhongrui,zy2421128}@buaa.edu.cn, wangzc_1997@outlook.com
[2] Hangzhou Innovation Institute of Beihang University, Hangzhou, China

Abstract. Zero-Shot Industrial Anomaly Detection (IAD) is critical for rapid deployment in agile manufacturing. We investigate a highly practical **Batch Zero-Shot IAD** setting, strictly constrained by small intra-batch references and common pose variations. Directly applying existing mutual-scoring methods fails because training-free features lack generalized anomaly semantics and spatial robustness, resulting in ambiguous scoring distances and an "anomaly-to-anomaly shortcut" during matching, which can lead to missed detections. To address this, we propose **DeMuS** (**De**coupled **Mu**tual **S**coring), a novel two-stage learning framework. By training two lightweight adapters via a two-stage auxiliary process, our method extracts highly discriminative anomaly semantics for accurate scoring (*how-to-score*), while explicitly decoupling robust part-level correspondence learning (*where-to-match*) to guarantee reliable physical alignment. This synergistic mechanism effectively prevents semantic confusion under unaligned inputs. Extensive experiments on MVTec AD, VisA, and Real-IAD demonstrate DeMuS achieves state-of-the-art performance and superior robustness against pose shifts, maintaining efficient inference for online inspection. Our code is available at https://github.com/evoLonation/DeMuS.

Keywords: Anomaly Detection · Batch Mutual Scoring · Zero-Shot Learning

1 Introduction

Industrial Anomaly Detection (IAD) deployment is chronically hindered by data and annotation costs. Anomalous samples are naturally scarce and highly diverse, while frequent product iterations require rapid algorithm deployment with minimal supervision [1, 11, 29]. Traditional unsupervised methods [4, 14, 30, 37, 44] still rely on large collections of normal samples per category,

* Corresponding author.

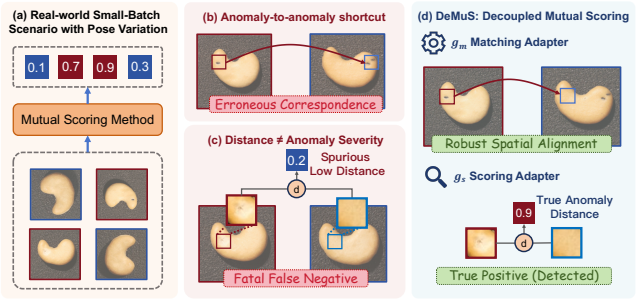

Fig. 1: Challenges in Batch Zero-Shot IAD and our DeMuS solution. (a) Practical scenario with unaligned intra-batch samples. (b-c) Key failure modes of baseline mutual scoring: (b) spatial shortcuts between anomalies and (c) lack of anomaly-aware semantics. (d) DeMuS resolves these via decoupled matching and scoring adapters.

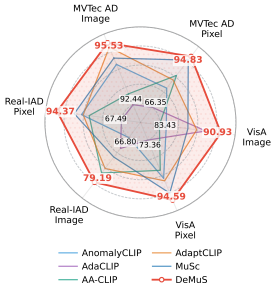

Fig. 2: Holistic superiority of DeMuS across three major industrial benchmarks at both image and pixel-level granularities.

entailing prohibitive data collection and maintenance costs in high-throughput lines. Consequently, Zero-Shot Anomaly Detection (ZSAD) [6, 33], which infers directly without target domain training data, has emerged as a more scalable solution for open industrial environments.

Benefiting from the advances in vision-language pre-training [36, 47], a surge of recent ZSAD works has followed a text-guided paradigm [9,16,20,31,49]. However, their generalization relies on text-driven semantic priors. Consequently, they struggle to identify structural defects, such as missing components, which are inherently indistinguishable without explicit normal visual references, thereby fundamentally restricting their fine-grained anomaly localization capabilities [31].

To bypass text-guidance limitations, recent methods leverage visual priors within the test set itself [2,3,23,27,48]. For instance, MuSc [27] establishes patch-level visual correspondences across unlabeled test images. However, it assumes access to the *entire* test set during inference, incurring prohibitive latency and memory overheads. In real production lines, inspection occurs sequentially in small batches. Motivated by this, we investigate **Batch Zero-Shot IAD** [2, 3, 48] (Fig. 1a), utilizing only intra-batch samples as references. Crucially, in this realistic setting, samples inevitably exhibit significant **Pose Variation** (e.g., rotation and translation), strictly requiring methods to establish stable local correspondences under unaligned inputs before identifying anomalies.

In this setting, pose variations force algorithms to rely on nearest neighbor (NN) matching for part correspondences [21, 25, 37]. However, directly applying baseline mutual scoring features exposes a key failure mode: the *anomaly-to-anomaly shortcut*. When abnormal samples coexist within a batch, a target anomaly can erroneously match a similar-looking anomaly from a different part in the reference image as illustrated in Fig. 1b [22]. Simultaneously, the pre-trained feature space lacks *anomaly-aware semantics*, meaning that distance variations often capture background perturbations rather than true anomaly

severity (Fig. 1c). Both issues converge to yield spuriously small distances, leading to inevitable missed detections (false negatives).

To address this, we propose **DeMuS** (**De**coupled **Mu**tual **S**coring), tailored for Batch Zero-Shot IAD, as shown in Fig. 1d. Our core idea achieves decoupled "part correspondence" (*where-to-match*) and "anomaly scoring" (*how-to-score*) via a two-stage auxiliary training process. Using a frozen backbone and trainable multi-scale adapters, the first stage trains a *Matching Adapter* via explicitly constructed alignment supervision. By strictly focusing on robust physical alignment against pose shifts, it mechanistically eradicates the anomaly-to-anomaly shortcut. Freed from spatial alignment burdens, the second stage trains a *Scoring Adapter* to mine generalized anomaly semantics. It forces normal-normal pairs to converge and anomalous pairs apart, drastically enhancing defect discriminability. During inference, a single NN matching step generates soft correspondences, followed by linear-weighted anomaly scoring. This decoupled design preserves the computational efficiency of baseline methods while satisfying the strict latency constraints of online inspection. Fig. 2 provides a compact overview of DeMuS's performance across three benchmarks.

The main contributions of this paper are summarized as follows:

- We formalize and study **Batch Zero-Shot IAD** under two practical deployment constraints: *limited intra-batch references* and *pose variation*, and analyze key failure modes of mutual-scoring-only inference in this protocol.
- We propose **DeMuS**, introducing auxiliary training to the mutual scoring framework. Via explicit alignment supervision and a two-stage decoupled learning paradigm, we effectively eradicate the anomaly-to-anomaly shortcut and extract highly discriminative anomaly semantics.
- Extensive experiments on **MVTec AD** [5], **VisA** [52], and **Real-IAD** [42] (including pose-variation evaluations and ablations) demonstrate that DeMuS consistently improves performance and robustness over strong zero-shot baselines with efficient inference.

