ADPRETRAIN++: DENSE DUAL-TASK ANOMALY REPRESENTATION PRETRAINING FOR INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised industrial anomaly detectors such as PatchCore, PaDiM and CFLOW score images and pixels with frozen features from generic backbones, which were never trained to separate normal from anomalous patterns. Residual anomaly-representation pretraining narrows this mismatch, but it optimises a single global objective on one feature space, even though image-level detection and pixel-level localisation require different evidence. We propose ADPretrain++, which pretrains residual features on an industrial dataset with real defect masks through a shared learnable-reference projector followed by an image adapter and a pixel adapter. The pixel adapter is trained with a dense norm-map loss and prototype shaping, and the image adapter with top-k multiple-instance learning and descriptor shaping. The pretrained features replace the backbone features of an unchanged detector. With a DINOv3-L backbone and averaged over four benchmarks, the features raise PatchCore by 4.8 points of image-level area under the ROC curve (I-AUROC) and 7.3 points of area under the per-region overlap curve (AUPRO), raise PaDiM by 1.1 and 1.6 points, and leave CFLOW essentially unchanged. Ablations attribute the gains to the dual-task objectives and the shared projector, and evaluation-time routing with PatchCore shows that the image adapter serves image scores while the pixel adapter serves localisation maps.

1 INTRODUCTION

Industrial anomaly detection (AD) must flag defective products and localise the defective region while training almost exclusively on normal images, because defects are rare, diverse and often unknown before deployment. The dominant unsupervised detectors therefore model normality in the feature space of a pretrained backbone. PaDiM fits per-location Gaussians (Defard et al., 2021), PatchCore retrieves nearest normal patches from a memory bank (Roth et al., 2022), CFLOW estimates normalising-flow likelihoods (Gudovskiy et al., 2022), and UniAD and GLASS reconstruct features or discriminate synthetic anomalies on top of them (You et al., 2022; Chen et al., 2024). Every one of these scoring rules succeeds only when anomalous regions are already distinguishable in the backbone features. Those features come from objectives unrelated to that requirement. ImageNet classification, contrastive and self-distillation learning, and vision–language alignment reward invariance to exactly the small local changes that constitute many industrial defects (Deng et al., 2009; He et al., 2020; Oquab et al., 2024; Radford et al., 2021).

ADPretrain++ keeps the normal-reference residual formulation and restructures what is learned on top of it (Figure 1). Low-rank adapters on the last backbone blocks adjust the encoder to industrial imagery without rewriting its generic features (Hu et al., 2022). Residuals from four backbone levels pass through a shared learnable-reference projector, which subtracts an attention read-out over learnable normal-like tokens, and then branch into two small task adapters. Real defect masks supervise the pixel route through a dense norm-map loss and prototype-based token shaping, whereas image labels supervise the image route through a top-k multiple-instance learning (MIL) loss and descriptor shaping. The ADPretrain objective is retained as a regulariser on the shared space. At evaluation time, a detector consumes the pixel route for localisation and, where the detector produces a separate image score, the image route for detection.

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.

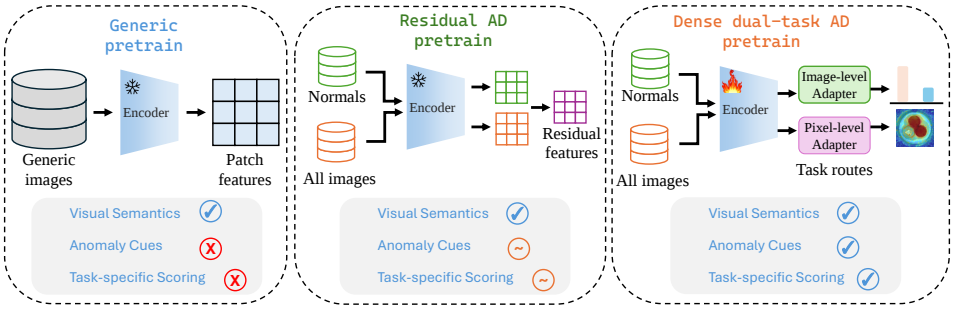

Figure 1: Three pretraining regimes for industrial AD features. Generic pretraining (left) supplies visual semantics without anomaly cues. Residual AD pretraining (middle) compares queries with normal references but trains one feature space for all uses. Dense dual-task pretraining (right, ADPretrain++) additionally adapts the encoder and serves image-level scores and pixel-level maps through separate task routes.

The paper makes three contributions. First, it identifies the single global objective of residual AD pretraining as a limitation for detectors that output both image scores and pixel maps, and it formulates detector-agnostic pretraining as learning one shared residual space with two task-specific routes. Second, it proposes dense dual-task objectives that supervise the pixel route with real defect masks and the image route with image labels. Third, it evaluates the pretrained features as a drop-in replacement in five detectors on five benchmarks (Section 4.2) and analyses the mechanism through component and objective ablations, evaluation-time routing, reference-shot and data-size studies, and token-norm statistics. The evaluation also reports the detectors and datasets on which the features do not help.

