# SUPERVISED ANOMALY DETECTION VIA EVIDENCE-AWARE RECONSTRUCTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reconstruction-based anomaly detection methods have achieved strong performance by exploiting the discrepancy between reconstructed and observed patterns. However, reconstruction discrepancies are inherently ambiguous: large residuals may arise from difficult normal content, while anomalous regions can remain undetected when they are well reconstructed. This ambiguity leads to overlapping score distributions and limits reliable anomaly decisions under practical operating conditions. In this work, we propose Evidence-aware Reconstruction for Anomaly Detection (EviAD), a framework that uses a small set of annotated anomalies to augment reconstruction-based detectors with uncertainty-aware anomaly scoring. EviAD models two complementary forms of uncertainty: reconstruction uncertainty characterizes how difficult normal content is to reconstruct, helping reduce false alarms caused by large normal residuals; evidential uncertainty characterizes how strongly the learned evidence supports an anomaly prediction, helping identify anomalies that reconstruction errors alone may miss. Together, they guide anomaly scoring to better distinguish difficult normal samples from well-reconstructed anomalies. Extensive experiments demonstrate that EviAD consistently improves anomaly detection and localization performance, achieving higher recall under strict false-positive constraints and supporting more reliable industrial anomaly inspection.

## 1 INTRODUCTION

Visual anomaly detection (AD) supports industrial quality assurance by identifying defects in manufactured products. Among existing paradigms, reconstruction-based methods have emerged as a prominent approach (Liu et al., 2024). These methods learn to reconstruct images or features from defect-free data and use reconstruction discrepancies as anomaly scores, assuming that normal patterns are reconstructed more faithfully than anomalies. With increasingly powerful pretrained representations (Guo et al., 2025b) and improved reconstruction architectures (Deng & Li, 2022; He et al., 2024), recent methods have achieved strong detection performance on established industrial benchmarks (Bergmann et al., 2019; Wang et al., 2024).

Despite this progress, performance is approaching saturation with increasingly small differences in commonly reported metrics (*e.g.,* image-level AUROC). In practice, anomaly scores must be converted into accept-or-reject decisions using a decision threshold. When missed defects are costly, these decisions must achieve high defect recall while limiting false alarms (Roth et al., 2022). Strong benchmark performance does not necessarily ensure this trade-off. Figure 1a illustrates this gap: despite an image-level AUROC of 98.13%, normal and anomalous samples occupy overlapping reconstruction-score regions, making reliable separation at an operating threshold challenging.

We argue that this gap originates from the intrinsic ambiguity of reconstruction discrepancies. Their magnitude reflects not only abnormality but also the difficulty of reconstructing normal content. Complex textures, boundaries, and fine details can produce large residuals even in defect-free regions, leading to false alarms on hard-normal regions (Baitieva et al., 2025). Conversely, anomalous regions may be reconstructed well and receive low anomaly scores (You et al., 2022). Reconstruction discrepancies therefore conflate normal reconstruction difficulty with abnormality, contributing to the overlap between normal and anomalous score distributions.



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.

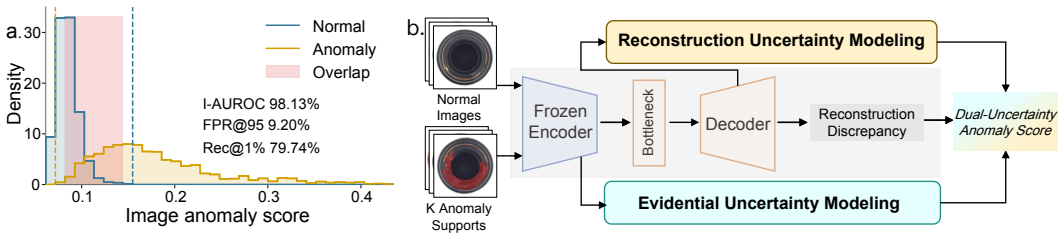



Figure 1: **a.** Image-level anomaly score distributions on the Real-IAD *switch* category. **b.** EviAD models reconstruction uncertainty and evidential uncertainty based on a reconstruction framework.

To address this ambiguity, we propose Evidence-aware reconstruction for Anomaly Detection (EviAD), a framework that augments feature-reconstruction detectors (Deng & Li, 2022; Guo et al., 2025b) through uncertainty-aware scoring (Figure 1b). EviAD models two complementary forms of uncertainty: *reconstruction uncertainty*, which characterizes the expected reconstruction-error scale of normal content, and *evidential uncertainty*, which reflects the sufficiency of evidence for distinguishing normal from anomalous patterns. Together, they inform anomaly scoring to better distinguish difficult normal samples from well-reconstructed anomalies. We adopt a supervised AD setting (Ding et al., 2022; Yao et al., 2023; Han et al., 2026) with a small set of anomalous supports and synthetic anomalies, while training the reconstruction pathway exclusively on normal data.

Specifically, EviAD predicts local reconstruction-error scales from decoder features using lightweight heads trained on observed normal reconstruction errors through a heteroscedastic objective. These estimates guide reconstruction refinement toward difficult normal regions and support difficulty-aware score correction. For evidential uncertainty, EviAD compares projected multi-level encoder features with learned reference subspaces for easy-normal, hard-normal, and anomalous patterns. These matches determine Dirichlet evidence, yielding anomaly belief and evidential uncertainty: strong matches with clear normal–anomaly separation provide more evidence, whereas weak or ambiguous matches indicate higher uncertainty. At inference, EviAD combines anomaly belief with reconstruction responses and uses the predicted reconstruction difficulty to suppress excessive responses when strong anomaly evidence is absent.

Our contributions are threefold: *1)* We propose EviAD, a framework that uses reconstruction and evidential uncertainty to guide anomaly-score correction under limited anomaly supervision, addressing hard-normal false alarms and well-reconstructed anomalies. *2)* We develop decoder-based reconstruction-error scale estimation and reference-based Dirichlet evidence estimation, supporting hard-normal refinement and uncertainty-aware anomaly scoring. *3)* Experiments on MVTec AD (Bergmann et al., 2019), Real-IAD (Wang et al., 2024), and MANTA (Fan et al., 2025) demonstrate improved detection and localization, with consistently reduced false-positive rates at high recall and improved recall under strict false-positive constraints.

## 2 Related Work

**Unsupervised Anomaly Detection.** These methods model normal appearance using defect-free images. For example, memory-based methods such as PatchCore and RAD detect anomalies by matching test features against stored normal features (Roth et al., 2022; Zhang et al., 2026). Reconstruction-based methods learn to reconstruct normal images or features and use reconstruction discrepancies as anomaly scores. Recent approaches advance this paradigm through contrastive feature learning (Guo et al., 2023), state-space modeling (He et al., 2024), and pretrained Transformers (Guo et al., 2025b).

**Supervised Anomaly Detection.** Supervised AD exploits labeled anomalies to provide defect-aware guidance (Baitieva et al., 2024; Yao et al., 2023). Under limited anomaly supervision, a central challenge is to learn from observed defects while maintaining sensitivity to unseen anomalies (Ding et al., 2022; Zhu et al., 2024). Recent methods address this challenge through normal–anomaly distribution modeling and boundary calibration. DPDL (Wang et al., 2025) learns Gaussian prototypes for normal features and uses a Schrödinger bridge to guide normal samples toward these prototypes while pushing anomalies away. ArcAD (Han et al., 2026) calibrates reconstruction-based detectors under cold-start conditions by forming compact normal clusters on a hypersphere and using real

and synthetic anomalies to refine the normal boundary. Within the supervised setting, our work models normal reconstruction difficulty and the reliability of supervised anomaly evidence through two complementary forms of uncertainty to guide anomaly scoring.