2 Related Work

2.1 Text-Guided ZSAD with CLIP

Mainstream zero-shot anomaly detection (ZSAD) methods build on vision–language pretraining (e.g., CLIP-style encoders [36, 41, 46]) and compute similarities between visual features and text features derived from normal/abnormal prompts to generate anomaly scores and localization maps. WinCLIP [20] systematically introduces compositional prompt ensembles and multi-scale aggregation for localization without target-domain tuning. However, since the original CLIP is biased towards category-level semantic alignment, it often lacks fine-grained anomaly awareness. To address this, recent works adapt the model using auxiliary annotated AD data in the source domain [7–10, 13, 16, 18, 24, 31, 34, 35, 49–51]. For instance, AnomalyCLIP [49] learns object-agnostic prompts to capture generic normality/abnormality. AdaCLIP [7] employs hybrid learnable prompts

for test-time conditional adaptation, while AA-CLIP [31] disentangles anomaly-aware anchors in the text space before aligning patch-level visual features. In summary, while this branch improves separability via prompt learning on auxiliary data, its heavy reliance on text priors fundamentally limits its ability to handle structural defects, which require explicit visual correspondences.

2.2 Leveraging Test-Set Information

Test-Set Visual Mutual Referencing. Unlike text-guided ZSAD, ACR [23] and MuSc [27] incorporate the visual information of the test set into the inference process, inspiring subsequent works [22, 28, 43]. MuSc utilizes nearest-neighbor matching and mutual scoring among unlabeled test samples for fine-grained anomaly localization. While its performance is compelling, MuSc implicitly relies on a large reference set (or even the entire test set). In contrast, we focus on the deployment-friendly Batch Zero-Shot scenario [3], restricting the reference set to a small batch to meet real-time constraints. More importantly, we identify that directly applying conventional mutual scoring in such restricted settings is highly vulnerable to pose variations, which we explicitly tackle via our decoupled spatial alignment mechanism.

Vision-Text Small-Batch Fusion. Some works explore pair-image or batch-based protocols without relying on the entire test set [2, 3, 48]. PatchCLIP [2] and Dual-Image Enhanced CLIP [48] introduce visual references via paired images while retaining the text-guided scoring framework. FiSeCLIP [3] discusses batch-based testing and introduces cross-modal filtering to suppress noise from unlabeled references. Unlike these early explorations that compensate small-batch mutual reference with text guidance, our approach is entirely orthogonal. We do not rely on text priors or cross-modal filtering. Instead, we propose a **pure visual solution** that directly tackles the fundamental challenges of spatial alignment and semantic confusion within the visual feature space, achieving robust mutual scoring under unaligned inputs via a decoupled learning paradigm.

Online TTA and Memory Updating. Another pipeline continuously adjusts model parameters or updates a memory bank during testing via test-time adaptation (TTA) or online updating [17, 26, 48]. While these methods can absorb new observations sequentially, they require strong protocol assumptions, such as cold-start accumulation. They are also more susceptible to feature drift and higher maintenance costs in fast-paced production lines with frequent lighting or product changes. Our method solely depends on immediate intra-batch references without maintaining long-term online states, making it more robust for scenarios with limited references and rapid distribution shifts.

2.3 FSAD with Memory Bank

This branch explicitly defines normality by introducing a few normal reference images, enabling rapid deployment via memory banks or retrieval mechanisms [9, 16, 19, 21, 37, 51]. Early works like APRIL-GAN [9] and WinCLIP [20] allow the provision of a few normal samples to complement their text-guided

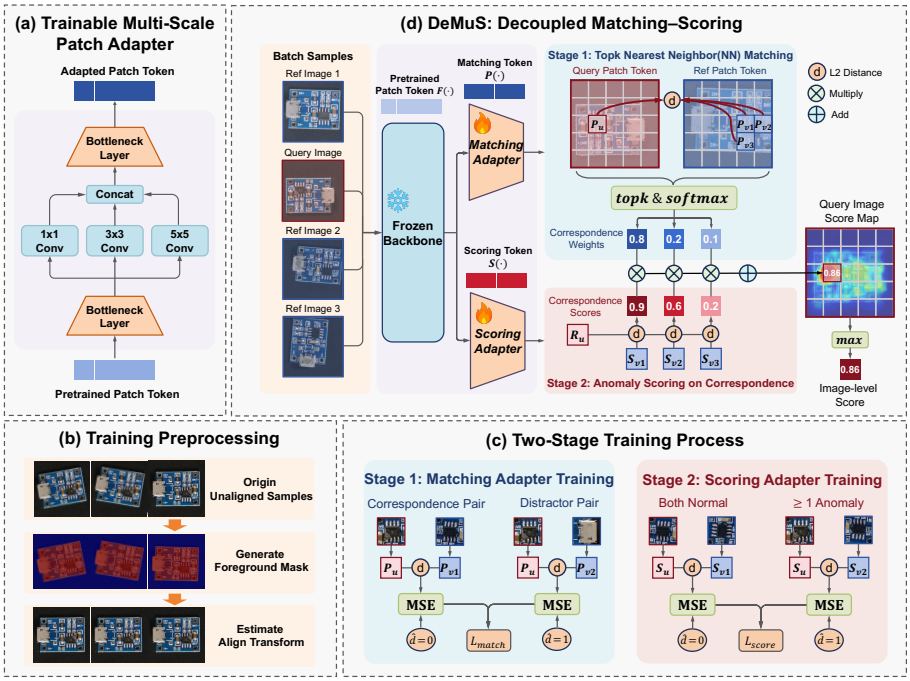

Fig. 3: Overview of the DeMuS framework. (a) Trainable Patch Adapter: Adapts frozen pre-trained features via multi-scale convolutions. **(b) Preprocessing**: Estimates spatial alignment on auxiliary data for robust "same-part" supervision. **(c) Two-Stage Training**: Matching Adapter learns pose-invariant alignment, while Scoring Adapter extracts discriminative anomaly semantics. **(d) Inference**: Matching features establish top-k soft correspondences, enabling scoring features to compute accurate semantic distances exclusively on aligned parts, effectively preventing shortcuts.

foundations without retraining for each target category. AdaptCLIP [16] achieves state-of-the-art performance by supporting zero-/few-shot generalization using lightweight adapters in a training-free manner on the target domain. From a "reference mechanism" perspective, our method advances this philosophy into a fully zero-shot and dynamic setting: rather than explicitly requiring offline-collected normal images and maintaining a persistent memory bank, we treat the unlabeled intra-batch samples as an **ephemeral reference set**. This essentially achieves the rapid deployment of FSAD without the strict requirement of few-shot normal templates.