2 Related Work

Unsupervised industrial AD is dominated by detectors that model normality in a fixed pretrained feature space. Embedding methods fit per-location Gaussians (Defard et al., 2021), retrieve nearest normal patches (Roth et al., 2022), or compact normal features around a centre (Tax & Duin, 2004), and later variants add structural guidance or noise-tolerant memory banks (Yang et al., 2024; Wang et al., 2025a). Flow-based detectors estimate likelihoods over feature pyramids (Gudovskiy et al., 2022) and have been extended with multi-scale attention and vector quantisation (Hu et al., 2025; Zhou et al., 2025). Reconstruction and distillation methods assume that anomalous regions are reconstructed or matched less accurately (You et al., 2022; Guo et al., 2025; Hu et al., 2026; Yu et al., 2025; Sakai et al., 2026), while GLASS, RealNet and text-guided generation train discriminators against synthesised anomalies (Chen et al., 2024; Zhang et al., 2024; Lee & Choi, 2024). These families differ in scoring rule but share one dependency, since anomalous evidence must already appear as distance, density or reconstruction error in the input features. ADPretrain++ leaves the detectors unchanged and replaces only their features, so five of them serve as probes of representation quality.

Foundation models now supply most of these features. Dense self-supervised backbones such as DINOv2 and DINOv3 provide spatially precise tokens (Oquab et al., 2024; Siméoni et al., 2025), and AnomalyDINO shows that nearest-neighbour matching of DINOv2 patches is already a strong few-shot detector (Damm et al., 2025). CLIP-based methods exploit text–image alignment and learned prompts for zero- and few-shot detection (Radford et al., 2021; Zhou et al., 2024; Ma et al., 2025; He et al., 2025; Zhu et al., 2025), although multimodal benchmarks report that category-level language describes small industrial defects poorly (Jiang et al., 2025). Low-rank adaptation makes it possible to tune such backbones with few trainable parameters (Hu et al., 2022). In all of these cases the pretraining objective remains generic, and anomaly sensitivity is obtained through the detector or the prompt rather than through the representation.

AD-oriented pretraining changes the representation itself. SPD learns from a spot-the-difference pretext task (Zou et al., 2022), ToCoAD combines synthetic anomalies with contrastive stages (Liang et al., 2025), dense pretraining targets localisation-oriented features (He et al., 2024), and AGSSP uses

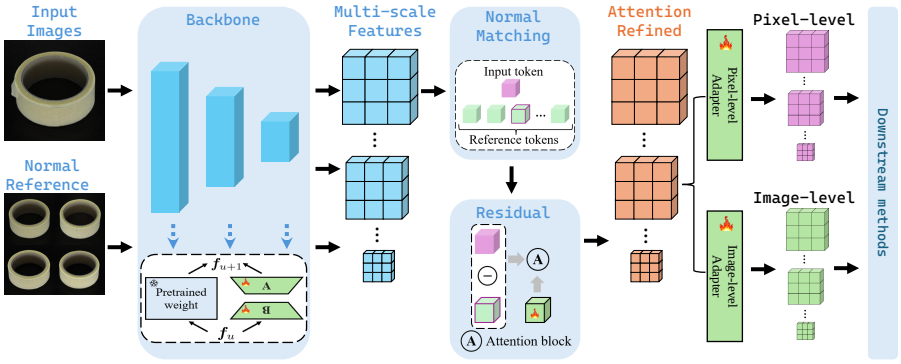

Figure 2: Overview of ADPretrain++. Multi-scale features of the query image and of K normal references are extracted by a backbone with LoRA adapters. Each query token is matched to its nearest reference token and the difference forms a residual token. A shared learnable-reference attention projector refines the residuals, and an image-level adapter A_I and a pixel-level adapter A_P produce the two feature routes consumed by downstream detectors.

pseudo defect maps on a large metallic-surface corpus (Liu et al., 2026). These pretexts depend on how pseudo anomalies are synthesised or on a single defect domain. Residual learning offers a more transferable signal, because subtracting features of normal references suppresses category appearance, which underlies class-generalisable and cross-domain detectors (Yao et al., 2024; Zhou et al., 2026; Rahmaniar & Suzuki, 2026; Wang et al., 2025b) and the pretraining method ADPretrain (Yao et al., 2025), trained on the real anomalies of Real-IAD (Wang et al., 2024). ADPretrain++ adopts this residual formulation and replaces its single global objective with separately supervised image and pixel routes. Its image-route objective follows the multiple-instance view of weakly labelled anomaly detection, in which a bag is anomalous if its highest-scoring instances are (Sultani et al., 2018); here the bag is an image and the instances are its tokens.

3 Method

ADPretrain++ learns a feature extractor, not a detector. Pretraining uses an industrial dataset with image labels $y \in \{0, 1\}$ and defect masks $\boldsymbol{M} \in \{0, 1\}^{H \times W}$; at evaluation time, the pretrained features replace the backbone features of an existing detector whose scoring rule and hyperparameters are unchanged. Figure 2 summarises the pipeline, and Appendix A.1 lists the notation.

3.1 Residual Features from Normal Reference Images

A backbone E_θ maps an RGB image $\boldsymbol{I}$ of size $H \times W$ to features at four levels, $\boldsymbol{F} = E_\theta(\boldsymbol{I})$. For readability the equations below describe one level, whose N tokens $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N$ of dimension C form the rows of $\boldsymbol{F}$; the same operations are applied to every level. The encoder is adapted with low-rank adapters on the linear layers of its last four blocks, so that each frozen weight $\boldsymbol{W}$ is used as $\boldsymbol{W} + (\alpha_r/r)\boldsymbol{B}\boldsymbol{A}$ with trainable rank-r factors $\boldsymbol{A}$ and $\boldsymbol{B}$ and scaling α_r (Hu et al., 2022).