# 3 EVIDENCE-AWARE RECONSTRUCTION

## 3.1 OVERVIEW

We consider a supervised $K$-anomaly support setting, where training uses a normal dataset $\mathcal{X}_N = \{x_i^N\}_{i=1}^N$ and a small set of $K$ anomalous images $\mathcal{X}_A^K = \{x_j^A\}_{j=1}^K$, each accompanied by a pixel-level anomaly mask. These annotated images serve as anomaly supports, providing defect-specific supervision. We build on a feature-reconstruction architecture consisting of a frozen encoder $E$, a trainable bottleneck $B$, and a trainable decoder $D$. Given a normal training image $x_n \in \mathcal{X}_N$, the encoder extracts features from $L$ selected layers, which are reconstructed by the bottleneck and decoder. We denote the encoder and reconstructed features at layer $l$ as $f_l^e$ and $f_l^d$, respectively, with $\{f_l^e\}_{l=1}^L = E(x_n)$ and $\{f_l^d\}_{l=1}^L = D(B(E(x_n)))$. Let $\mathcal{P}$ denote the shared spatial patch grid. For each spatial patch $p \in \mathcal{P}$, the layer-wise reconstruction error is defined as $r_l(p) = 1 - \cos(f_l^e(p), f_l^d(p))$. The bottleneck and decoder are initially optimized on normal images using $\mathcal{L}_{\text{rec}} = \frac{1}{L|\mathcal{P}|}\sum_{l=1}^L \sum_{p\in\mathcal{P}} r_l(p)$.

Building on this reconstruction pathway, EviAD models two complementary forms of uncertainty, as shown in Figure 2. Reconstruction uncertainty learns from normal data to predict the expected reconstruction-error scale, guiding reconstruction refinement toward difficult normal regions and supporting difficulty-aware score correction. Evidential uncertainty learns normal and anomalous reference patterns from encoder features using normal images, annotated anomaly supports, and synthetic anomalies, yielding anomaly belief and evidential uncertainty. At inference, evidential uncertainty modulates the strength of anomaly belief, while reconstruction uncertainty supports an additional correction for difficult normal content.

## 3.2 RECONSTRUCTION UNCERTAINTY MODELING (RUM)

Normal regions can exhibit different reconstruction difficulties, making the same reconstruction error carry different implications for anomaly detection. We estimate the expected reconstruction-error scale of normal content and use it as reconstruction uncertainty to guide hard-normal refinement and difficulty-aware anomaly scoring.

**Normal reconstruction-error scale estimation.** After the initial reconstruction training, we keep the encoder $E$ frozen and freeze the bottleneck $B$ and decoder $D$. We attach a lightweight scale head $H_l$ to each selected decoder layer. Given the reconstructed feature $f_l^d(p)$ at patch $p$, the head predicts a positive error scale $\sigma_l^2(p) = \exp\big(H_l(f_l^d(p))\big)$. The scale heads are trained exclusively on normal images using the observed reconstruction errors $r_l(p)$ through a heteroscedastic objective:

$$\mathcal{L}_{\text{diff}} = \frac{1}{2L|\mathcal{P}|}\sum_{l=1}^{L}\sum_{p\in\mathcal{P}}\left[\frac{r_l(p)}{\sigma_l^2(p)} + \log\sigma_l^2(p)\right]. \tag{1}$$

With the reconstruction pathway fixed, the heads learn how the expected error scale varies with normal content. We aggregate the layer-wise predictions into patch-level reconstruction uncertainty:

$$u_{\text{rec}}(p) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\sigma_l^2(p)}. \tag{2}$$

Larger values indicate greater predicted reconstruction difficulty.

**Difficulty-guided hard-normal refinement.** We next use the predicted scales to emphasize difficult normal patches during reconstruction refinement. For each normal image and decoder layer, patches in the top-$q$ fraction of predicted error scales receive additional weight:

$$w_l(p) = 1 + \gamma\,\mathbb{1}\left[\sigma_l^2(p) \geq Q_{1-q}\big(\sigma_l^2\big)\right], \tag{3}$$

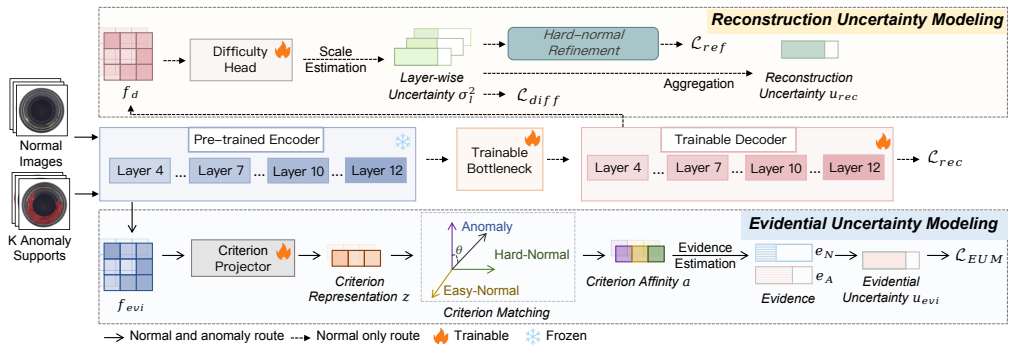



Figure 2: **EviAD Overview.** Based on a reconstruction framework, EviAD disentangles anomaly detection uncertainty into two complementary aspects: Reconstruction Uncertainty Modeling estimates normal reconstruction difficulty, and Evidential Uncertainty Modeling learns anomaly evidence with uncertainty awareness.

where $\mathbb{1}[\cdot]$ is the indicator function, $q \in (0,1)$ is the selected fraction, and $Q_{1-q}(\sigma_l^2)$ denotes the $(1-q)$-quantile of predicted scales within the image at layer $l$. The parameter $\gamma > 0$ controls the additional weight assigned to difficult patches. Keeping the encoder and scale heads fixed, we refine the bottleneck and decoder on normal images using the difficulty-weighted reconstruction objective:

$$\mathcal{L}_{\text{ref}} = \frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{p\in\mathcal{P}} w_l(p) r_l(p)}{\sum_{p\in\mathcal{P}} w_l(p)}. \tag{4}$$

This refinement emphasizes reducing reconstruction errors in difficult normal regions. At inference, the learned scale heads provide reconstruction uncertainty for difficulty-aware score correction.

### 3.3 Evidential Uncertainty Modeling (EUM)

Although reconstruction uncertainty accounts for difficult normal content, anomalies with small reconstruction errors may still be missed. We learn normal and anomalous reference patterns from the available supervision and use feature-to-reference matching to estimate anomaly belief.

**Criterion representation and matching.** We construct a shared criterion space using normal images and a limited set of mask-annotated anomaly supports. The masks identify anomalous patches, while mask-aware defect-paste augmentation (Li et al., 2021) provides additional synthetic anomalies for training. For each image $x$, we aggregate features from the selected encoder layers as $f_{\text{evi}}(x,p) = \frac{1}{L}\sum_{l=1}^{L} f_l^e(x,p)$ and project them into the criterion space: $z(x,p) = P_\phi(f_{\text{evi}}(x,p))$, where $P_\phi$ is a learnable projection head.

To represent normal patterns with different reconstruction difficulties, we learn separate reference subspaces $\mathcal{B}_{\text{easy}}$ and $\mathcal{B}_{\text{hard}}$, together with an anomaly reference subspace $\mathcal{B}_A$. Specifically, normal patches are divided into easy and hard groups according to their reconstruction residuals, while mask-covered defect patches constitute the anomaly criterion set. Each subspace is initialized using the leading SVD directions of its projected criterion features. Each reference subspace $\mathcal{B}_c$ is represented by a basis matrix $B_c$, whose rows correspond to its basis vectors. The basis vectors are subsequently optimized as trainable parameters and orthonormalized through QR decomposition before matching (see Appendix A.3.1). For the normalized feature $\hat{z}(x,p) = z(x,p)/(\|z(x,p)\|_2 + \epsilon)$, its criterion affinity for $c \in \mathcal{C} = \{\text{easy}, \text{hard}, A\}$ is defined as

$$a_c(x,p) = \max_j \left|\hat{z}(x,p)^\top B_{c,j}\right|, \tag{5}$$

where $B_{c,j}$ denotes the $j$-th basis vector of $\mathcal{B}_c$. This affinity measures the strongest match to the reference directions. We combine the two normal affinities as $a_N(x,p) = \max(a_{\text{easy}}(x,p), a_{\text{hard}}(x,p))$, obtaining normal and anomaly affinities for evidence estimation.