3 Method

3.1 Problem Formulation

Unlike conventional single-image zero-shot AD, we introduce **Batch Zero-Shot IAD** to mirror real-world manufacturing, where products are continuously in-

spected in small batches without target-domain annotations. Formally, testing occurs in batches $\mathcal{B}_t = \{I_1^{(t)}, \ldots, I_B^{(t)}\}$, where a small B (e.g., 4 or 8) ensures low latency. Since anomalies are rare, normal patches dominate [27]. Each target $I_a^{(t)} \in \mathcal{B}_t$ is evaluated exclusively against the remaining intra-batch references:

$$\mathcal{R}\Big(I_a^{(t)}\Big) = \mathcal{B}_t \setminus \{I_a^{(t)}\}, \qquad \Big|\mathcal{R}\Big(I_a^{(t)}\Big)\Big| = B - 1. \tag{1}$$

Real-World Constraints. This protocol imposes two stringent real-world constraints: (i) *Restricted References*: The model cannot access any pre-collected, normal memory banks; it must rely entirely on the $B-1$ immediate, unlabeled neighbors. (ii) *Pose Variation*: Intra-batch samples inevitably exhibit spatial misalignments (e.g., rotation, translation). *This is the fundamental catalyst that necessitates robust part-level matching before anomaly semantic comparison* [25].
Auxiliary Training Data. To overcome the semantic ambiguity of pre-trained features, we leverage a target-disjoint auxiliary dataset $\mathcal{D}_{aux}$ with anomaly annotations [7, 9, 49]. Training is conducted exclusively on $\mathcal{D}_{aux}$ to learn discriminative anomaly semantics and robust spatial alignment. During target-domain inference, the model adheres to the zero-shot protocol (no further tuning) and detects defects referencing only the instantaneous batch $\mathcal{B}_t$.

3.2 Overview

To tackle unaligned intra-batch inference, we propose **DeMuS** (**De**coupled **Mu**tual **S**coring) as illustrated in Fig. 3. Building upon a frozen pre-trained backbone, we introduce trainable multi-scale adapters to adaptively capture diverse defect granularities. To provide reliable spatial supervision, we construct a correspondence oracle via explicit alignment estimation on auxiliary data. Guided by this oracle, our framework is optimized via a two-stage decoupled paradigm: a **Matching Adapter** (g_m) first learns pose-invariant physical correspondences (*where-to-match*) to substantially mitigate spatial shortcuts; subsequently, freed from alignment burdens, a **Scoring Adapter** (g_s) extracts highly discriminative anomaly semantics (*how-to-score*). During inference, g_m computes top-k soft correspondence weights via nearest-neighbor search, enabling g_s to aggregate true semantic anomaly distances exclusively on these aligned parts, thereby ensuring precise and efficient online inspection.

3.3 Frozen Backbone and Multi-layer Tokens

Given an input image I, we extract multi-layer patch tokens using a frozen pre-trained visual backbone. For a selected layer $\ell \in \mathcal{L}$, the token grid is $F^{(\ell)}(I) \in \mathbb{R}^{h \times w \times c}$. We introduce an independent, trainable lightweight adapter $g^{(\ell)}(\cdot)$ for each layer, yielding $Z^{(\ell)}(I) = g^{(\ell)}(F^{(\ell)}(I))$. To eliminate the impact of feature scales, we apply ℓ_2 normalization: $\tilde{Z}_u^{(\ell)} = Z_u^{(\ell)} / (\|Z_u^{(\ell)}\|_2 + \epsilon)$, where u is the patch index and ϵ ensures numerical stability. We adopt distance-level fusion to

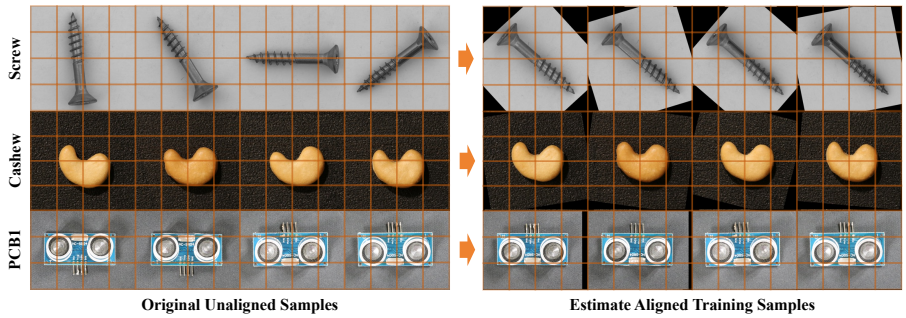

Fig. 4: Visualization of alignment preprocessing. Examples of original unaligned samples and their aligned results from the MVTec AD and VisA datasets.

aggregate multi-layer distances. For brevity, we encapsulate the distance metric induced by the multi-layer features $\{Z^{(\ell)}(\cdot)\}_{\ell\in\mathcal{L}}$ into a unified notation:

$$\|Z_u(A) - Z_v(B)\|_2 \triangleq \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \left\|\tilde{Z}_u^{(\ell)}(A) - \tilde{Z}_v^{(\ell)}(B)\right\|_2. \tag{2}$$

3.4 Trainable Multi-Scale Patch Adapter

Unlike the static multi-scale pooling utilized in MuSc [27], we propose a **trainable multi-scale Inception-style adapter** [39, 40] (see Fig. 3a) to adaptively capture diverse defect granularities. This architecture is instantiated independently for the matching branch (g_m) and the scoring branch (g_s). Given $F \in \mathbb{R}^{h\times w\times c}$, a 1×1 convolution first reduces the channel dimension to c_b, yielding F_b. We then apply three parallel convolutional branches (1×1, 3×3, and 5×5) on F_b, concatenate their outputs, and project back to c channels via a 1×1 convolution to obtain ΔF. Finally, a residual connection is employed to preserve pre-trained stability: $\hat{F} = F + \alpha\cdot\Delta F$, where α is a scaling constant (e.g., 0.1). Thus, the matching and scoring features are defined as $P(I) = g_m(F(I))$ and $S(I) = g_s(F(I))$, respectively.

3.5 Training Preprocessing and Augmentation

Foreground Mask. Background regions disrupt matching. We generate a foreground mask $M(I) \in \{0,1\}^{h\times w}$ via 1D PCA on the pre-trained patch features followed by coarse thresholding, computing losses exclusively on foreground patches. This preprocessing is only required during auxiliary training.