A set of K normal reference images of the same product is encoded by the same backbone into a token bank $\boldsymbol{R}$ with rows $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M$. Each query token is explained by its nearest normal token, and the unexplained part forms the residual token:

$$\hat{\boldsymbol{r}}_u = \arg\min_{\boldsymbol{r}_m \in \boldsymbol{R}} \|\boldsymbol{f}_u - \boldsymbol{r}_m\|_2, \qquad \boldsymbol{x}_u = \boldsymbol{f}_u - \hat{\boldsymbol{r}}_u, \tag{1}$$

where u indexes query tokens and the residual tokens $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ form the rows of $\boldsymbol{X}$. Subtracting the nearest normal token removes the appearance shared with normal samples of the product, so the residual carries mainly the deviation from normality, following Yao et al. (2025). During pretraining K is drawn from $\{1, 4, 8\}$ for each batch; evaluation uses $K = 8$.

3.2 Learnable-Reference Projector and Task Adapters

A residual computed against a few references still contains normal variation that the bank does not cover, such as pose or illumination differences. The projector therefore removes a learned normal-like component before mapping residuals into the pretrained space:

$$\boldsymbol{H} = \boldsymbol{X} - \mathrm{Attn}(\boldsymbol{X}, \boldsymbol{Q}, \boldsymbol{Q}), \qquad \boldsymbol{Y} = P_\phi(\boldsymbol{H}) \in \mathbb{R}^{N \times D}, \tag{2}$$

where $\boldsymbol{Q} \in \mathbb{R}^{N_r \times C}$ holds $N_r = 2048$ learnable reference tokens. $\mathrm{Attn}(\cdot,\cdot,\cdot)$ is multi-head attention with 16 heads; its queries are the residual tokens and its keys and values are $\boldsymbol{Q}$. P_ϕ is a one-layer MLP residual update followed by a channel projection to width D. The four feature levels are processed in parallel and aligned by P_ϕ.

The shared features $\boldsymbol{Y}$ then branch into two task adapters:

$$\boldsymbol{Z}^I = A_I(\boldsymbol{Y}), \qquad \boldsymbol{Z}^P = A_P(\boldsymbol{Y}), \tag{3}$$

where the image adapter A_I and the pixel adapter A_P are each a residual depth-wise convolution block with group normalisation and a learnable output scale. The adapters are deliberately small (Appendix G). They reshape the shared space for one output granularity without introducing a detector-specific head, so that differences between routes can be attributed to the representation.

3.3 Dense Dual-Task Objectives

Each route receives the supervision that matches its output. The training objective is

$$\begin{aligned} \mathcal{L} &= \lambda_{\mathrm{pix}}\mathcal{L}_{\mathrm{pix}} + \lambda_{\mathrm{img}}\mathcal{L}_{\mathrm{img}} + \lambda_{\mathrm{legacy}}\mathcal{L}_{\mathrm{legacy}}, \\ \mathcal{L}_{\mathrm{pix}} &= \lambda_{\mathrm{dn}}\mathcal{L}_{\mathrm{dn}} + \lambda^P_{\mathrm{shape}}\mathcal{L}^P_{\mathrm{shape}}, \qquad \mathcal{L}_{\mathrm{img}} = \lambda_{\mathrm{mil}}\mathcal{L}_{\mathrm{mil}} + \lambda^I_{\mathrm{desc}}\mathcal{L}^I_{\mathrm{desc}}, \end{aligned} \tag{4}$$

where the λ coefficients are fixed loss weights (Appendix B) and $\mathcal{L}_{\mathrm{legacy}}$ is the angle-and-norm objective of ADPretrain (Yao et al., 2025), retained on the shared space as a regulariser. Figure 3

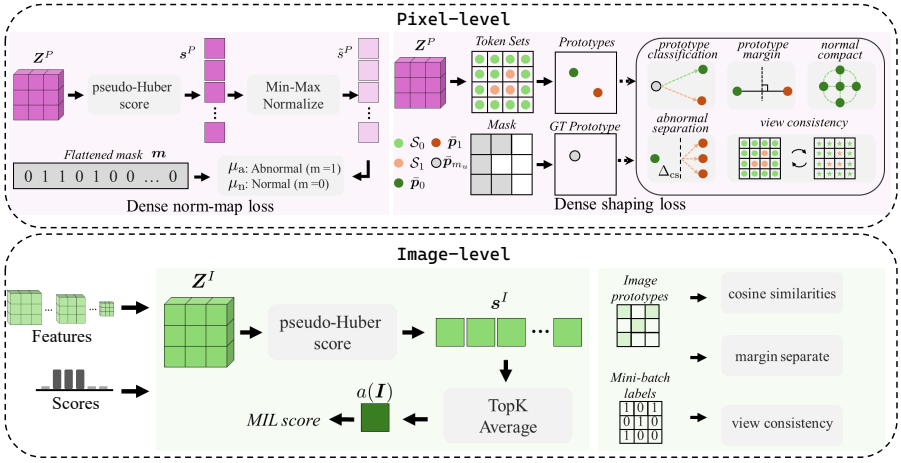

Figure 3: Training objectives of ADPretrain++. The pixel-level objectives comprise the dense norm-map loss, computed from normalised pseudo-Huber scores and the resized mask, and the dense shaping loss, built from mask-derived token sets and prototypes. The image-level objectives comprise the top-k MIL score and descriptor shaping with in-batch image prototypes.

illustrates how the four route-specific objectives act on the pixel-route and image-route features, and Appendix A.1 gives their complete definitions.