**Criterion-guided evidence estimation.** The relative affinities indicate whether a patch is more consistent with normal or anomalous patterns. We convert them into evidence allocation weights:

$$\boldsymbol{\pi}(x,p) = \mathrm{softmax}\left(\frac{[a_N(x,p), a_A(x,p)]}{\tau}\right), \tag{6}$$

where $\tau > 0$ is a temperature parameter. To determine the amount of supporting evidence, we measure reference coverage as $c(x,p) = \max(a_N(x,p), a_A(x,p))$ and normal–anomaly separation as $m(x,p) = |a_N(x,p) - a_A(x,p)|$. Strong matches with clear separation provide more evidence, whereas weak or ambiguous matches indicate higher uncertainty. The evidence strength and class-wise evidence are defined as

$$\begin{aligned} S(x,p) &= S_{\max} g_c(c(x,p)) g_m(m(x,p)), \\ [e_N(x,p), e_A(x,p)] &= S(x,p)\boldsymbol{\pi}(x,p), \end{aligned} \tag{7}$$

where $S_{\max}$ controls the evidence scale, and $g_c$ and $g_m$ are sigmoid mappings of coverage and separation, respectively.

Following evidential deep learning (Sensoy et al.), we parameterize a Dirichlet distribution with $\alpha_N = e_N + 1$ and $\alpha_A = e_A + 1$. Since $S = e_N + e_A$, anomaly belief and evidential uncertainty are

$$b_A(x,p) = \frac{e_A(x,p)}{S(x,p)+2}, \qquad u_{\mathrm{evi}}(x,p) = \frac{2}{S(x,p)+2}. \tag{8}$$

**Training objective.** We jointly learn the projection head and reference subspaces using criterion and evidential objectives. The criterion objective encourages matching to the assigned reference, discourages overlap between reference subspaces, and separates competing affinities:

$$\mathcal{L}_{\mathrm{cri}} = \sum_{c\in\mathcal{C}} \mathbb{E}_{p\in\Omega_c}\left[1 - a_c(p)\right] + \sum_{c<c'} \left\| B_c B_{c'}^{\top} \right\|_F^2 + \sum_{c\in\mathcal{C}} \mathbb{E}_{p\in\Omega_c}\left[\delta + \max_{c'\neq c} a_{c'}(p) - a_c(p)\right]_+, \tag{9}$$

where $\Omega_c$ denotes the patches assigned to reference group $c$, $\delta$ is the affinity margin, and $[v]_+ = \max(v, 0)$. The image argument is omitted for brevity.

For evidence learning, let $S_\alpha(p) = \alpha_N(p) + \alpha_A(p)$ and $\hat{y}_A(p) = \alpha_A(p)/S_\alpha(p)$. We optimize

$$\mathcal{L}_{\mathrm{evi}} = \ell_{\mathrm{sup}}(\hat{y}_A, y_A) + \lambda_{\mathrm{EDL}} \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \sum_{k\in\{N,A\}} y_k(p)\left[\psi(S_\alpha(p)) - \psi(\alpha_k(p))\right] + \eta_t \mathcal{R}_{\mathrm{KL}}, \tag{10}$$

where $y_k(p)$ is the class target of patch $p$, and $\psi$ is the digamma function. The digamma-based evidential term encourages evidence for the target class, with its contribution controlled by $\lambda_{\mathrm{EDL}}$. The supervised term $\ell_{\mathrm{sup}}$ is the sum of Dice loss and focal loss applied to the expected anomaly probability $\hat{y}_A$. The KL regularizer $\mathcal{R}_{\mathrm{KL}}$ suppresses unsupported non-target evidence by regularizing the adjusted Dirichlet distribution toward a uniform Dirichlet prior. Its coefficient $\eta_t$ is linearly annealed during the initial training iterations (see Appendix A.3.2). The overall objective for anomaly evidence learning is:

$$\mathcal{L}_{\mathrm{EUM}} = \mathcal{L}_{\mathrm{evi}} + \beta\, \mathcal{L}_{\mathrm{cri}}, \tag{11}$$

where $\beta$ controls the contribution of the criterion objective. The encoder and reconstruction pathway remain frozen during this stage. The learned anomaly belief is subsequently combined with reconstruction responses, as described in Sec. 3.4.

### 3.4 Training and Uncertainty-Guided Inference

**Staged optimization.** The optimization is performed in two stages. In the first stage, the bottleneck and decoder are optimized using $\mathcal{L}_{\mathrm{rec}}$ on normal images. In the second stage, RUM and EUM are optimized concurrently using separate optimizers and disjoint gradient paths. Within RUM, the scale heads are trained using $\mathcal{L}_{\mathrm{diff}}$ with the reconstruction pathway fixed, while the predicted scales guide reconstruction refinement through $\mathcal{L}_{\mathrm{ref}}$, which updates the bottleneck and decoder while keeping the encoder and scale heads fixed. Meanwhile, EUM is trained using $\mathcal{L}_{\mathrm{EUM}}$ on normal images, the $K$ annotated anomaly supports, and synthetic anomalies. Its objective updates only the projection head and reference subspaces. Anomaly supervision is confined to EUM and does not enter the reconstruction objectives.

**Dual-uncertainty inference.** At inference, evidential uncertainty is reflected in anomaly belief through $b_A(x,p) = (1 - u_{\text{evi}}(x,p))\pi_A(x,p)$, reducing belief when the supporting evidence is insufficient. Reconstruction uncertainty guides difficulty-aware correction of reconstruction responses. We standardize each patch-level score component using normal training statistics, denoted by $z_N(v) = (v - \mu_v)/\sigma_v$, where $\mu_v$ and $\sigma_v$ are computed from the top-1% pooled scores of normal training images and fixed during testing.

We first derive an image-level gate from anomaly belief:

$$g_N(x) = \mathbb{1}\left[B_A(x) \leq Q_{q_N}\left(B_A(\mathcal{X}_N)\right)\right], \tag{12}$$

where $B_A(x)$ is obtained by averaging the highest-scoring $1\%$ of patches in the anomaly-belief map $b_A(x,p)$, and $B_A(\mathcal{X}_N)$ denotes the collection of these image-level scores over normal training images. The hyperparameter $q_N \in (0,1)$ specifies the quantile level used to set the gate threshold.

The gate enables difficulty-aware correction when the image-level anomaly belief does not exceed the normal reference threshold. The corrected reconstruction response is

$$s_{\text{RUM}}(x,p) = \lambda_R z_N\left(r_{\text{RUM}}(x,p)\right) - \lambda_D g_N(x)\left[z_N\left(u_{\text{rec}}(x,p)\right)\right]_+, \tag{13}$$

where $r_{\text{RUM}}(x,p)$ is obtained by averaging the group-wise cosine reconstruction errors from the RUM-refined reconstruction, $u_{\text{rec}}(x,p)$ is the predicted reconstruction uncertainty, and $\lambda_R$ and $\lambda_D$ control the reconstruction contribution and difficulty-correction strength, respectively.

The final patch-level anomaly score combines calibrated anomaly belief with the corrected reconstruction response:

$$s_{\text{EviAD}}(x,p) = z_N\left(b_A(x,p)\right) + s_{\text{RUM}}(x,p). \tag{14}$$

Finally, we bilinearly upsample the patch-level score map to the input image resolution and compute the image-level anomaly score by averaging its top-$1\%$ responses.

# 4 Experiments

## 4.1 Experimental Setup

**Dataset.** We conduct experiments on three industrial AD datasets. MVTec AD (Bergmann et al., 2019) contains 15 object and texture categories, with 3,629 normal training images and 1,725 test images. Real-IAD (Wang et al., 2024) contains 30 industrial categories, including 36,465 normal training images and 114,585 test images. MANTA (Fan et al., 2025) contains 38 object categories. We use its Tiny version, which provides 107,415 training and 44,585 test images.

**K-anomaly support setting.** We follow the single-class evaluation protocol for all experiments. For the supervised methods, including EviAD, we randomly select $K = 10$ anomalous images per category from the original test set as anomaly supports. These samples are removed from the test set and used only for training. To keep the total number of original training images unchanged, we remove $K$ normal samples per category from the normal training set. Unsupervised methods use the full normal training set without anomaly support and are evaluated on the same reduced test set.

**Evaluation Metrics.** Following previous works (Guo et al., 2025b; He et al., 2024), we evaluate image-level performance using the Area Under the Receiver Operating Characteristic Curve (I-AUROC) and pixel-level localization using Pixel-level AUROC (P-AUROC), Average Precision (AP), and maximum $F_1$ score ($F_1$-max). To evaluate image-level operating-point performance, we report False Positive Rate at 90%/95% True Positive Rate (FPR@90/FPR@95) and Recall at 1%/5% False Positive Rate (Rec@1/Rec@5), which measure false alarms under high recall and recall under strict false-positive constraints, respectively.

