A Metacognitive Perspective on Anomaly Detection with Vision Foundation Models

Matteo Galiazzo[1], **Gabriele Onorato**[1,*], **Dario Loi**[1],
Nathaniel D. Bastian[2], **Francesco Restuccia**[1]

[1]Northeastern University
[2]Pathfinder Applied Research, LLC
[*]onorato.g@northeastern.edu

Abstract

Anomaly Detection (AD) in images identifies defective regions by scoring against a model of normality learned from defect-free data. Current state of the art methods encode images via a frozen Vision Foundation Model (VFM) backbone and score them by their Nearest-Neighbor (NN) distance to a memory bank of normal samples. Recently, due to the parallel evolution of VFM encoders and scoring mechanism, the need to disentangle their individual contributions has emerged, in order to guarantee the deployment of the best performing model in specific industrial production environments. In our work, we analyze AD within a metacognitive perspective, we freeze the backbone and swap the scorer, replacing NN distance with a Mahalanobis score over a Gaussian distribution. By doing that, we can evaluate two different metacognitive functions of the scoring mechanisms: monitoring how far an observation is from the learned model of normality via a threshold-free metric, and deciding when the difference justifies an action according to a threshold. We show that these two metacognitive functions can dissociate. We further notice that replacing a NN memory bank with a simpler Mahalanobis model improves pixel-level monitoring ($+8.00$ AU-PRO$_{0.05}$ on MVTec AD 2) while degrading the segmentation performance after the thresholding step (-13.46 F_1). Lastly, we show that a threshold derived by Extreme Value Theory (EVT) recovers that loss without using anomalous data: the F_1 difference between the two scorers narrows from 13.46 to 0.61 points on MVTec AD 2, with consistent gains on MVTec AD and VisA.

Code — https://github.com/Restuccia-Group/metacog-ad

Introduction

The work of Flavell (1979) proposes a model of cognitive monitoring to understand how humans access their inner thoughts and knowledge to supervise and control their own cognitive progress. This concept was later decomposed into a two-level framework by Nelson and Narens (1990), where they formalize metacognition as two interacting levels: an object level that performs the cognition and a meta level that holds a model of the object level. The two levels are connected by monitoring, the flow of information from the object level to the meta level, and control, the commitment to an action based on that information. Monitoring and control can fail independently: a subject could monitor correctly by producing an accurate judgment, but still fail to produce a good decision (Ackerman and Thompson 2017).

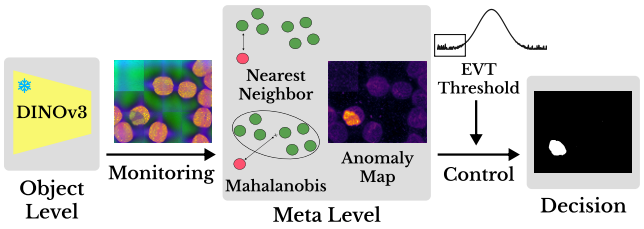


Figure 1: *The metacognitive reading*: Anomaly Detection is a two-level system: a frozen encoder as the object level, and a scorer as the meta level that monitors it. The threshold commits to a decision from the meta level's output.

This paradigm appears in modern Artificial Intelligence (AI) systems: a model operates on the knowledge that it acquired during the training, but it can't recognize *when* a new input falls outside that knowledge. Out-of-Distribution (OOD) detection addresses this by measuring how much the model's outputs deviate from the training distribution (Hendrycks and Gimpel 2017; Sayyed et al. 2026). In structural Anomaly Detection (AD) the objective is to localize defective regions. To do so, the anomaly scorer compares the patch features of a frozen Vision Foundation Model (VFM) with its learned representation of normal data and produces a per-pixel anomaly map (Roth et al. 2022; Zhang et al. 2025; Roming et al. 2026).
Key Issue. The best performing scorers in structural AD are often based on memory banks: collections of patch features extracted from normal training images and queried by Nearest-Neighbor (NN) distance. Since the encoder and the scorer are evaluated together it is difficult to separate their individual impact on anomaly ranking and localization.
Metacognitive Perspective. In this setup we can see the frozen VFM as the object level: it produces patch features but it is not able to score them. The scorer is the meta level, holding a learned model of the object level's outputs. The patch features are the monitoring signal described by Nelson and Narens (1990), since they carry the information from the object level to the meta level, and the anomaly score

[*]Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

is the judgment that the meta level produces from them. Threshold-free metrics such as AU-PRO evaluate the quality of the monitoring signal by integrating the per-region overlap over a range of false-positive rates. Deploying the system however requires choosing a single threshold, and even if the ranking is correct, the threshold can be placed at the wrong level. Fleming and Lau (2014) identify this as the dissociation between metacognitive sensitivity and bias, and recommend evaluating them with dedicated metrics. Segmentation F_1 depends on both the ranking and the threshold, so it can't show which of the two is responsible when the performance changes. In this work, we report both AU-PRO and segmentation F_1, and show that their performance gap can be shrunk by improving the threshold estimation.