Both routes score a token by a pseudo-Huber radial energy, which is quadratic near the origin and linear for large residuals:

$$\psi_\delta(\boldsymbol{z}) = \delta^2\big((1 + \|\boldsymbol{z}\|_2^2/\delta^2)^{1/2} - 1\big), \qquad \delta = 1, \tag{5}$$

so that $s_u^P = \psi_\delta(\boldsymbol{z}_u^P)$ and $s_u^I = \psi_\delta(\boldsymbol{z}_u^I)$ for pixel-route and image-route tokens. Because each token is a residual against normal references, a larger radial energy indicates more local evidence that the references do not explain.

The dense norm-map loss $\mathcal{L}_{\mathrm{dn}}$ trains the pixel route to reproduce the defect mask in its norm map. With $\tilde{s}_u^P$ the per-image min–max normalised pixel score and $\boldsymbol{m}$ the mask resized to the token grid, the loss is

$$\mathcal{L}_{\mathrm{dn}} = \mathcal{L}_{\mathrm{focal}}(\tilde{\boldsymbol{s}}^P, \boldsymbol{m}) + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}}(\tilde{\boldsymbol{s}}^P, \boldsymbol{m}) + \big[\mu_{\mathrm{n}} + \Delta_{\mathrm{dn}} - \mu_{\mathrm{a}}\big]_+, \tag{6}$$

where $\mathcal{L}_{\mathrm{focal}}$ is the focal loss (Lin et al., 2017), $\mathcal{L}_{\mathrm{dice}}$ is the soft Dice loss (Milletari et al., 2016), μ_{a} and μ_{n} are the mean normalised scores over anomalous and normal tokens, Δ_{dn} is a margin and $[\cdot]_+ = \max(\cdot, 0)$. The hinge raises anomalous scores and lowers normal scores until their gap reaches Δ_{dn}. The normalisation is computed within each image, so $\mathcal{L}_{\mathrm{dn}}$ constrains the ordering of tokens inside an image but not the absolute norm across images; Section 4.6 examines this consequence. The dense shaping loss $\mathcal{L}_{\mathrm{shape}}^P$ acts on the direction of pixel-route tokens. It classifies each token against mask-derived normal and anomalous prototypes by cosine similarity and keeps normal tokens compact around the normal prototype. It also pushes anomalous tokens away from that prototype and enforces consistency between two augmented views (Eq. 12).

The image route is supervised with image labels only. An anomalous image may contain a defect covering a few tokens, so averaging all token scores would dilute the evidence. Following the multiple-instance view that a bag is positive if its strongest instances are (Sultani et al., 2018), the image score averages the k largest token energies:

$$a(\boldsymbol{I}) = \frac{1}{k} \sum_{u \in \mathrm{TopK}(\boldsymbol{s}^I, k)} s_u^I, \qquad \mathcal{L}_{\mathrm{mil}} = -y \log \sigma\big(a(\boldsymbol{I})\big) - (1-y) \log\big(1 - \sigma(a(\boldsymbol{I}))\big), \tag{7}$$

where $\mathrm{TopK}(\boldsymbol{s}^I, k)$ returns the indices of the k largest entries of $\boldsymbol{s}^I$, k is 5% of the tokens, and σ is the logistic sigmoid. The descriptor shaping loss $\mathcal{L}_{\mathrm{desc}}^I$ complements this scalar decision. It pools image-route tokens with score-dependent softmax weights into a descriptor and classifies the descriptor against in-batch normal and anomalous prototypes. It further separates the two prototypes by a margin and aligns descriptors of two augmented views (Eq. 15).

3.4 Plugging the Features into Downstream Detectors

A downstream detector receives the four-level token maps of the chosen route in place of its backbone features and builds its normal model from the training images of each category exactly as with the original features. The route follows the output a detector computes. PaDiM, CFLOW, GLASS and UniAD derive both scores from dense feature maps and consume the pixel route $\boldsymbol{Z}^P$. PatchCore consumes $\boldsymbol{Z}^P$ for its anomaly map and $\boldsymbol{Z}^I$ for its image score, since its image score is the maximum patch distance and benefits from the image-route training in Eq. 7. We additionally report a detector-free feature-norm probe (FeatureNorm), which scores each token by its ℓ_2 norm; it measures how much anomaly information the magnitude of the pretrained features carries and uses the image route by default. The residual construction of Eq. 1 requires K normal images of the product at test time, which the unsupervised protocol already provides through the training set.

4 Experiments

4.1 Setup

Pretraining uses Real-IAD (Wang et al., 2024), a multi-view industrial dataset with image labels and pixel-level defect masks. Downstream evaluation uses five benchmarks whose categories do not occur in Real-IAD, namely MVTec AD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), MVTec 3D-AD (Bergmann et al., 2022) and MPDD (Jezek et al., 2021). Averages that exclude MPDD are referred to as four-benchmark averages. Each benchmark is used with its official split, in which training images are normal and test images carry image labels and defect masks.

The pretrained features are evaluated with five unsupervised detectors that cover the main scoring families, namely PaDiM (Gaussian embedding), PatchCore (memory-bank retrieval), CFLOW (normalising flow), GLASS (discrimination against synthesised anomalies) and UniAD (feature

Table 1: Main results with a DINOv3-L backbone on five industrial anomaly-detection benchmarks. Each cell reports I-AUROC/AUPRO in % (both ↑). Original uses the raw frozen backbone features; ADPretrain and ADPretrain++ replace them with pretrained residual features while every detector is left unchanged. Highlighting compares the feature sources within each block; Avg. is the unweighted average over the five datasets, and the last block averages the five detectors. The four-benchmark averages quoted in the text exclude the MPDD column.