## 4.2 Comparison with State-of-the-art Methods

We compare EviAD with both unsupervised (Deng & Li, 2022; Guo et al., 2023; Liu et al., 2023; Guo et al., 2025b; Wei et al., 2025; Lendering et al., 2026) and supervised anomaly detection (Ding et al., 2022; Yao et al., 2023; Han et al., 2026) methods. To evaluate robustness to anomaly support selections, we repeat the experiments with three random seeds, each selecting a different set of $K$ anomaly supports per category. The results are reported in Table 1.

Table 1: **Comparisons with Unsupervised and Supervised Baselines.** Image-level results are reported using I-AUROC↑, FPR@TPR90↓, FPR@TPR95↓, Recall@FPR1↑, and Recall@FPR5↑. Pixel-level results are reported using P-AUROC↑, P-AP↑, and P-$F_1$-max↑. All results are macro-averaged over categories. Small values at the lower right denote standard deviations over 3 seeds. The best performance is highlighted in bold.

| Setting | Method | Image-level | | | | | Pixel-level | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUROC↑ | FPR@90↓ | FPR@95↓ | Rec@1↑ | Rec@5↑ | AUROC↑ | AP↑ | $F_1$-max↑ |
| **MVTec AD** | | | | | | | | | |
| *Unsupervised* | Dinomaly | $99.7_{\pm 0.0}$ | $0.6_{\pm 0.0}$ | $1.8_{\pm 0.2}$ | $96.1_{\pm 0.1}$ | $98.9_{\pm 0.1}$ | $98.5_{\pm 0.0}$ | $65.3_{\pm 0.0}$ | $66.7_{\pm 0.1}$ |
| | RD4AD | $98.8_{\pm 0.0}$ | $1.6_{\pm 0.1}$ | $6.8_{\pm 0.3}$ | $91.0_{\pm 0.3}$ | $93.8_{\pm 0.2}$ | $97.7_{\pm 0.0}$ | $57.2_{\pm 0.1}$ | $60.4_{\pm 0.1}$ |
| | ReContrast | $99.2_{\pm 0.0}$ | $1.2_{\pm 0.1}$ | $6.7_{\pm 1.0}$ | $93.6_{\pm 0.1}$ | $96.2_{\pm 0.1}$ | $98.0_{\pm 0.0}$ | $57.8_{\pm 0.2}$ | $60.8_{\pm 0.1}$ |
| | UniNet | $98.3_{\pm 0.1}$ | $4.1_{\pm 0.1}$ | $8.2_{\pm 0.3}$ | $86.3_{\pm 0.2}$ | $90.9_{\pm 0.1}$ | $98.2_{\pm 0.0}$ | $61.2_{\pm 0.1}$ | $62.0_{\pm 0.1}$ |
| | SimpleNet | $95.0_{\pm 0.1}$ | $13.2_{\pm 0.2}$ | $17.7_{\pm 0.6}$ | $78.4_{\pm 0.3}$ | $86.7_{\pm 0.1}$ | $95.9_{\pm 0.1}$ | $51.0_{\pm 0.3}$ | $53.1_{\pm 0.4}$ |
| | SubspaceAD | $98.3_{\pm 0.0}$ | $5.9_{\pm 0.4}$ | $9.5_{\pm 0.5}$ | $88.6_{\pm 0.2}$ | $92.8_{\pm 0.2}$ | $97.9_{\pm 0.0}$ | $58.9_{\pm 0.3}$ | $61.5_{\pm 0.3}$ |
| *Supervised* | DRA | $96.5_{\pm 0.5}$ | $9.0_{\pm 2.4}$ | $17.7_{\pm 4.0}$ | $86.8_{\pm 0.9}$ | $90.0_{\pm 0.9}$ | $82.5_{\pm 2.1}$ | $24.3_{\pm 0.6}$ | $30.9_{\pm 1.0}$ |
| | BGAD | $80.0_{\pm 1.8}$ | $44.3_{\pm 3.3}$ | $52.9_{\pm 4.4}$ | $52.1_{\pm 1.8}$ | $55.2_{\pm 2.0}$ | $92.1_{\pm 0.3}$ | $43.6_{\pm 1.6}$ | $44.7_{\pm 1.4}$ |
| | ArcAD | $\mathbf{99.9}_{\pm 0.1}$ | $0.4_{\pm 0.1}$ | $0.6_{\pm 0.0}$ | $97.7_{\pm 0.0}$ | $99.3_{\pm 0.1}$ | $99.0_{\pm 0.0}$ | $66.8_{\pm 0.0}$ | $67.7_{\pm 0.2}$ |
| | **EviAD (Ours)** | $\mathbf{99.9}_{\pm 0.1}$ | $\mathbf{0.3}_{\pm 0.3}$ | $\mathbf{0.5}_{\pm 0.2}$ | $\mathbf{98.6}_{\pm 0.2}$ | $\mathbf{99.4}_{\pm 0.5}$ | $\mathbf{99.1}_{\pm 0.1}$ | $\mathbf{70.7}_{\pm 1.2}$ | $\mathbf{73.8}_{\pm 1.1}$ |
| **Real-IAD** | | | | | | | | | |
| *Unsupervised* | Dinomaly | $91.5_{\pm 0.0}$ | $24.1_{\pm 0.1}$ | $34.6_{\pm 0.1}$ | $45.9_{\pm 0.1}$ | $64.4_{\pm 0.1}$ | $99.1_{\pm 0.0}$ | $40.4_{\pm 0.1}$ | $45.4_{\pm 0.0}$ |
| | RD4AD | $89.7_{\pm 0.0}$ | $30.3_{\pm 0.1}$ | $43.0_{\pm 0.1}$ | $43.7_{\pm 0.1}$ | $61.6_{\pm 0.1}$ | $99.0_{\pm 0.0}$ | $34.2_{\pm 0.1}$ | $41.4_{\pm 0.0}$ |
| | ReContrast | $90.7_{\pm 0.0}$ | $27.1_{\pm 0.1}$ | $38.4_{\pm 0.1}$ | $44.5_{\pm 0.1}$ | $62.4_{\pm 0.1}$ | $99.1_{\pm 0.0}$ | $34.6_{\pm 0.1}$ | $41.9_{\pm 0.1}$ |
| | UniNet | $87.2_{\pm 0.0}$ | $32.9_{\pm 0.1}$ | $44.6_{\pm 0.1}$ | $28.7_{\pm 0.1}$ | $46.4_{\pm 0.0}$ | $98.1_{\pm 0.0}$ | $26.7_{\pm 0.1}$ | $35.0_{\pm 0.1}$ |
| | SimpleNet | $88.4_{\pm 0.0}$ | $34.7_{\pm 0.1}$ | $48.3_{\pm 0.1}$ | $41.0_{\pm 0.0}$ | $60.5_{\pm 0.1}$ | $96.7_{\pm 0.0}$ | $21.5_{\pm 0.1}$ | $29.0_{\pm 0.1}$ |
| | SubspaceAD | $86.6_{\pm 0.0}$ | $36.8_{\pm 0.1}$ | $49.3_{\pm 0.1}$ | $31.3_{\pm 0.0}$ | $51.8_{\pm 0.1}$ | $98.9_{\pm 0.0}$ | $37.8_{\pm 0.1}$ | $43.0_{\pm 0.1}$ |
| *Supervised* | DRA | $81.5_{\pm 0.0}$ | $49.8_{\pm 0.7}$ | $64.2_{\pm 0.7}$ | $24.6_{\pm 0.2}$ | $39.9_{\pm 0.4}$ | $81.6_{\pm 0.2}$ | $4.3_{\pm 0.4}$ | $9.9_{\pm 0.5}$ |
| | BGAD | $51.2_{\pm 1.7}$ | $87.9_{\pm 1.5}$ | $93.7_{\pm 0.7}$ | $1.8_{\pm 0.4}$ | $7.3_{\pm 1.4}$ | $86.5_{\pm 1.3}$ | $1.5_{\pm 0.4}$ | $3.7_{\pm 0.8}$ |
| | ArcAD | $94.2_{\pm 0.2}$ | $17.7_{\pm 0.1}$ | $28.1_{\pm 0.2}$ | $59.6_{\pm 0.1}$ | $75.6_{\pm 0.1}$ | $99.2_{\pm 0.0}$ | $41.8_{\pm 0.2}$ | $46.9_{\pm 0.1}$ |
| | **EviAD (Ours)** | $\mathbf{95.3}_{\pm 0.0}$ | $\mathbf{14.7}_{\pm 0.3}$ | $\mathbf{24.8}_{\pm 0.1}$ | $\mathbf{64.4}_{\pm 0.1}$ | $\mathbf{80.6}_{\pm 0.3}$ | $\mathbf{99.4}_{\pm 0.0}$ | $\mathbf{42.9}_{\pm 0.2}$ | $\mathbf{48.3}_{\pm 0.2}$ |
| **MANTA** | | | | | | | | | |
| *Unsupervised* | Dinomaly | $94.5_{\pm 0.0}$ | $13.3_{\pm 0.1}$ | $19.5_{\pm 0.1}$ | $59.2_{\pm 0.1}$ | $80.7_{\pm 0.1}$ | $95.4_{\pm 0.0}$ | $55.4_{\pm 0.1}$ | $56.6_{\pm 0.1}$ |
| | RD4AD | $91.7_{\pm 0.0}$ | $22.5_{\pm 0.1}$ | $32.5_{\pm 0.2}$ | $48.7_{\pm 0.0}$ | $72.7_{\pm 0.1}$ | $95.4_{\pm 0.0}$ | $43.1_{\pm 0.1}$ | $47.2_{\pm 0.0}$ |
| | ReContrast | $94.3_{\pm 0.0}$ | $15.2_{\pm 0.1}$ | $23.4_{\pm 0.2}$ | $51.7_{\pm 0.0}$ | $76.1_{\pm 0.1}$ | $96.4_{\pm 0.0}$ | $46.9_{\pm 0.1}$ | $50.3_{\pm 0.0}$ |
| | UniNet | $78.9_{\pm 0.1}$ | $55.9_{\pm 0.3}$ | $70.6_{\pm 0.3}$ | $22.9_{\pm 0.1}$ | $40.3_{\pm 0.1}$ | $83.3_{\pm 0.0}$ | $18.1_{\pm 0.0}$ | $24.1_{\pm 0.0}$ |
| | SimpleNet | $89.1_{\pm 0.1}$ | $28.6_{\pm 0.1}$ | $41.8_{\pm 0.5}$ | $50.3_{\pm 0.1}$ | $69.4_{\pm 0.2}$ | $90.1_{\pm 0.0}$ | $28.5_{\pm 0.0}$ | $34.8_{\pm 0.0}$ |
| | SubspaceAD | $91.2_{\pm 0.0}$ | $24.0_{\pm 0.1}$ | $36.2_{\pm 0.1}$ | $37.8_{\pm 0.1}$ | $64.5_{\pm 0.1}$ | $95.0_{\pm 0.0}$ | $45.2_{\pm 0.0}$ | $48.9_{\pm 0.0}$ |
| *Supervised* | DRA | $91.6_{\pm 1.0}$ | $25.3_{\pm 3.5}$ | $40.5_{\pm 4.1}$ | $48.6_{\pm 3.3}$ | $71.7_{\pm 2.6}$ | $70.8_{\pm 1.7}$ | $9.5_{\pm 1.2}$ | $17.0_{\pm 1.4}$ |
| | BGAD | $76.9_{\pm 2.0}$ | $60.1_{\pm 3.9}$ | $73.1_{\pm 3.3}$ | $27.7_{\pm 3.4}$ | $41.0_{\pm 3.9}$ | $86.4_{\pm 1.2}$ | $17.8_{\pm 3.3}$ | $22.9_{\pm 3.0}$ |
| | ArcAD | $96.3_{\pm 0.3}$ | $10.3_{\pm 0.1}$ | $17.6_{\pm 0.0}$ | $67.3_{\pm 0.1}$ | $85.7_{\pm 0.0}$ | $96.7_{\pm 0.2}$ | $58.9_{\pm 0.1}$ | $60.1_{\pm 0.1}$ |
| | **EviAD (Ours)** | $\mathbf{97.1}_{\pm 0.1}$ | $\mathbf{8.2}_{\pm 0.5}$ | $\mathbf{14.7}_{\pm 0.7}$ | $\mathbf{70.0}_{\pm 0.5}$ | $\mathbf{87.5}_{\pm 0.1}$ | $\mathbf{97.8}_{\pm 0.1}$ | $\mathbf{69.5}_{\pm 0.4}$ | $\mathbf{65.0}_{\pm 0.5}$ |