Correspondence Oracle. To construct "same-part" supervision (Fig. 3b), we estimate an alignment transformation $T_{A\to B}$ between intra-class image pairs (A, B) using OpenCV's `findTransformECC` on their masks (as visualized in Fig. 4). For a patch $u \in \Omega_A$, we map its center to B and select its 4 nearest discrete neighbors to form a candidate correspondence set $\mathcal{P}(u)$. This localized neighbor selection accommodates slight deformations and minor alignment

errors between objects. For non-alignable categories (e.g., repetitive textures), $\mathcal{P}(u)$ is derived via simple NN matching on average-pooled pre-trained features.
Pose Augmentation. During training, we apply random affine augmentations (rotation and translation) to prevent overfitting, decouple absolute positional encodings, and simulate production line pose shifts. If augmentations G_A and G_B are applied to images A and B to yield A' and B', the alignment matrix is updated via $T_{A'\to B'} = G_B\, T^0_{A\to B}\, G_A^{-1}$, ensuring consistency between spatial supervision and data augmentation.

3.6 Two-Stage Decoupled Training

Stage-1: Matching Adapter. g_m learns stable same-part correspondences, focusing on "where-to-match" while remaining invariant to anomalies and pose shifts, as shown in Fig. 3c Left. For an intra-class pair (A, B) and $u \in \Omega_A$, we select the patch $v^\star(u) \in \mathcal{P}(u)$ with the minimum feature distance as the *positive correspondence pair* (target distance $\hat{d} = 0$). We further randomly sample $v^-(u)$ from $\Omega_B \setminus \{v^\star(u)\}$ to form a *distractor pair* ($\hat{d} = 1$), which prevents feature collapse and enhances discriminability. g_m is optimized via Mean Squared Error (MSE) regression:

$$\mathcal{L}_{match} = \sum_{u\in\Omega_A} \Big(\underbrace{\big(\|P_u(A) - P_{v^\star(u)}(B)\|_2 - 0\big)^2}_{\text{correspondence regression}} + \underbrace{\big(\|P_u(A) - P_{v^-(u)}(B)\|_2 - 1\big)^2}_{\text{distractor regression}} \Big). \tag{3}$$

Stage-2: Scoring Adapter. As depicted in Fig. 3c Right, g_s learns to distinguish anomaly semantics based on the established correspondences. We downsample the pixel-level anomaly annotations to the patch grid, yielding $y_u(I) \in \{0, 1\}$ (1 indicates anomaly). For each candidate pair (u, v), the target distance is defined as $\hat{d}_{uv} = 0$ if both $y_u(A) = 0$ and $y_v(B) = 0$, and 1 otherwise. g_s is optimized via MSE regression on the same-part candidate set:

$$\mathcal{L}_{score} = \sum_{u\in\Omega_A} \sum_{v\in\mathcal{P}(u)} \big(\|S_u(A) - S_v(B)\|_2 - \hat{d}_{uv}\big)^2. \tag{4}$$

This enforces anomaly-awareness: normal-normal pairs converge, while anomalous pairs are pushed apart.

3.7 Inference: Decoupled Mutual Scoring

Given a target image A and a reference set $\mathcal{R}(A) = \{R_1, \ldots, R_n\}$ (where $n = B - 1$), inference proceeds in two steps, as detailed in Fig. 3d.
Step 1: Matching-based soft correspondence (top-k NN). We compute the $N \times N$ distance matrix $D^{(i)}_{u,v} = \|P_u(A) - P_v(R_i)\|_2$ on the flattened patch grid ($N = hw$). For each target patch u, we retrieve the top-k nearest indices in R_i, denoted as $\mathcal{K}_i(u)$, and compute the softmax-normalized weights:

$$w_{i,u,v} = \frac{\exp\big(-D^{(i)}_{u,v}\big)}{\sum_{v'\in\mathcal{K}_i(u)} \exp\big(-D^{(i)}_{u,v'}\big)}, \qquad v \in \mathcal{K}_i(u). \tag{5}$$

Step 2: Anomaly Scoring on Correspondence. Using the scoring features $S(\cdot)$, the anomaly score referenced by R_i is aggregated exclusively on the top-k set:

$$s_u^{(i)} = \sum_{v \in \mathcal{K}_i(u)} w_{i,u,v} \cdot \|S_u(A) - S_v(R_i)\|_2. \tag{6}$$

We fuse scores across all references by adopting the most "normal" explanation: $\mathcal{M}_u(A) = \min_i s_u^{(i)}$, yielding the patch-level anomaly map. The image-level score is $\max_u \mathcal{M}_u(A)$.

Complexity. The primary $O(nN^2)$ overhead stems from computing the g_m distance matrix, whereas the g_s scoring adds only $O(nNk)$ ($k \ll N$). Thus, DeMuS inherently preserves the same efficiency order as baseline mutual scoring [27], easily satisfying small-batch online latency requirements.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three popular industrial anomaly detection benchmarks: MVTec AD [5], VisA [52], and Real-IAD [42]. These encompass 57 distinct categories and 118,472 test samples, covering general object images, textures, cluttered objects, and multi-view manufacturing scenarios.

Baselines. We compare against five state-of-the-art text-guided zero-shot methods: APRIL-GAN [9], AnomalyCLIP [49], AdaCLIP [7], AA-CLIP [31], and AdaptCLIP [16]. For mutual scoring methods, we include MuSc [27] and its variant equipped with DINOv3 [38], denoted as MuSc(DINOv3). MuSc and MuSc(DINOv3) are the most direct comparisons, while text-guided methods are included for comprehensive positioning. For fairness under our protocol, we also implement intra-batch mutual-scoring variants of APRIL-GAN and AdaptCLIP by replacing their fixed memory bank with the remaining samples in the current batch (reported as '-ms').

Metrics. Following standard protocols [49], we report image-level Area Under the ROC Curve (AUROC) and Average Precision (AP), alongside pixel-level AUROC and Per-Region Overlap (AUPRO) [5], compactly tabulated as AUC, AP, PAUC, and PRO.

Implementation Details. Following AA-CLIP [31], we use VisA as the primary auxiliary training set, but switch to MVTec AD when evaluating VisA. For MuSc, the reference set size is aligned with our batch size. For our method, we employ the frozen DINOv3 (ViT-L/16) [15, 38, 45] as the backbone, utilizing features from layers 6, 12, 18, and 24. We set $k = 3$ for top-k matching. All experiments are conducted on a single H100 GPU. Input images are resized and cropped to 512×512 to ensure exact division by the patch size (16). Unless otherwise specified, the batch size is set to 4 for MVTec AD and VisA, and 8 for Real-IAD (accounting for its 5 different capturing views per category).