Our Analysis. Starting from SuperADD (Roming et al. 2026), the SOTA memory-bank model, we fix the VFM backbone and replace NN distance to the patches with Mahalanobis distance to a Gaussian fitted on normal patches. Despite the reduction in FLOPs and memory usage, the Mahalanobis scorer remains competitive with the memory bank baseline. On MVTec AD 2 (Heckler-Kram et al. 2026) the new scorer improves monitoring (AU-PRO$_{0.05}$) but degrades control (segmentation F_1) under the dataset's threshold: as in the metacognitive literature, the two dissociate. An Extreme Value Theory (EVT)-based threshold recovers most of the gap without using anomalous data. In summary, our main contributions are as follows:

❶ We associate the Nelson–Narens metacognitive architecture to AD from frozen VFMs, identifying analogies. Following Fleming and Lau (2014), we separate monitoring sensitivity (threshold-free metrics) from decision calibration (thresholded metrics), and show that they can dissociate.

❷ We show that fitting a single Gaussian per layer over the patch tokens achieves spatial anomaly localization without a memory bank, with ≈ 78 times lower scoring cost per token. The Mahalanobis scorer improves mean AU-PRO$_{0.05}$ from 56.81 of the NN model to 64.81 ($+8.00$) on MVTec AD 2.

❸ We propose an unsupervised segmentation threshold based on EVT. On MVTec AD 2 it improves segmentation F_1 score from 19.88 to 49.39 for the Mahalanobis scorer ($+29.51$), and from 33.34 to 50.00 for the NN one ($+16.66$), with larger gains on MVTec AD ($+21.48$ and $+29.57$) and VisA ($+32.68$ and $+39.95$).

❹ We conduct comprehensive evaluations across the MVTec AD, MVTec AD 2, and VisA datasets, showing that the calibration gap between published thresholds and achievable performance is consistent across datasets.

Related Work

Memory Bank methods. Memory-bank based approaches store encoded patches of normal samples, and score them via NN distance to this bank. PatchCore (Roth et al. 2022) introduces k-Nearest Neighbor (kNN) matching multi-resolution Convolutional Neural Network (CNN) feature maps and makes the approach practical at scale by combining locally-aware mid-level wide ResNet (Zagoruyko and Komodakis 2016) features with greedy coreset subsampling, shrinking the memory bank by up to $100\times$ with negligible loss in accuracy. Recent foundation-model approaches such as AnomalyDINO (Damm et al. 2025), SuperAD (Zhang et al. 2025), and SuperADD (Roming et al. 2026) apply the same coreset subsampling idea directly to Vision Transformer (ViT) patch tokens extracted from self-supervised DINO family backbones (Oquab et al. 2024; Siméoni et al. 2026), thus improving few-shot and zero-shot performance without any task-specific training. While effective, memory-bank approaches suffer from a memory and computational bottleneck during inference; despite the coreset selection, the model still requires storing thousands of patch descriptors, and looking them up via NN search becomes a bottleneck, particularly in latency- or memory-constrained deployment.

Parametric Modeling methods. Parametric approaches avoid storing a memory bank, and instead summarize normal representations directly as the parameters of a statistical model, so that the inference cost is independent of the number of normal training images. Rippel et al. (2021) model the distribution of CNN features with a single multivariate Gaussian per feature stage and score queries via Mahalanobis distance. Due to global average pooling, they operate in a rank-deficient regime, thus they resort to data augmentation techniques and Ledoit-Wolf shrinkage (Ledoit and Wolf 2004) to stabilize covariance matrix estimation. To introduce spatial localization that Rippel et al. (2021) lacks, PaDiM (Defard et al. 2021) fits position-conditioned Gaussians to the multi-layer CNN feature vector. SubspaceAD (Lendering, Akdag, and Bondarau 2026) evaluates parametric modeling in the few-shot setting: rather than fitting a full-rank Gaussian per patch position, it fits a single low-rank PCA subspace to the covariance matrix, scoring anomalies by the magnitude of the reconstruction residual orthogonal to that subspace.