Across the three datasets, EviAD achieves the strongest overall performance among both unsupervised and supervised baselines. At the image level, EviAD achieves the best or tied-best I-AUROC on MVTec AD, Real-IAD, and MANTA, reaching 99.9%, 95.3%, and 97.1%, respectively. Compared with the strongest unsupervised and supervised counterparts, EviAD improves I-AUROC by 0.2% over the best unsupervised method Dinomaly and matches the best supervised performance of ArcAD on MVTec AD, while achieving gains of 3.8%/1.1% on Real-IAD and 2.6%/0.8% on MANTA, respectively. More importantly, the advantages become more pronounced under practical operating-point evaluation. On the 30-category Real-IAD dataset, compared with the strongest supervised baseline ArcAD, EviAD reduces image-level FPR@90 and FPR@95 by 3.0% and 3.3%, respectively, while improving Recall@1 and Recall@5 by 4.8% and 5.0%. These improvements indicate that EviAD provides better defect recall under stricter false-positive constraints by improving score separation around practical decision boundaries, reducing false alarms from hard normal samples while maintaining sensitivity to well-reconstructed anomalies.

EviAD also shows consistent improvements in pixel-level anomaly localization, achieving P-AUROC/P-AP/P-$F_1$-max of 99.1%/70.7%/73.8% on MVTec AD, 99.4%/42.9%/48.3% on Real-IAD, and 97.8%/69.5%/65.0% on MANTA. Compared with the strongest competing results, this corresponds to average improvements of 0.5%, 5.2%, and 4.1% in P-AUROC, P-AP, and P-$F_1$-max,

respectively. In particular, the gains in P-AP and P-$F_1$-max indicate that EviAD improves pixel-level anomaly ranking and localization quality.

The consistent improvements in both image-level and pixel-level metrics, together with the low variance across different random seeds, demonstrate the robustness of EviAD under different anomaly support selections. These results can be attributed to explicitly modeling two complementary sources of reconstruction ambiguity. RUM estimates the expected reconstruction difficulty of normal content, reducing false alarms caused by intrinsically difficult normal regions. In contrast, EUM provides reconstruction-independent anomaly evidence to recover defects with weak reconstruction discrepancies, while evidential uncertainty reduces anomaly belief when supporting evidence is insufficient. Together, the two uncertainty modeling components improve the separation between difficult normal samples and true anomalies, leading to more reliable detection under strict operating constraints and improved pixel-level localization.

### 4.3 Ablation Study

**Dual-uncertainty Ablation.** We first conducted ablation experiments on the Real-IAD dataset to investigate the contribution of the two uncertainty modeling components in EviAD by removing them individually. As shown in Figure 3, compared with the reconstruction baseline, incorporating reconstruction uncertainty modeling improves I-AUROC and Recall@5 by 0.6% and 1.8%, respectively, while reducing FPR@95 by 2.7%. Incorporating evidential uncertainty modeling into the reconstruction baseline (*Recon.+EUM*) achieves larger gains, improving I-AUROC and Recall@5 by 3.7% and 15.8%, respectively, while reducing FPR@95 by 9.2%. These results demonstrate that the two uncertainty modeling components address different sources of ambiguity in reconstruction-based anomaly detection. Reconstruction uncertainty modeling reduces false alarms from hard-normal regions by characterizing their intrinsic reconstruction difficulty, whereas evidential uncertainty modeling provides additional anomaly evidence from limited anomalous supports to identify defects with weak reconstruction discrepancies. Combining both uncertainty modeling components achieves the best performance, improving I-AUROC and Recall@5 over the baseline by 3.8% and 16.2%, respectively, while reducing FPR@95 by 9.8%, demonstrating their complementary effects.