Table 1: Performance comparison with zero-shot methods on MVTec AD, VisA, and Real-IAD. **Top:** Image-level metrics (AUC, AP). **Bottom:** Pixel-level metrics (PAUC, PRO). Best and second-best results are highlighted in bold and underlined. The methods are categorized into Text-only and Vision-only based on the utilized information. Since the evaluation of mutual-scoring methods is sensitive to the sample order, we report the mean and standard deviation (mean±std) across 5 random seeds.

		MVTec AD		VisA		Real-IAD		Average	
Type	Method	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Text only	APRIL-GAN	86.10	93.50	78.00	81.40	68.60	64.60	77.57	79.83
	AnomalyCLIP	91.50	96.20	82.10	85.40	69.40	64.19	81.00	81.93
	AdaCLIP	89.20	95.68	85.80	88.71	70.86	67.26	81.95	83.88
	AA-CLIP	90.50	95.05	84.60	82.25	75.41	73.54	83.50	83.61
	AdaptCLIP	93.50	96.60	84.80	87.20	74.20	71.90	84.17	85.23
Vision only	MuSc	92.16 ±0.57	96.32 ±0.18	84.26 ±0.44	85.80 ±0.44	73.87 ±0.13	66.74 ±0.14	83.43	82.95
	MuSc(DINOv3)	91.53 ±0.56	95.74 ±0.37	84.39 ±0.61	85.85 ±0.37	73.21 ±0.11	63.90 ±0.13	83.04	81.83
	DeMuS	**94.17** ±0.34	**96.89** ±0.28	**90.16** ±0.19	**91.70** ±0.11	**80.78** ±0.10	**77.60** ±0.08	**88.37** +4.20	**88.73** +3.50
Type	Method	PAUC	PRO	PAUC	PRO	PAUC	PRO	PAUC	PRO
Text only	APRIL-GAN	87.60	44.00	94.20	86.80	94.60	43.60	92.13	58.13
	AnomalyCLIP	91.10	81.40	95.50	87.00	95.08	83.96	93.89	84.12
	AdaCLIP	88.70	44.00	95.50	51.21	95.84	39.14	93.35	44.78
	AA-CLIP	91.90	87.23	95.50	84.77	95.04	81.55	94.15	84.52
	AdaptCLIP	90.90	85.90	95.60	87.60	94.90	83.60	93.80	85.70
Vision only	MuSc	96.07 ±0.19	91.04 ±0.40	96.38 ±0.50	90.22 ±0.29	96.53 ±0.04	84.84 ±0.16	96.33	88.70
	MuSc(DINOv3)	**96.40** ±0.18	90.12 ±0.32	96.65 ±0.31	86.22 ±0.67	95.77 ±0.05	80.26 ±0.17	96.27	85.53
	DeMuS	96.15 ±0.12	**93.50** ±0.11	**96.94** ±0.28	**92.23** ±0.06	**97.64** ±0.03	**91.10** ±0.04	**96.91** +0.58	**92.28** +3.58

4.2 Main Results

As presented in Tab. 1, our method comprehensively outperforms all text-guided baselines across both image-level and pixel-level metrics. Notably, compared to the current SOTA method AdaptCLIP, we achieve an improvement of 4.20% in AUROC and 6.58% in AUPRO. When compared with the mutual scoring baseline MuSc, our approach yields an average increase of 4.94% in image AUROC and a remarkable 10.86% boost in image AP on the challenging multi-view Real-IAD dataset. Pixel-level AUPRO is also improved by an average of 3.58%, demonstrating superior fine-grained defect localization.

Furthermore, we benchmark against the highly competitive intra-batch variants of APRIL-GAN and AdaptCLIP, which utilize vision-language bimodal information, as presented in Tab. 2. Despite relying exclusively on visual references, our method remains superior, outperforming them by 1.51% in image AUROC and 3.07% in AUPRO on average. Specifically, we observe substantial gains on Real-IAD and MVTec AD, while remaining on par with the baselines on VisA due to their strong text-guided priors.

Table 2: Performance comparison with multi-modal mutual-scoring variants. We benchmark our vision-only DeMuS against the Text&Vision mutual-scoring variants of competitive few-shot methods (denoted with an '-ms' suffix) to demonstrate its cross-modal superiority.

		MVTec AD		VisA		Real-IAD		Average	
Type	Method	AUC	AP	AUC	AP	AUC	AP	AUC	AP
T&V	APRIL-GAN-ms	91.66 ±0.10	95.86 ±0.10	**90.44** ±0.17	**92.82** ±0.15	78.48 ±0.04	76.74 ±0.05	86.86	88.47
T&V	AdaptCLIP-ms	93.58 ±0.56	**97.04** ±0.34	89.20 ±0.36	90.50 ±0.37	77.68 ±0.07	73.16 ±0.10	86.82	86.90
Vis.	**DeMuS**	**94.17** ±0.34	96.89 ±0.28	90.16 ±0.19	91.70 ±0.11	**80.78** ±0.10	**77.60** ±0.08	**88.37** +1.51	**88.73** +0.26
Type	Method	PAUC	PRO	PAUC	PRO	PAUC	PRO	PAUC	PRO
T&V	APRIL-GAN-ms	94.48 ±0.16	90.30 ±0.11	96.04 ±0.08	89.70 ±0.09	96.52 ±0.04	87.62 ±0.07	95.68	89.21
T&V	AdaptCLIP-ms	93.64 ±0.10	89.36 ±0.10	**96.94** ±0.08	90.72 ±0.13	96.40 ±0.00	85.84 ±0.05	95.66	88.64
Vis.	**DeMuS**	**96.15** ±0.12	**93.50** ±0.11	**96.94** ±0.28	**92.23** ±0.06	**97.64** ±0.03	**91.10** ±0.04	**96.91** +1.23	**92.28** +3.07

Table 3: Performance comparison under the Pose Variation scenario. We report absolute metrics and the performance retention rate (defined as the ratio of the pose-varied performance to the default performance in Tab. 1). To simulate unaligned real-world production lines, test samples are perturbed with random arbitrary-angle rotations and translations within $[-128, 128]$ pixels.