Detector	Features	MVTec AD	VisA	BTAD	MVTec 3D-AD	MPDD	Avg.
PaDiM	Original	99.0/93.0	93.5/84.7	96.7/74.2	83.4/87.7	88.7/89.6	92.3/85.8
	ADPretrain	**99.1/93.3**	94.5/85.7	**96.8**/74.6	84.7/88.4	90.0/90.1	93.0/86.4
	ADPretrain++	**99.1/93.3**	**95.2/87.3**	**96.8/75.5**	**85.9/90.0**	**91.2/91.4**	**93.6/87.5**
PatchCore	Original	98.0/84.7	88.5/71.5	94.2/67.1	76.5/78.4	85.4/82.3	88.5/76.8
	ADPretrain	98.6/85.1	91.6/73.9	94.6/67.4	81.8/81.4	87.7/84.4	90.9/78.4
	ADPretrain++	**99.3/89.3**	**95.2/85.3**	**95.0/68.1**	**87.0/88.3**	**90.0/89.2**	**93.3/84.0**
CFLOW	Original	98.2/92.1	94.9/89.5	**95.9/74.2**	**90.0/92.2**	91.8/92.5	94.2/88.1
	ADPretrain	98.3/92.4	95.3/89.8	95.5/72.9	89.9/92.1	92.0/92.6	94.2/88.0
	ADPretrain++	**98.4/93.2**	**95.7/90.6**	95.4/72.6	89.9/92.1	**92.1/92.8**	**94.3/88.3**
GLASS	Original	98.9/93.4	94.8/88.6	95.6/73.8	86.2/90.1	93.4/92.1	93.8/87.6
	ADPretrain	99.2/93.7	96.0/89.4	95.8/74.0	87.8/90.7	94.5/92.6	94.7/88.1
	ADPretrain++	**99.5/94.4**	**97.2/91.2**	**96.0/74.5**	**89.3/92.0**	**95.6/93.6**	**95.5/89.1**
UniAD	Original	95.6/88.9	91.1/81.9	92.8/70.8	79.3/83.1	86.6/84.2	89.1/81.8
	ADPretrain	96.7/89.9	93.2/83.7	93.3/71.3	82.0/84.8	88.4/85.5	90.7/83.0
	ADPretrain++	**97.7/92.3**	**95.3/88.0**	**93.8/72.4**	**84.6/88.9**	**90.2/88.6**	**92.3/86.0**
All detectors	Original	97.9/90.4	92.6/83.2	95.0/72.0	83.1/86.3	89.2/88.1	91.6/84.0
	ADPretrain	98.4/90.9	94.1/84.5	95.2/72.0	85.2/87.5	90.5/89.0	92.7/84.8
	ADPretrain++	**98.8/92.5**	**95.7/88.5**	**95.4/72.6**	**87.3/90.3**	**91.8/91.1**	**93.8/87.0**

ADPretrain entries of PaDiM and PatchCore on MVTec AD and VisA come from the ADPretrain-objective checkpoint of the ablation study (Table 2).

reconstruction) (Defard et al., 2021; Roth et al., 2022; Gudovskiy et al., 2022; Chen et al., 2024; You et al., 2022). All features are extracted from a DINOv3-L backbone (Siméoni et al., 2025) at an input resolution of 224×224. Three feature sources are compared, namely the original backbone features (Original), residual features pretrained with the ADPretrain objective on the same backbone and data, and ADPretrain++ features. Only the features change between rows; each detector keeps its scoring rule and its default hyperparameters, and no hyperparameter is tuned on a downstream benchmark. Image-level detection is measured by I-AUROC and localisation by AUPRO, both in percent; Appendix C reports five further image- and pixel-level metrics. Each configuration is run once, and averages are unweighted over categories within a dataset and over datasets across columns. In every table, bold and underlined entries mark the best and second-best values within the comparison group named in the caption. Appendix B lists the remaining implementation details.

The Section 4.3 ablations use a lighter protocol so that every variant can be trained under the same schedule. Each variant is pretrained with the ablation schedule and evaluated with PaDiM, PatchCore and the FeatureNorm probe on MVTec AD and VisA, and the six detector–dataset scores are averaged. This checkpoint differs from the main-table checkpoint, so ablation numbers are compared only with each other.

4.2 Main Results

Table 1 reports I-AUROC and AUPRO for every detector, benchmark and feature source, and Figure 4 isolates the change relative to the original features. The gain depends on the detector. PatchCore, whose scores are raw distances in feature space, benefits most. In the four-benchmark average, its I-AUROC rises from 89.3 to 94.1 and AUPRO from 75.4 to 82.8, with the largest changes on VisA (+6.7 and +13.8) and MVTec 3D-AD (+10.5 and +9.9). PaDiM gains 1.1 I-AUROC and 1.6 AUPRO points over the same four benchmarks, and its gains are again concentrated on VisA and MVTec 3D-AD. GLASS and UniAD improve by 1.7/1.5 and 3.2/4.2 points on the five-benchmark average.