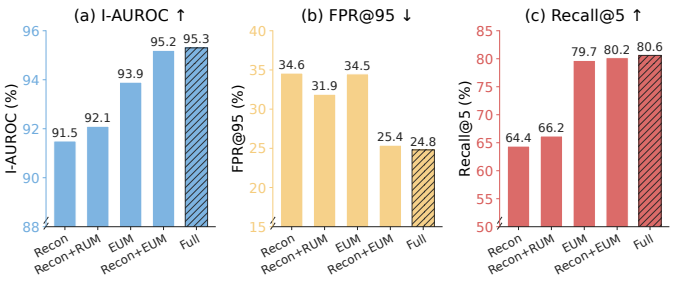



Figure 3: **Main component ablation of EviAD.**

We evaluated whether evidential uncertainty alone is sufficient without reconstruction information. Compared with Recon+EUM, EUM alone reduces I-AUROC and Recall@5 by 1.3% and 0.5%, respectively, while increasing FPR@95 by 9.1%. Compared with full EviAD, EUM alone further reduces I-AUROC and Recall@5 by 1.4% and 0.9%, respectively, while increasing FPR@95 by 9.7%. These results show that reconstruction and anomaly evidence provide complementary information, while RUM further improves detection by reducing false responses from hard-normal regions.

**EUM Ablation.** On the Real-IAD dataset, we further analyze the contribution of each component in EUM, with results reported in Table 2. Removing CutPaste decreases I-AUROC, Rec@1, and Rec@5 by 0.4%, 2.7%, and 0.4%, respectively, indicating that synthetic anomaly augmentation provides additional anomaly variations under limited anomaly supervision. Removing $\mathcal{L}_{\text{cri}}$ leads to drops of 3.0%, 15.8%, and 13.6%, respectively, demonstrating the importance of learning a criterion space that captures discriminative normal and anomalous patterns. Similarly, removing $\mathcal{L}_{\text{evi}}$ decreases the three metrics by 3.0%, 16.0%, and 13.7%, confirming the contribution of evidence estimation to reliable anomaly prediction. Together, these results demonstrate the complementary contributions of the criterion and evidential objectives.

Finally, replacing EUM with a binary classifier and fusing its softmax predictions with reconstruction scores under the same strategy reduces I-AUROC, Rec@1, and Rec@5 by 2.1%, 11.3%, and 7.9%, respectively. This comparison suggests that explicitly modeling criterion compatibility and evidence

Table 2: **Ablation of the EUM.** The best performance is highlighted in bold.

| Group | Image-level | | | | | Pixel-level | | |
|---|---|---|---|---|---|---|---|---|
| | AUROC↑ | FPR@90↓ | FPR@95↓ | Rec@1↑ | Rec@5↑ | AUROC↑ | AP↑ | $F_1$-max↑ |
| EviAD (Binary head) | 93.2 | 20.7 | 31.5 | 53.1 | 72.7 | 99.3 | 34.5 | 44.6 |
| EviAD (w/o $\mathcal{L}_{\text{evi}}$) | 92.3 | 21.8 | 31.9 | 48.4 | 66.9 | 99.3 | 42.4 | 47.7 |
| EviAD (w/o $\mathcal{L}_{\text{cri}}$) | 92.3 | 21.9 | 31.6 | 48.6 | 67.0 | 99.3 | 42.2 | 47.4 |
| EviAD (w/o CutPaste) | 94.9 | 16.5 | 28.0 | 61.7 | 80.2 | 99.2 | 36.4 | 40.0 |
| EviAD | **95.3** | **14.7** | **24.8** | **64.4** | **80.6** | **99.4** | **42.9** | **48.3** |

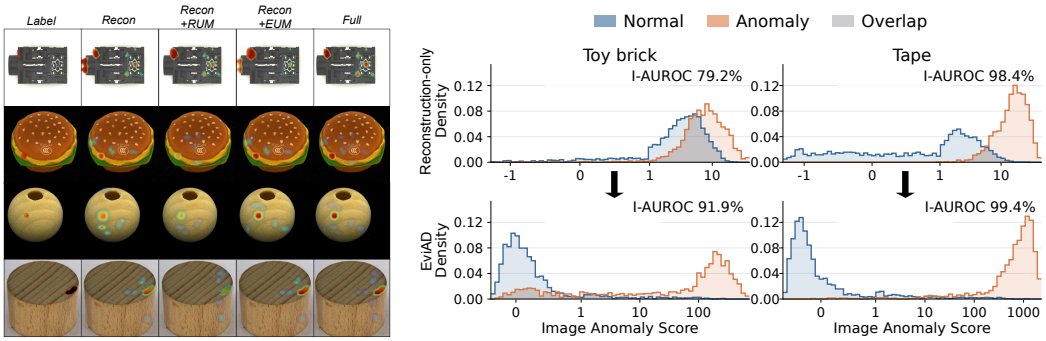



Figure 4: **Left:** Pixel-level anomaly maps produced by different branches of EviAD, including reconstruction baseline, +RUM, +EUM, and full EviAD. **Right:** Image-level anomaly score distributions on the *Toy brick* and *Tape* categories of Real-IAD, comparing the reconstruction baseline with EviAD.

reliability provides additional benefits over directly predicting binary anomaly probabilities, especially under low false-positive operating points.

### 4.4 Visualization

**Dual-Uncertainty Visualization.** To visualize the individual effects of the dual uncertainty modeling, we present pixel-level anomaly maps under different branches of EviAD. As shown in Figure 4.Left, adding RUM suppresses false responses caused by hard-normal regions. For example, in the first row, the incorrectly highlighted region caused by reconstruction errors is reduced after introducing RUM, indicating that reconstruction uncertainty helps distinguish difficult normal patterns from true anomalies. In contrast, introducing EUM enhances the anomaly evidence on defect regions and improves the localization of anomalous areas. These observations demonstrate that RUM and EUM provide complementary functions: RUM reduces false positives from hard-normal content, while EUM strengthens anomaly evidence for defect localization. Combining both uncertainties further achieves more accurate pixel-level anomaly localization.

**Anomaly Score Distribution.** We further visualize the image-level anomaly score distributions in Figure 4 (Right). On *Toy brick*, EviAD achieves higher I-AUROC with clearer separation between normal and anomalous samples. On *Tape*, although the I-AUROC differs from the reconstruction baseline by only about 1%, EviAD assigns lower scores to normal samples and higher scores to anomalies, substantially reducing their distribution overlap. Such improved separation around the decision boundary is particularly beneficial for operating-point performance.

## 5 Conclusion

We proposed EviAD, a framework that enhances reconstruction-based anomaly detection by modeling two complementary uncertainties: reconstruction uncertainty for normal-content difficulty and evidential uncertainty for anomaly evidence reliability. By disentangling these uncertainty sources, EviAD enables uncertainty-aware integration of reconstruction responses and anomaly evidence, improving the separation between difficult normal regions and well-reconstructed anomalies.

### AI USE STATEMENT

Generative AI tools were used to assist with language polishing and LaTeX composing. All AI-assisted text and derivations were manually reviewed and checked for correctness by the authors. The authors take responsibility for the final content of this work, including content produced with the aid of generative AI.

## REFERENCES

Aimira Baitieva, David Hurych, Victor Besnier, and Olivier Bernard. Supervised anomaly detection for complex industrial images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17754–17762. IEEE, 2024.

Aimira Baitieva, Yacine Bouaouni, Alexandre Briot, Dick Ameln, Souhaiel Khalfaoui, and Samet Akcay. Beyond academic benchmarks: Critical analysis and best practices for visual industrial anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4015–4025. IEEE, 2025.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9727–9736. IEEE, 2022.

Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7378–7388. IEEE, 2022.

Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25518–25527. IEEE, 2025.

Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36: 10721–10740, 2023.

Jia Guo, Shuai Lu, Lei Fan, Zelin Li, Donglin Di, Yang Song, Weihang Zhang, Wenbing Zhu, Hong Yan, Fang Chen, et al. One dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection. *arXiv preprint arXiv:2510.17611*, 2025a.

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 Computer Vision and Pattern Recognition Conference*, pp. 20405–20415, 2025b.

Ningning Han, Lei Fan, Jia Guo, Yunkang Cao, Xiu Su, Feng Cao, Donglin Di, and Tonghua Su. Arcad: Anomaly-rectified calibration for cold-start supervised anomaly detection. In *European Conference on Computer Vision*, pp. 460–477. Springer, 2026.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024.