Threshold calibration. Producing a segmentation mask from a continuous anomaly map requires choosing a threshold. Many papers only report $F_1^{\max}$, sweeping the threshold over the labeled test set (Roth et al. 2022; Akcay et al. 2022; Zhang et al. 2025). This metric evaluates ranking quality, not a deployable decision rule. MVTec AD 2 (Heckler-Kram et al. 2026) instead proposes thresholding at $\mu+3\sigma$, estimated on defect-free validation images. This is a simple idea, but it implicitly assumes that the score tail is Gaussian. MEBin (Huang et al. 2025) avoids using labels and a validation set by sweeping thresholds on each test image and picking the range where the count of connected components plateaus. Their criterion depends on the test data rather than a model of normality. U-Flow (Tailanian, Pardo, and Musé 2024) calibrates the threshold just on normal data against a fixed budget of expected false detections per image. Their threshold derivation is tied to the normalizing flow likelihood, and doesn't transfer to other scoring methods. We use an FPR budget without tying the threshold to a specific scoring model. We fit the tail of the score distribution on normal images using EVT, as described in the thresholding section.

Metacognition. Nelson and Narens (1990) split cognition into an object level and a meta level, connected by monitoring and control. The relation is asymmetric: the object level has no model of the meta level, the same way as a VFM that produces features but cannot score them. Since monitoring is logically independent of control, the two can succeed or fail independently. Fleming and Daw (2017) formalize

self-evaluation as a second-order inference on a coupled but distinct decision system, which is why the quality of the meta level is not fixed by the quality of the object level. Monitoring is cue-based and inferential: Koriat (1997) shows that judgments are driven by proxy cues, not direct access to the underlying state, and Ackerman and Thompson (2017) generalize this into a meta-reasoning framework. An anomaly scorer works in the same way: it never observes defectiveness, just features that correlate with it. Fleming and Lau (2014) decompose metacognitive performance into sensitivity and bias, recommending that they are evaluated separately. Koriat (1997) reports that repeated study improved resolution while impairing calibration, the same dissociation we observe between AU-PRO and thresholded F_1. Fixing the backbone and varying only the scorer isolates the contribution of the meta level from that of the object level.

Methodology

Our pipeline is divided into three main stages: a frozen encoder that turns an image into a grid of patch tokens, a scorer that maps each token to a scalar anomaly score, and a threshold that turns the resulting score map into a binary mask. We keep the encoder stage identical to SuperADD (Roming et al. 2026), the winner of the VAND 4.0 CVPR challenge on the MVTec AD 2 dataset, which we use as our baseline by running their published code unmodified.

Shared pipeline. Given a test image, we downscale and tile it into 640×640 patches with $128\,\mathrm{px}$ overlap. Each tile is processed by a frozen DINOv3 backbone (Siméoni et al. 2026). We extract patch tokens at four equally spaced depths ($25\%, 50\%, 75\%,$ and 100% of the network depth). Overlapping regions are coupled, and per-tile token grids are unified into a single spatial feature map $Z_\ell \in \mathbb{R}^{H\times W\times d}$ per layer, where $z_\ell(p) \in \mathbb{R}^d$ is the token at position p. The scorer maps each token $z_\ell(p)$ to a scalar anomaly score $s_\ell(p)$, producing one anomaly map per token extraction layer. The four per-layer maps are bilinearly upsampled to the original image resolution and then combined by unweighted mean, $S(p) = \frac{1}{4}\sum_{\ell\in\mathcal{L}} s_\ell(p)$. To turn S into a binary segmentation mask, a threshold t is applied and the resulting mask is post-processed with anomaly region filling.

Mahalanobis distance scorer. We replace the memory bank of the baseline with a single parametric density model per layer: a d-dimensional Gaussian $\mathcal{N}(\mu_\ell, \Sigma_\ell)$ fitted to all normal training tokens at layer ℓ, pooled over every spatial position and every training image. Unlike Rippel et al. (2021), who pool features globally, and PaDiM (Defard et al. 2021), which fits one Gaussian per position, we fit a single Gaussian over all grid tokens while retaining spatial localization; unlike SubspaceAD (Lendering, Akdag, and Bondarau 2026), the full-rank distance requires no choice of subspace dimension.

Fitting. The standard covariance estimator requires storing all n tokens to compute the mean and covariance, using $\mathcal{O}(nd)$ space. We accumulate the token count, the sum $\sum_i z_i$ and the sum of outer products $\sum_i z_i z_i^\top$ to compute them without storing all the tokens. Then we can calculate the mean and covariance in closed form: $\mu_\ell = \frac{1}{n}\sum_i z_i$ and $\Sigma_\ell = \frac{1}{n}\sum_i z_i z_i^\top - \mu_\ell\mu_\ell^\top$. The time complexity is $\mathcal{O}(nd^2)$ in both cases, but by doing it in one pass we use $\mathcal{O}(d^2)$ space ($52.47\,\mathrm{MB}$ across the four layers) instead of $\mathcal{O}(nd)$ (4.7–$16.8\,\mathrm{GB}$ for ViT-H+ on MVTec AD 2). Because each training image contributes $H \times W$ tokens, $n \gg d$, the covariance matrix is guaranteed to be full-rank without the synthetic data augmentations and shrinkage of Rippel et al. (2021) or the augmentations SubspaceAD requires to compensate the few-shot setting (Lendering, Akdag, and Bondarau 2026).