	MVTec AD		VisA		Real-IAD		Avg. Perf.		Avg. Ret.	
Method	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MuSc	81.30 ±1.16	91.24 ±0.43	71.41 ±1.10	75.16 ±1.00	65.94 ±0.02	62.16 ±0.06	72.88	76.19	87.41	91.82
MuSc(DINOv3)	78.80 ±1.23	89.18 ±0.67	70.49 ±1.50	73.85 ±1.26	60.59 ±0.06	55.75 ±0.09	69.96	72.93	84.13	88.81
DeMuS	**90.79** ±0.58	**94.77** ±0.32	**83.60** ±0.50	**85.74** ±0.32	**76.54** ±0.06	**73.94** ±0.16	**83.64** +10.76	**84.82** +8.63	**94.63** +7.22	**95.53** +3.71
Method	PAUC	PRO	PAUC	PRO	PAUC	PRO	PAUC	PRO	PAUC	PRO
MuSc	93.57 ±0.13	84.76 ±0.23	92.40 ±0.79	83.96 ±0.26	94.98 ±0.02	79.65 ±0.08	93.65	82.79	97.22	93.35
MuSc(DINOv3)	92.35 ±0.48	76.55 ±0.45	90.17 ±0.20	68.74 ±0.69	92.51 ±0.04	70.22 ±0.08	91.68	71.84	95.23	84.05
DeMuS	**94.48** ±0.28	**90.55** ±0.10	**96.18** ±0.15	**90.19** ±0.46	**96.87** ±0.02	**88.66** ±0.02	**95.84** +2.19	**89.80** +7.01	**98.90** +1.68	**97.32** +3.97

4.3 Robustness to Pose Variation

To simulate unaligned scenarios in actual production lines, we apply random rotations and slight translations to the test samples. As shown in Tab. 3, our method exhibits exceptional robustness under the Pose Variation setting. Compared to mutual scoring baselines, our AUROC improves by an average of 10.74%. Remarkably, our method retains 95.53% of its original image AP and 97.32% of its original AUPRO despite severe spatial misalignment.

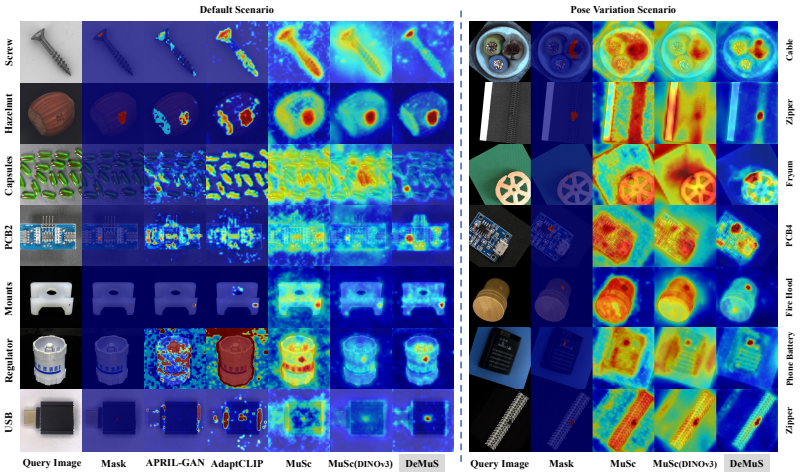

Fig. 5: Qualitative segmentation results under default and Pose Variation scenarios across MVTec AD, VisA, and Real-IAD. Columns from left to right display the query image, ground truth, and the globally normalized anomaly heatmaps.

4.4 Qualitative Analysis

Figure 5 visualizes the pixel-level anomaly maps across different datasets. By referencing intra-batch samples, our approach successfully identifies *subtle anomalies indistinguishable in isolation*, which text-guided methods frequently miss. Moreover, compared to MuSc, our generated maps exhibit significantly higher contrast between anomalous and normal regions, proving that our training effectively disentangles normality from abnormality within the visual feature space.

4.5 Ablation Study

We conduct comprehensive ablation studies on MVTec AD and VisA to validate each component, as shown in Tab. 4.

Pre-trained Backbone Selection. Tab.4 (rows 1–2) compares CLIP [36] and DINOv3 [38] in the training-free setting. DINOv3 yields clear improvements over CLIP on MVTec AD (+2.88% pixel AUROC, +3.18% AUPRO). This is attributed to DINOv3's self-supervised visual features being spatially smooth, whereas CLIP's global contrastive learning often introduces noisy local activations, as illustrated in Fig. 6. DINOv3 is also more amenable to the PCA-based foreground separation used in our training preprocessing [12, 32].

Decoupling of Matching and Scoring. Tab.4 (rows 3–4) evaluate two variants: (1) *Matching-only* (training only g_m, using pre-trained features for scoring) and (2) *Scoring-only* (training only g_s, using it for both matching and scoring). The *Matching-only* variant performs worse than the pre-trained baseline, lacking anomaly awareness. Conversely, the *Scoring-only* variant suffers a 2.49% drop in AP on VisA due to degraded correspondence matching. To investigate this,

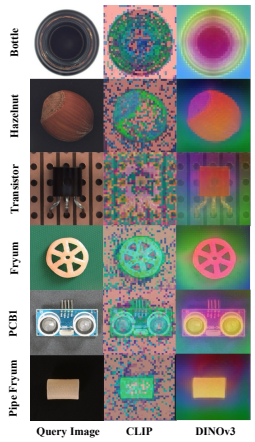

Fig. 6: 3D PCA comparison of features on MVTec AD and VisA.

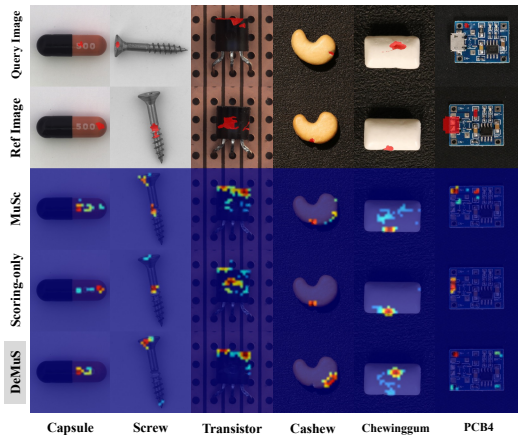

Fig. 7: Matching heatmaps for anomalous patches onto the reference images. Comparison across MuSc, *Scoring-only*, and **DeMuS** on MVTec AD and VisA.

Table 4: Ablation study of key components on MVTec AD and VisA. We evaluate the impact of backbone selection (rows 1-2), decoupling strategy (rows 3-4), adapter architecture (rows 5-6), and training preprocessing (rows 7-8).