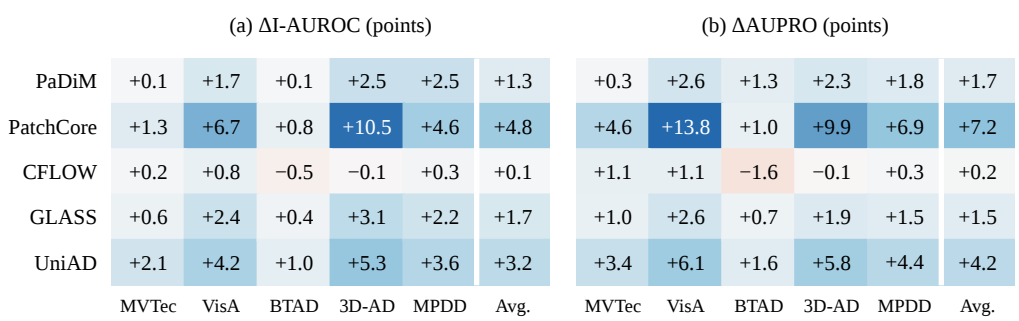

Figure 4: Change produced by ADPretrain++ features relative to the Original DINOv3-L features, in percentage points, for each detector and benchmark. Blue cells mark improvements and red cells mark degradations; Avg. is the unweighted average over the five benchmarks, whereas the four-benchmark averages quoted in the text exclude MPDD.

Table 2: Component and objective ablation on MVTec AD and VisA. Each entry averages PaDiM, PatchCore and FeatureNorm over both datasets (six detector–dataset pairs); all pretrained variants share one ablation checkpoint schedule that differs from the main-table checkpoint. I-AUROC and AUPRO in % (↑); Mean is their average.

Variant	I-AUROC	AUPRO	Mean
Components			
Original DINOv3-L features	80.7	56.4	68.6
ADPretrain objective and projector	95.6	84.7	90.1
w/o shared projector	94.1	85.0	89.6
w/o LoRA on the backbone	96.1	86.1	91.1
one shared route for all tasks	96.1	**87.5**	**91.8**
Objectives			
w/o $\mathcal{L}_{\text{dn}}$	96.3	85.5	90.9
w/o $\mathcal{L}^{P}_{\text{shape}}$	<u>96.4</u>	86.1	91.2
w/o $\mathcal{L}_{\text{mil}}$	95.9	86.6	91.3
w/o $\mathcal{L}^{I}_{\text{desc}}$	96.2	86.7	91.5
w/o $\mathcal{L}_{\text{legacy}}$	96.3	86.4	91.3
$\mathcal{L}_{\text{pix}}$ only ($\mathcal{L}_{\text{dn}} + \mathcal{L}^{P}_{\text{shape}}$)	95.8	86.5	91.2
$\mathcal{L}_{\text{img}}$ only ($\mathcal{L}_{\text{mil}} + \mathcal{L}^{I}_{\text{desc}}$)	96.3	84.9	90.6
ADPretrain++ (full)	**96.5**	<u>87.0</u>	<u>91.7</u>

CFLOW is the exception. Its average change over the four benchmarks is 0.1 points in both metrics; it improves slightly on MVTec AD and VisA but loses 0.5 I-AUROC and 1.6 AUPRO on BTAD and 0.1 points on MVTec 3D-AD. A normalising flow learns its own density transform on top of the input features, so a representation that already separates normal from anomalous tokens leaves it little to add. On BTAD, the pretrained space appears to discard part of the variation that the flow had exploited, and PaDiM and PatchCore also gain at most 1.3 points there. Averaged over all five detectors and five benchmarks, the original features reach 91.6/84.0, ADPretrain features 92.7/84.8 and ADPretrain++ features 93.8/87.0. The advantage over ADPretrain is larger in AUPRO than in I-AUROC, which is the pattern expected if the additional objectives mainly improve dense localisation; Section 4.3 tests this interpretation directly. Appendix E repeats the comparison with CLIP-L and DINOv2-L backbones, where the ranking of the three feature sources is the same on average but the margin over ADPretrain shrinks to a few tenths of a point.

4.3 Component and Objective Ablation

Table 2 removes one component or objective at a time under the ablation protocol of Section 4.1. Pretraining is responsible for most of the improvement over the original features, but its form matters.

Table 3: Feature routing at evaluation time with the main ADPretrain++ checkpoint. Entries are I-AUROC/AUPRO in % (↑), averaged over the four benchmarks other than MPDD. Default routes image-level scores through A_I and pixel-level maps through A_P; FeatureNorm scores the image-route norm by default. Highlighting compares the routes within each column; – denotes a route that does not apply to, or was not evaluated for, that detector.

Route	PatchCore	PaDiM	FeatureNorm
Task-specific adapters (default)	**94.1/82.8**	**94.3/86.5**	**93.6**/16.4
Shared features for all tasks	93.2/81.6	93.9/**86.5**	92.9/**83.6**
w/o pixel adapter A_P	**94.1**/81.6	93.9/**86.5**	–
w/o image adapter A_I	93.2/**82.8**	–	–

The ADPretrain objective with its projector already raises the mean from 68.6 to 90.1, and the full ADPretrain++ objective adds 0.9 I-AUROC and 2.3 AUPRO points on top of it. The asymmetry between the two metrics supports the motivation in Section 1, namely that the global objective leaves more room in localisation than in detection. Removing the learnable-reference projector lowers I-AUROC by 2.4 points and AUPRO by 2.0 points relative to the full model, the largest drop among the components, which indicates that subtracting a learned normal-like component is needed in addition to subtracting the nearest reference token. Removing LoRA costs 0.4 I-AUROC and 0.9 AUPRO, so encoder adaptation contributes but is not the main source of the gain.