Camile Lendering, Erkut Akdag, and Egor Bondarau. Subspacead: Training-free few-shot anomaly detection via subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28557–28566, 2026.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9659–9669. IEEE, 2021.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1): 104–135, 2024.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20402–20411. IEEE, 2023.

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

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.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty.

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 *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 22883–22892. IEEE, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20416–20426. IEEE, 2025.

Shun Wei, Jielin Jiang, and Xiaolong Xu. Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9994–10003. IEEE, 2025.

Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24490–24499. IEEE, 2023.

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

Xingwu Zhang, Guanxuan Li, Paul Henderson, Gerardo Aragon-Camarasa, and Zijun Long. Is training necessary for anomaly detection? *arXiv e-prints*, pp. arXiv–2601, 2026.

Jiawen Zhu, Choubo Ding, Yu Tian, and Guansong Pang. Anomaly heterogeneity learning for open-set supervised anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17616–17626. IEEE, 2024.

# A APPENDIX

In this appendix, we provide additional experimental analyses, the reconstruction-based preliminary, methodological details, implementation information for EviAD, and additional pixel-level visualizations. The appendix is organized as follows:

- **Additional ablation studies**, including the influence of the number of anomalous supports $K$ and the sensitivity to the inference hyperparameters. (See Sec. A.1)
- **Reconstruction-based preliminary**, describing the detailed reconstruction framework used as our baseline. (See Sec. A.2)
- **Additional method details**, including criterion-subspace construction, and KL regularization and annealing. (See Sec. A.3)
- **Implementation details**, including architectures and training settings, and evaluation protocols. (See Sec. A.4)
- **Additional pixel-level visualizations**, illustrating pixel-level localization performance of EviAD. (See Sec. A.5)

## A.1 ADDITIONAL ABLATION STUDIES

### A.1.1 EFFECT OF THE NUMBER OF ANOMALOUS SUPPORTS

We study the effect of the number of mask-annotated anomalous support images on Real-IAD. For each category, we first sample a fixed set of 10 anomalous images from the official test split and exclude all ten images from evaluation. The settings with $K \in \{1, 5, 10\}$ use nested subsets of these ten images as anomalous supports, while all three settings are evaluated on the same remaining test set. We remove $K$ normal training images per category to keep the total number of original training images unchanged. Table A.1 reports the results. Increasing $K$ from 1 to 10 consistently improves image-level separation, especially in the low-FPR operating region.

Table A.1: Effect of the number of anomalous supports on Real-IAD. All values are percentages.

| | Image-level | | | | | Pixel-level | | |
|---|---|---|---|---|---|---|---|---|
| $K$ | **I-AUROC↑** | **FPR@90↓** | **FPR@95↓** | **Rec@1↑** | **Rec@5↑** | **P-AUROC↑** | **P-AP↑** | **P-F1↑** |
| 1 | 91.8 | 23.1 | 34.4 | 47.0 | 66.1 | 99.2 | 42.0 | 46.8 |
| 5 | 94.5 | 15.5 | 28.7 | 57.9 | 75.8 | 99.4 | 42.7 | 48.2 |
| 10 | **95.3** | **14.7** | **24.8** | **64.4** | **80.6** | **99.4** | **42.9** | **48.3** |

### A.1.2 SENSITIVITY TO INFERENCE HYPERPARAMETERS

We evaluate the sensitivity of EviAD to the reconstruction weight $\lambda_R$, difficulty-correction weight $\lambda_D$, and normal-gate quantile $q_N$. We vary one parameter at a time while fixing the other two to the reported setting $(\lambda_R, \lambda_D, q_N) = (0.1, 0.1, 0.5)$. All standardization statistics and gate thresholds are estimated exclusively from normal training images. As shown in Table A.2, EviAD remains stable within a local neighborhood of the selected setting.

## A.2 PRELIMINARY: RECONSTRUCTION-BASED ANOMALY DETECTION

Following Dinomaly (Guo et al., 2025b;a), our reconstruction backbone consists of the frozen encoder $E$, the trainable bottleneck $B$, and the trainable decoder $D$ introduced in the main paper. We instantiate $E$ using DINOv2-ViT-B/14 with register tokens (Oquab et al., 2023). The encoder contains 12 Transformer blocks with an embedding dimension of 768. We collect the outputs of the third to tenth blocks,

$$\mathcal{M} = \{3, 4, \ldots, 10\}, \qquad L = |\mathcal{M}| = 8, \tag{15}$$

and re-index these selected representations as $\{f_l^e\}_{l=1}^L$, following the notation in the main paper.

Each encoder representation contains a class token, register tokens, and spatial patch tokens. The complete token sequence is retained when passing through the bottleneck and decoder, allowing the

Table A.2: Parameter sensitivity of EviAD on Real-IAD. Each block varies one parameter while fixing the remaining parameters to their reported values.

| Parameter | Value | I-AUROC↑ | FPR@90↓ | FPR@95↓ | Rec@1↑ | Rec@5↑ | P-AUROC↑ | P-AP↑ | P-F1↑ |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_R$ | 0.05 | 95.08 | 15.10 | 25.94 | 64.46 | 80.03 | 99.38 | 42.82 | 48.27 |
| | 0.10 | 95.28 | 14.71 | 24.75 | 64.49 | 80.61 | 99.37 | 42.91 | 48.28 |
| | 0.15 | 95.35 | 14.28 | 24.39 | 64.50 | 80.58 | 99.36 | 42.96 | 48.29 |
| $\lambda_D$ | 0.05 | 95.28 | 14.70 | 24.81 | 64.49 | 80.61 | 99.37 | 42.91 | 48.28 |
| | 0.10 | 95.28 | 14.71 | 24.75 | 64.49 | 80.61 | 99.37 | 42.91 | 48.28 |
| | 0.15 | 95.27 | 14.70 | 24.79 | 64.49 | 80.62 | 99.37 | 42.91 | 48.28 |
| $q_N$ | 0.40 | 95.28 | 14.73 | 24.76 | 64.49 | 80.61 | 99.37 | 42.91 | 48.28 |
| | 0.50 | 95.28 | 14.71 | 24.75 | 64.49 | 80.61 | 99.37 | 42.91 | 48.28 |
| | 0.60 | 95.28 | 14.70 | 24.68 | 64.49 | 80.61 | 99.37 | 42.91 | 48.28 |

global tokens to provide context during reconstruction. The class and register tokens are removed before constructing the spatial reconstruction-error maps.

The selected encoder representations are first averaged across layers and passed through the bottleneck:

$$f^b = B\left(\frac{1}{L}\sum_{l=1}^{L} f_l^e\right). \tag{16}$$

The bottleneck is implemented as a two-layer MLP with GELU activation, an expansion ratio of 4, and a dropout rate of 0.4. It maps the aggregated encoder representation back to the same 768-dimensional token space. The decoder $D$ contains eight Transformer blocks, each composed of a linear attention layer and a feed-forward network. Intermediate decoder outputs are collected and ordered in reverse depth so that they correspond to the selected encoder representations, yielding $\{f_l^d\}_{l=1}^{L}$.

To preserve information at different semantic depths, the eight encoder and decoder representations are divided into two groups, $\mathcal{G}_1 = \{1, 2, 3, 4\}$ and $\mathcal{G}_2 = \{5, 6, 7, 8\}$. Features within each group are averaged separately on the encoder and decoder sides before reconstruction comparison. The first group mainly preserves relatively shallow appearance information, whereas the second captures deeper structural semantics. The resulting spatial features are used to compute the cosine reconstruction errors defined in the main paper. During reconstruction training, $E$ remains frozen and only $B$ and $D$ are optimized on normal images.

## A.3 Additional Method Details

### A.3.1 Criterion-subspace construction

We construct the three reference subspaces from the projected patch features $z(x, p)$ defined in the main paper. For each normal image $x$, we first aggregate its reconstruction residual across the selected layers:

$$\bar{r}(x, p) = \frac{1}{L}\sum_{l=1}^{L} r_l(x, p). \tag{17}$$

Let $\Omega_c$ denote the set of image–patch pairs assigned to criterion $c \in \mathcal{C} = \{\text{easy}, \text{hard}, A\}$. Normal patches in the highest $10\%$ of $\bar{r}(x, p)$ are assigned to $\Omega_{\text{hard}}$, while the remaining normal patches form $\Omega_{\text{easy}}$. For each anomalous support image, we average-pool its pixel-level mask onto the patch grid and assign defect-dominant patches to $\Omega_A$.