Method	MVTec AD				VisA				Average			
	AUC	AP	PAUC	PRO	AUC	AP	PAUC	PRO	AUC	AP	PAUC	PRO
CLIP-pretrain	88.3	94.2	93.6	88.9	82.1	84.2	96.1	89.1	85.2	89.2	94.8	89.0
DINOv3-pretrain	89.3	94.2	96.5	92.1	81.9	83.8	95.9	83.5	85.6	89.0	96.2	87.8
Matching-only	88.4	94.0	96.4	91.8	76.9	79.7	95.3	80.3	82.6	86.8	95.8	86.1
Scoring-only	93.3	96.3	96.4	93.7	88.8	89.4	96.6	92.0	91.0	92.9	96.5	92.9
MLP	89.3	94.4	**96.7**	93.0	89.1	90.1	96.7	91.7	89.2	92.2	96.7	92.3
MLP+Avgpool	89.0	93.9	96.0	89.2	88.9	89.4	**97.7**	92.2	89.0	91.7	**96.9**	90.7
w/o Pose Aug.	91.6	95.0	95.9	92.9	88.3	89.1	97.2	91.4	89.9	92.1	96.5	92.2
Unaligned Oracle	91.2	94.9	96.1	93.6	87.7	89.4	96.7	91.8	89.5	92.1	96.4	92.7
DeMuS	**94.2**	**97.0**	96.4	**93.7**	**90.4**	**91.9**	97.0	**92.3**	**92.3**	**94.5**	96.7	**93.0**

Figure 7 visualizes the top-k nearest neighbor matching heatmaps (Eq. 5). Without g_m, it fails to focus on corresponding physical parts, leading to inaccurate semantic comparisons. Our decoupled design effectively mitigates this.

Multi-Scale Convolutional Adapter. Tab.4 (rows 5–6) shows that replacing our multi-scale convolutional adapter with a standard Multilayer Perceptron (MLP) bottleneck drops image-level performance by 4.91% on MVTec AD (where large-area anomalies are common). Even when applying MuSc-style multi-scale aggregation to the MLP adapter, improvements remain marginal and unstable. This confirms that trainable multi-scale convolutions are crucial for accommodating defects of varying granularities.

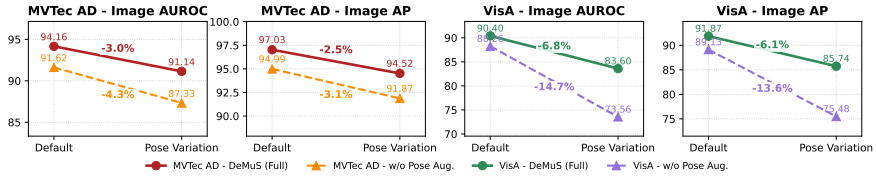

Fig. 8: Robustness degradation analysis under pose variations. We compare the performance trajectories of our full **DeMuS** and the *w/o Pose Aug.* variant across default and pose-varied scenarios.

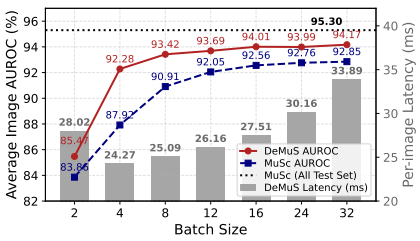

Fig. 9: Accuracy-efficiency trade-off. Image AUROC (left) of DeMuS and MuSc, and inference latency (right) across varying batch sizes.

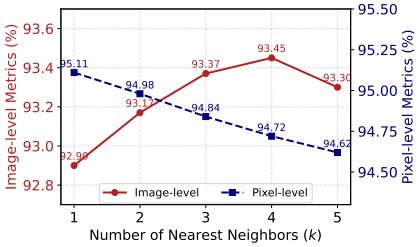

Fig. 10: Nearest neighbors k analysis. Image-level (left) and pixel-level (right) performance metrics evaluated under varying values of k.

Alignment Supervision and Augmentation. Simplifying the Correspondence Oracle to average-pooled NN matching (*Unaligned Oracle*) degrades performance (Tab. 4, rows 7–8). Furthermore, Figure 8 reveals removing affine transformations (*w/o Pose Aug.*) triggers catastrophic failure under pose-varied scenarios, as the model overfits to absolute positional semantics. Crucially, such spatial robustness is not merely a byproduct of augmentation. Instead, our preconstructed alignment supervision acts as the fundamental anchor, enabling the model to digest aggressive pose shifts and learn pose-invariant representations.

Impact of Batch Size. We analyze the accuracy-efficiency trade-off across batch sizes B (Fig. 9). Extremely small settings ($B = 2$) yield sub-optimal accuracy due to insufficient normal references. A substantial leap occurs at $B = 4$ (92.28% average AUROC), concurrently delivering the lowest per-image latency (24.27 ms) via GPU parallelization. For $B > 4$, performance gains saturate while the quadratic matching complexity steadily inflates latency, violating real-time industrial constraints. Notably, DeMuS consistently outperforms MuSc across all sizes. Thus, we adopt $B = 4$ to balance detection precision and speed.

Impact of Top-k Correspondences. As shown in Fig. 10, increasing nearest neighbors (k) yields a clear trade-off: image-level metrics improve while pixel-level metrics degrade. Although a larger k enhances image-level robustness by smoothing out isolated noisy matches, it introduces looser spatial correspondences that dilute exact defect boundaries, hindering fine-grained localization. We adopt $k = 3$ to balance detection reliability and segmentation precision.

5 Conclusion

In this paper, we investigate the practical Batch Zero-Shot IAD setting to address stringent latency and unaligned spatial constraints in agile manufacturing. To overcome the semantic ambiguity and spatial vulnerability of pre-trained features, we propose DeMuS, a decoupled mutual scoring framework. DeMuS mechanistically separates robust spatial alignment (*where-to-match*) from anomaly discrimination (*how-to-score*) via a two-stage auxiliary training paradigm. Extensive experiments demonstrate DeMuS achieves state-of-the-art performance and exceptional robustness against severe pose shifts, offering a highly scalable solution for online inspection.

Limitations and Future Work. DeMuS primarily targets rigid or semi-rigid pose variations. For highly deformable products (e.g., textiles or tangled cables), establishing precise local correspondences remains challenging. Furthermore, our framework still requires an annotated auxiliary dataset to optimize the adapters. Future work will explore unsupervised flow-based matching for complex non-rigid deformations and investigate self-supervised synthetic anomaly generation to completely bypass the reliance on auxiliary annotations.

Acknowledgements

This work was supported by National Key Research and Development Program of China under Grant No. 2024YFB3309602, National Natural Science Foundation of China under Grant No. 62472017.