The objective rows separate the two routes. Removing the dense norm-map loss causes the largest AUPRO drop among single objectives (1.5 points), and training with the image-level objectives alone yields the lowest AUPRO of all objective variants (84.9), whereas removing the MIL loss causes the largest I-AUROC drop among single objectives (0.6 points). Each group of objectives therefore mainly serves the output it supervises, and neither group alone reaches the full model. A single shared route trained with all objectives is the one variant that is competitive with the full model. It loses 0.4 I-AUROC but gains 0.5 AUPRO under this protocol. The dual-task objectives thus account for most of the localisation gain even without the split, and the split mainly benefits the image score. Section 4.4 examines this trade-off with the main checkpoint.

4.4 Evaluation-Time Routing

Table 3 keeps the main checkpoint fixed and changes only which features a detector reads. For PatchCore the two adapters dissociate cleanly. Replacing the image-route input to the image score with shared features lowers I-AUROC from 94.1 to 93.2 and leaves AUPRO unchanged, and replacing the pixel-route input to the anomaly map lowers AUPRO from 82.8 to 81.6 and leaves I-AUROC unchanged. PaDiM, which derives both outputs from the same dense map, gains 0.4 I-AUROC from the pixel adapter and nothing in AUPRO. The FeatureNorm probe shows that a route can separate images without delineating regions. The image-route norm gives the best I-AUROC but only 16.4 AUPRO, whereas the shared-feature norm reaches 83.6 AUPRO. The top-k MIL objective rewards the few largest token energies in an anomalous image and places no constraint on where the remaining energy lies, which is consistent with this behaviour. Appendix J gives the per-dataset routing results and Table 7 the FeatureNorm results per benchmark.

4.5 Reference Shots and Training Data

The residual construction depends on normal reference images, and the downstream detectors depend on normal training images; Figure 5 varies both. With a single reference image, PatchCore already reaches 93.9 I-AUROC and 82.2 AUPRO, within 0.6 points of the default $K = 8$, and PaDiM is within 0.4 points. The benefit of the pretrained features therefore does not rely on a large reference bank. FeatureNorm behaves differently. Its I-AUROC rises from 90.8 to 93.6 as K grows from 1 to 8 while its AUPRO falls from 21.4 to 16.4, which is consistent with a larger bank lowering the residual energy of normal tokens in the image score without making the image-route map more precise.

When the detectors see only 10% of the VisA training images, ADPretrain++ features give 93.7/85.3 with PatchCore, 94.1/87.9 with PaDiM and 95.6/91.0 with CFLOW; all three exceed the scores of the original features trained on the full set (88.5/71.5, 93.5/84.7 and 94.9/89.5). On VisA, the pretrained

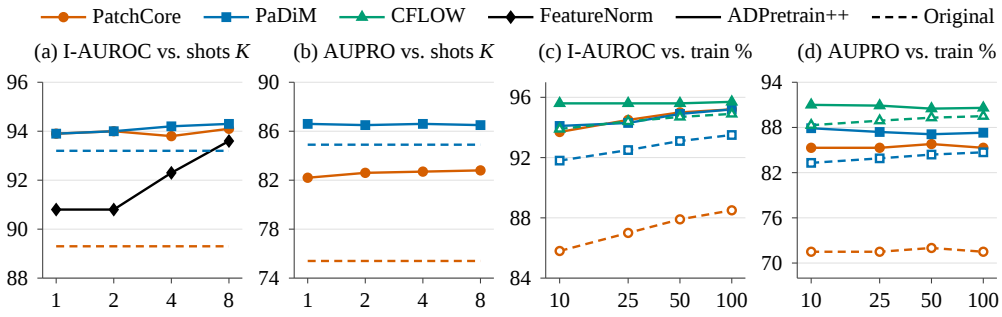

Figure 5: Sensitivity to the number of normal reference images K at evaluation time, as four-benchmark averages (a, b), and to the fraction of normal training images available to the detector on VisA (c, d). Solid lines use ADPretrain++ features and dashed lines the Original features, which do not use reference images and are therefore constant in (a) and (b).

features therefore reduce the amount of normal data a detector needs for a new product. The effect of pretraining data amount is reported in Appendix F.

4.6 Token-Norm Statistics

The dense norm-map loss constrains token scores after per-image normalisation (Eq. 6), which leaves the absolute norm free across images. Pooling all tokens of one category from each of BTAD, MVTec AD and VisA (Figure 6, Appendix H) confirms this consequence. On the shared and image routes, the token norm separates anomalous from normal tokens with an AUROC of 0.85 to 0.92, whereas on the pixel route the pooled AUROC falls below chance, to 0.35 to 0.44. The pixel route still carries anomaly information, since the detectors that consume it gain in AUPRO (Section 4.2). Its norm orders tokens within an image, whereas the evidence that holds across images resides in the full token vector, which PaDiM, PatchCore and CFLOW model.