For each criterion $c$, we stack its normalized projected features into

$$Z_c = \left[\hat{z}(x, p)^\top\right]_{(x,p)\in\Omega_c} \in \mathbb{R}^{n_c \times d}, \tag{18}$$

where $n_c = |\Omega_c|$ is the number of image–patch pairs assigned to criterion $c$, and $d$ is the dimension of the projected criterion feature. We then compute the singular value decomposition

$$Z_c = U_c \Sigma_c V_c^\top, \qquad B_c^{(0)} = \left(V_c[:, 1 : r_c]\right)^\top, \tag{19}$$

where $r_c$ is the rank of reference subspace $\mathcal{B}_c$. The leading $r_c$ right singular vectors therefore initialize the rows of the basis matrix $B_c$.

After initialization, the basis matrices $\{B_c\}_{c\in\mathcal{C}}$ and the projection head $P_\phi$ are jointly optimized by $\mathcal{L}_{\mathrm{EUM}}$. Before criterion matching, we orthonormalize each basis by applying QR decomposition to its transpose:

$$B_c^\top = Q_c R_c, \qquad B_c \leftarrow Q_c^\top, \qquad B_c B_c^\top = I. \tag{20}$$

Consequently, the basis used for criterion matching has unit-norm and mutually orthogonal rows within each reference subspace.

### A.3.2 KL REGULARIZATION AND ANNEALING

The KL regularizer penalizes unsupported evidence assigned to the non-target class. For each patch $p$, we construct the adjusted Dirichlet parameters as

$$\widetilde{\alpha}_k(p) = y_k(p) + (1 - y_k(p))\,\alpha_k(p), \qquad k \in \{N, A\}. \tag{21}$$

For the target class, $y_k(p) = 1$ sets $\widetilde{\alpha}_k(p) = 1$, thereby removing its evidence from the regularizer. For the non-target class, $y_k(p) = 0$ retains $\widetilde{\alpha}_k(p) = \alpha_k(p)$. The regularizer therefore acts only on evidence that incorrectly supports the non-target class.

Let $\widetilde{\boldsymbol{\alpha}}(p) = [\widetilde{\alpha}_N(p), \widetilde{\alpha}_A(p)]$ and $\widetilde{S}_\alpha(p) = \widetilde{\alpha}_N(p) + \widetilde{\alpha}_A(p)$. We regularize the adjusted distribution toward the uniform Dirichlet prior:

$$\mathcal{R}_{\mathrm{KL}} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} D_{\mathrm{KL}}\left[\mathrm{Dir}\left(\widetilde{\boldsymbol{\alpha}}(p)\right) \,\|\, \mathrm{Dir}(\mathbf{1})\right]. \tag{22}$$

For the binary Dirichlet distribution used in EviAD, this term is computed as

$$\begin{aligned}\mathcal{R}_{\mathrm{KL}} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \Bigg[ &\log\Gamma\Big(\widetilde{S}_\alpha(p)\Big) - \sum_{k\in\{N,A\}} \log\Gamma(\widetilde{\alpha}_k(p)) \\ &+ \sum_{k\in\{N,A\}} (\widetilde{\alpha}_k(p) - 1)\left(\psi(\widetilde{\alpha}_k(p)) - \psi\Big(\widetilde{S}_\alpha(p)\Big)\right)\Bigg].\end{aligned} \tag{23}$$

To avoid suppressing evidence before the model has learned meaningful normal–anomaly predictions, the KL coefficient is linearly annealed:

$$\eta_t = \lambda_{\mathrm{KL}} \min\left(1, \frac{t+1}{T_{\mathrm{KL}}}\right), \tag{24}$$

where $t$ denotes the current EUM training iteration, $\lambda_{\mathrm{KL}} = 0.01$, and $T_{\mathrm{KL}} = 500$. After the first $T_{\mathrm{KL}}$ iterations, the coefficient remains fixed at $\lambda_{\mathrm{KL}}$.

## A.4 IMPLEMENTATION DETAILS

**Reconstruction backbone.** We use DINOv2-ViT-B/14 with registers as the frozen encoder. Its embedding dimension is 768. The reconstruction targets are obtained from eight middle layers and divided into two groups of four. The bottleneck contains one MLP with expansion ratio 4 and dropout rate 0.4. The decoder contains eight Transformer blocks with linear attention, 12 attention heads, and MLP ratio 4. Input images are resized to $448 \times 448$ and center-cropped to $392 \times 392$.

**Reconstruction training.** The encoder $E$ is frozen throughout training, while the bottleneck $B$ and decoder $D$ are optimized using StableAdamW with learning rate $2 \times 10^{-3}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay $10^{-4}$, and a warm-up cosine schedule whose final learning rate is $2 \times 10^{-4}$. We train this base reconstruction pathway for 5,000 iterations using normal images only.

**Dual-uncertainty modeling.** After the initial reconstruction training, RUM and EUM are optimized concurrently for 800 iterations using separate optimizers and disjoint gradient paths. For RUM, the encoder $E$ remains frozen and only normal images are used. The reconstruction-difficulty heads are optimized by $\mathcal{L}_{\text{diff}}$ using AdamW with learning rate $10^{-3}$, while the bottleneck $B$ and decoder $D$ are refined by $\mathcal{L}_{\text{ref}}$ with learning rate $2 \times 10^{-4}$. The predicted reconstruction scales determine the weights assigned to hard-normal patches during refinement.

In parallel, EUM is optimized by $\mathcal{L}_{\text{EUM}} = \mathcal{L}_{\text{evi}} + \beta \mathcal{L}_{\text{cri}}$ using AdamW with learning rate $5 \times 10^{-4}$. This objective updates only the projection head $P_\phi$ and criterion bases $\{B_c\}_{c \in \mathcal{C}}$, while the encoder and the complete reconstruction pathway remain frozen with respect to EUM. The easy-normal, hard-normal, and anomaly subspace ranks are set to 8, 4, and 4, respectively. Each normal batch is paired with the $K = 10$ mask-annotated anomaly supports, and mask-aware defect-paste augmentation is applied with probability 0.5. Consequently, anomaly supervision is confined to EUM and does not back-propagate into RUM or alter the reconstruction model.

**Batch size and inference.** We use a training batch size of 16 and the same batch size for evaluation. All score standardization and normal-gate thresholds are computed separately for each category from its normal training set. The reported inference setting is $(\lambda_R, \lambda_D, q_N) = (0.1, 0.1, 0.5)$.

### A.5 Additional Pixel-level Visualization

Additional qualitative results compare the pixel-level anomaly localization maps of Dinomaly, ArcAD, and EviAD on the Real-IAD dataset. As shown in Figure A.1, EviAD produces more accurate anomaly localization with fewer irrelevant activations on normal regions compared with Dinomaly and ArcAD. In particular, EviAD better preserves the spatial structure of anomalous regions while suppressing responses caused by complex normal patterns. Compared with reconstruction-based predictions from Dinomaly, EviAD reduces excessive activations in normal areas and provides clearer localization of defect regions. Compared with ArcAD, EviAD achieves more consistent anomaly maps across different defect types, indicating its effectiveness in handling diverse industrial anomalies. These qualitative results are consistent with the quantitative improvements in pixel-level metrics reported in the main paper.

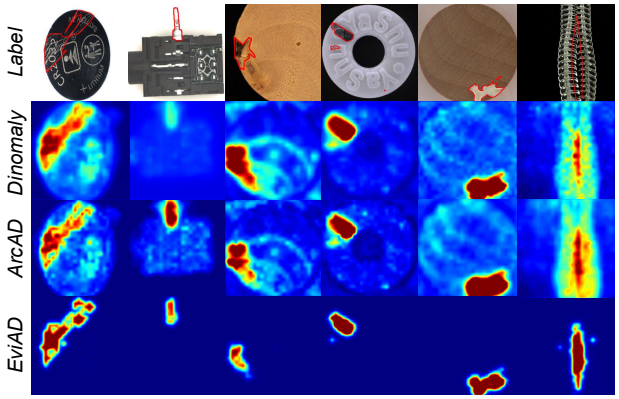



Figure A.1: **Pixel-level anomaly localization comparison on Real-IAD.**