References

1. Alzarooni, A., Iqbal, E., Khan, S.U., Javed, S., Moyo, B., Abdulrahman, Y.: Anomaly detection for industrial applications, its challenges, solutions, and future directions: A review. arXiv preprint arXiv:2501.11310 (2025)
2. Bai, Y., Dong, Y., Ma, H., Yu, B., Chen, Z., Ma, L., Tian, G.: Patchclip: Enhencing clip with pair-images for zero-shot anomaly detection. In: 2024 China Automation Congress (CAC). pp. 4162–4167. IEEE (2024)
3. Bai, Y., Zhang, J., Cao, Y., Lu, G., He, Q., Li, X., Tian, G.: Bridge feature matching and cross-modal alignment with mutual-filtering for zero-shot anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3534–3543 (2025)
4. Batzner, K., Heckler, L., König, R.: Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 128–138 (2024)
5. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: 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)
6. Cao, W., Yao, X., Xu, Z., Liu, Y., Pan, Y., Ming, Z.: A survey of zero-shot object detection. Big Data Mining and Analytics **8**(3), 726–750 (2025)

7. Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In: European conference on computer vision. pp. 55–72. Springer (2024)
8. Chen, Q., Qu, Z., Luo, W., Yao, H., Cao, Y., Jiang, Y., Duan, Y., Luo, H., Lv, C., Zhang, Z.: Cops: Conditional prompt synthesis for zero-shot anomaly detection. arXiv preprint arXiv:2508.03447 (2025)
9. Chen, X., Han, Y., Zhang, J.: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. arXiv preprint arXiv:2305.17382 (2023)
10. Chen, X., Zhang, J., Tian, G., He, H., Zhang, W., Wang, Y., Wang, C., Liu, Y.: Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In: International joint conference on artificial intelligence. pp. 17–33. Springer (2024)
11. Cheng, Y., Cao, Y., Yao, H., Luo, W., Jiang, C., Zhang, H., Shen, W.: A comprehensive survey for real-world industrial surface defect detection: Challenges, approaches, and prospects. Journal of Manufacturing Systems **84**, 152–172 (2026)
12. Damm, S., Laszkiewicz, M., Lederer, J., Fischer, A.: Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1319–1329. IEEE (2025)
13. Fang, Q., Lv, W., Su, Q.: Af-clip: Zero-shot anomaly detection via anomaly-focused clip adaptation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4846–4855 (2025)
14. Fang, Z., Wang, X., Li, H., Liu, J., Hu, Q., Xiao, J.: Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17481–17490 (2023)
15. Fu, C., Fang, H., Zheng, X., Wei, W., Li, Y., Sun, H., Li, X.: Ssvp: Synergistic semantic-visual prompting for industrial zero-shot anomaly detection. arXiv preprint arXiv:2601.09147 (2026)
16. Gao, B.B., Zhou, Y., Yan, J., Cai, Y., Zhang, W., Wang, M., Liu, J., Liu, Y., Wang, L., Wang, C.: Adaptclip: Adapting clip for universal visual anomaly detection. In: AAAI (2026)
17. He, J., Cao, M., Peng, S., Xie, Q.: Rareclip: Rarity-aware online zero-shot industrial anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24478–24487 (2025)
18. Hua, L., Su, X., Luo, Y., You, S., Long, J.: Hieclip: Hierarchical clip with explicit alignment for zero-shot anomaly detection. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
19. Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, X., Wang, Y.: Few-shot anomaly detection via category-agnostic registration learning. IEEE Transactions on Neural Networks and Learning Systems **36**(7), 13527–13541 (2024)
20. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: 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)
21. Kim, D., Park, C., Cho, S., Lee, S.: Fapm: Fast adaptive patch memory for real-time industrial anomaly detection. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
22. Le-Gia, T., Jaehyun, A.: On the problem of consistent anomalies in zero-shot industrial anomaly detection. arXiv preprint arXiv:2510.10456 (2025)

23. Li, A., Qiu, C., Kloft, M., Smyth, P., Rudolph, M., Mandt, S.: Zero-shot anomaly detection via batch normalization. Advances in Neural Information Processing Systems **36**, 40963–40993 (2023)
24. Li, C., Zhou, S., Kong, J., Qi, L., Xue, H.: Kanoclip: Zero-shot anomaly detection through knowledge-driven prompt learning and enhanced cross-modal integration. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
25. Li, H., Hu, J., Li, B., Chen, H., Zheng, Y., Shen, C.: Target before shooting: Accurate anomaly detection and localization under one millisecond via cascade patch retrieval. IEEE Transactions on Image Processing **33**, 5606–5621 (2024)
26. Li, S., Li, Z., Wang, W., Zheng, L., Lu, Y.: Dnpr: Zero-shot industrial anomaly detection via dynamic normal prototype refinement. Expert Systems with Applications p. 131331 (2026)
27. Li, X., Huang, Z., Xue, F., Zhou, Y.: Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In: The Twelfth International Conference on Learning Representations (2024)
28. Li, X., Xue, F., Zhou, Y.: Musc-v2: Zero-shot multimodal industrial anomaly classification and segmentation with mutual scoring of unlabeled samples. arXiv preprint arXiv:2511.10047 (2025)
29. Li, Z., Yan, Y., Wang, X., Ge, Y., Meng, L.: A survey of deep learning for industrial visual anomaly detection. Artificial Intelligence Review **58**(9), 279 (2025)
30. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., Yu, W.: Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9974–9983 (2025)
31. Ma, W., Zhang, X., Yao, Q., Tang, F., Wu, C., Li, Y., Yan, R., Jiang, Z., Zhou, S.K.: Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4744–4754 (2025)
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
33. Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z., Wu, Q.J.: A review of generalized zero-shot learning methods. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4051–4070 (2022)
34. Qu, Z., Tao, X., Gong, X., Qu, S., Chen, Q., Zhang, Z., Wang, X., Ding, G.: Bayesian prompt flow learning for zero-shot anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30398–30408 (2025)
35. Qu, Z., Tao, X., Prasad, M., Shen, F., Zhang, Z., Gong, X., Ding, G.: Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In: European Conference on Computer Vision. pp. 301–317. Springer (2024)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
37. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)

38. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
41. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
42. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22883–22892 (2024)
43. Wang, S., Wang, G., Liu, X., Liu, J., Liu, J., Wang, S.: Scad: A self-constrained solution to automate context-guided zero-shot image anomaly detection. Neural Networks p. 108577 (2026)
44. Wei, S., Jiang, J., Xu, X.: Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 9994–10003 (June 2025)
45. Yuan, J., Ye, J., Chen, W., Gao, C.: Ad-dinov3: Enhancing dinov3 for zero-shot anomaly detection with anomaly-aware calibration. arXiv preprint arXiv:2509.14084 (2025)
46. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
47. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024)
48. Zhang, Z., Deng, H., Bao, J., Li, X.: Dual-image enhanced clip for zero-shot anomaly detection. arXiv preprint arXiv:2405.04782 (2024)
49. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: The Twelfth International Conference on Learning Representations (2023)
50. Zhu, J., Ong, Y.S., Shen, C., Pang, G.: Fine-grained abnormality prompt learning for zero-shot anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22241–22251 (2025)
51. Zhu, J., Pang, G.: Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17826–17836 (2024)
52. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European conference on computer vision. pp. 392–408. Springer (2022)