5 Limitations and Conclusion

Several limitations qualify the results. The pretrained features do not make memory-based detectors robust to contaminated training data. When a small fraction of the images used to build the normal model is anomalous, PatchCore and PaDiM lose about a quarter of their I-AUROC at 5% contamination and fall below chance level from 10%, whereas CFLOW is nearly unaffected (Appendix I). A plausible explanation is that anomalous training tokens enter the memory bank or Gaussian model as normal entries, after which similar test defects are matched to them and scored as normal. The pixel-route norm is calibrated only within an image (Section 4.6), so the pixel route should be consumed by a detector that models the full feature vector rather than by a raw norm threshold. All results rely on a single pretraining source, Real-IAD, and on a single seed per configuration; the smallest differences in Tables 2 and 3, a few tenths of a point, may lie within the range that seed variation produces. Finally, the evaluation is quantitative only, and per-pixel anomaly maps have not been inspected for systematic failure patterns such as boundary bleeding or texture false positives.

Within these limits, the experiments show that residual anomaly pretraining benefits from supervising image-level and pixel-level outputs separately. Dense objectives driven by real defect masks improve localisation beyond what a single global objective achieves, top-k MIL on image labels serves the image score, and a learnable-reference projector is needed on top of nearest-reference subtraction. The route split matters less than the objectives: it serves the separate image score of PatchCore, whereas one shared route trained with all objectives is on par in the ablation average. As a drop-in replacement for backbone features, ADPretrain++ helps PatchCore, which scores raw feature distances, the most, improves the Gaussian, discriminative and reconstruction-based detectors by smaller margins, and leaves a normalising-flow detector essentially unchanged.

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. 9584–9592, 2019.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pp. 202–213, 2022.

Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 37–54. Springer, 2024.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1319–1329, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, Lecture Notes in Computer Science, pp. 475–489. Springer, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20405–20415, 2025.

Haitian He, Sarah Erfani, Mingming Gong, and Qiuhong Ke. Learning transferable representations for image anomaly localization using dense pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1113–1122, 2024.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ronglong Hu, Jizheng Yi, Xiaoyao Li, Aibin Chen, and Guangjie Han. Efficient industrial anomaly detection via cross-scale distillation with enhanced feature compression. *Pattern Recognition*, 173: 112761, 2026.

Zhengnan Hu, Xiangrui Zeng, Yiqun Li, Zhouping Yin, Erli Meng, Ziyu Wei, Leyan Zhu, and Zitian Wang. MSAttnFlow: Normalizing flow for unsupervised anomaly detection with multi-scale attention. *Pattern Recognition*, 161:111220, 2025.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. In *Proceedings of the 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 66–71, 2021.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *International Conference on Learning Representations*, 2025.

Mingyu Lee and Jongwon Choi. Text-guided variational image generation for industrial anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26519–26528, 2024.

Yun Liang, Zhiguang Hu, Junjie Huang, Donglin Di, Anyang Su, and Lei Fan. ToCoAD: Two-stage contrastive learning for industrial anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, 74:1–9, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017.

Chuni Liu, Hongjie Li, Jiaqi Du, Yangyang Hou, Qian Sun, Lei Jin, and Ke Xu. Advancing metallic surface defect detection via anomaly-guided pretraining on a large industrial dataset. *Pattern Recognition*, 179:113788, 2026.

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*, pp. 4744–4754, 2025.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

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 *Proceedings of the IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 1–6, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Wahyu Rahmaniar and Kenji Suzuki. Multi-AD: Cross-domain unsupervised anomaly detection for medical and industrial applications. *Pattern Recognition*, 172:112486, 2026.

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.

Shunsuke Sakai, Xiangteng He, Chunzhi Gu, Leonid Sigal, and Tatsuhito Hasegawa. InvAD: Inversion-based reconstruction-free anomaly detection with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21389–21398, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, 2004.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. 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.

Chengjie Wang, Xi Jiang, Bin-Bin Gao, Zhenye Gan, Yong Liu, Feng Zheng, and Lizhuang Ma. SoftPatch+: Fully unsupervised anomaly classification and segmentation. *Pattern Recognition*, 161:111295, 2025a.

Fuyun Wang, Tong Zhang, Yuanzhi Wang, Yide Qiu, Xin Liu, Xu Guo, and Zhen Cui. Distribution prototype diffusion learning for open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20416–20426, 2025b.

Minghui Yang, Jing Liu, Zhiwei Yang, and Zhaoyang Wu. SLSG: Industrial image anomaly detection with improved feature embeddings and one-class classification. *Pattern Recognition*, 156:110862, 2024.

Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. ResAD: A simple framework for class generalizable anomaly detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125287–125311, 2024.

Xincheng Yao, Yan Luo, Zefeng Qian, and Chongyang Zhang. ADPretrain: Advancing industrial anomaly detection via anomaly representation pretraining. In *Advances in Neural Information Processing Systems*, volume 38, pp. 136372–136400, 2025.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4571–4584, 2022.

Qien Yu, Shengxin Dai, Ran Dong, and Soichiro Ikuno. Attention-based vector quantized variational autoencoder for anomaly detection by using orthogonal subspace constraints. *Pattern Recognition*, 164:111500, 2025.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16699–16708, 2024.

Jiayin Zhou, Wai Keung Wong, and Fangjian Liao. One-shot unsupervised industrial anomaly detection: Enhanced performance under extreme data scarcity. *Pattern Recognition*, 173:112759, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 2024.

Yixuan Zhou, Xing Xu, Zhe Sun, Jingkuan Song, Andrzej Cichocki, and Heng Tao Shen. VQ-Flow: Taming normalizing flows for multi-class anomaly detection via hierarchical vector quantization. *IEEE Transactions on Multimedia*, 27:6884–6895, 2025.

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. 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.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pp. 392–408. Springer, 2022.