Scoring. Given the fitted parameters $(\mu_\ell, \Sigma_\ell^{-1})$, we score a token z by Mahalanobis distance

$$s_\ell(p) = \sqrt{(z-\mu_\ell)^\top \Sigma_\ell^{-1}(z-\mu_\ell)}. \tag{1}$$

Cost. The fitted model is a vector of size d and a $d \times d$ matrix per layer, over 4 layers, independent of the size of the training set. At inference, scoring a token has cost $\mathcal{O}(d^2)$. The memory bank baseline stores a coreset of m vectors per layer and scores each token by NN search with cost $\mathcal{O}(md)$ per query. In the default ViT-H+ case, with $d = 1280$ and $m = 10^5$ the Mahalanobis scorer is $\frac{md}{d^2} = \frac{10^5}{1280} \approx 78\times$ cheaper per token against brute-force search.

Thresholding. To deploy the model, it needs to be able to commit to a decision, so we need to binarize the anomaly map S. Let $\mathcal{V}_c$ be the set of defect-free images of category c, held out from the images used to fit μ_ℓ and Σ_ℓ, and let $H_c = \{S(p) : p \in I, I \in \mathcal{V}_c\}$ be the pool of per-pixel anomaly scores. We set an acceptable per-pixel FPR α on normal data and set the threshold at the $1-\alpha$ quantile of H_c. We write α as $\alpha_k = 1 - \Phi(k)$, where Φ is the standard normal cumulative distribution function. At this FPR budget, the quantile is far into the tail, where the empirical estimate is unreliable. Therefore we fit a Generalized Pareto Distribution (GPD) to the excesses above $u = Q_{0.99}(H_c)$ by peaks-over-threshold (Coles et al. 2001) and get the threshold from the fitted tail: $t_c = \hat{q}_{1-\alpha_k}$, the $1-\alpha_k$ quantile of the fitted GPD, estimated per category. We measure the FPR on held-out normal images by 5-fold cross-validation. On a 640×640 image, $k = 3$ admits ≈ 553 false-positive pixels, and $k = 5$ targets 0.1 pixels but delivers 0.9–3.6; we use $k = 4$ which admits $\approx$ 18–20 pixels, yielding the tightest reliable budget.

Experiments

Experimental setup. We evaluate the models on three datasets. MVTec AD (Bergmann et al. 2019), MVTec AD 2 (Heckler-Kram et al. 2026), introduced as its successor to address the saturation on the original one, and VisA (Zou et al. 2022). We run each model configuration with three seeds and report their mean and standard deviation. As the backbone we use DINOv3 (Siméoni et al. 2026) ViT-S, ViT-B, ViT-L, ViT-H+, all with patch size 16; Fig. 2 reports every size, while Tables 1 and 2 uses ViT-H+. We evaluate the monitoring sensitivity with AU-PRO and image AUROC, and the decision calibration with segmentation F_1 with four thresholding rules: percentile (Q_{99}), 3σ rule ($\mu+3\sigma$), EVT tail rule (EVT t_c), and the oracle $F_1^{\max}$ as the performance ceiling. All the table comparisons are made with a two-sided Welch's t-test with Holm-Bonferroni correction, applied to the per-category differences.

Dataset	Q_{99}		$\mu{+}3\sigma$		EVT t_c		$F_1^{\max}$	
	NN	Mahalanobis	NN	Mahalanobis	NN	Mahalanobis	NN	Mahalanobis
VisA	$11.08_{\pm 0.03}$	$11.67_{<0.01}$	$14.54_{\pm 0.01}$	$16.01_{\pm 0.01}$	$\mathbf{47.22}^{\dagger}_{\pm 0.19}$	$\mathbf{55.96}^{\dagger}_{\pm 0.01}$	$55.45_{\pm 0.04}$	$60.52_{<0.01}$
MVTec AD	$22.65_{\pm 0.03}$	$25.66_{<0.01}$	$30.72_{\pm 0.05}$	$33.41_{\pm 0.02}$	$\mathbf{52.20}^{\dagger}_{\pm 0.74}$	$\mathbf{62.98}^{\dagger}_{\pm 0.03}$	$66.68_{\pm 0.04}$	$67.92_{<0.01}$
MVTec AD 2	$10.82_{\pm 0.09}$	$12.71_{<0.01}$	$33.34_{\pm 0.05}$	$19.88_{\pm 0.01}$	$\mathbf{50.00}^{\dagger}_{\pm 0.50}$	$\mathbf{49.39}^{\dagger}_{<0.01}$	$63.15_{\pm 0.08}$	$56.42_{<0.01}$

Table 1: Segmentation F_1 (%) across thresholds compared to the $F_1^{\max}$ oracle, values reported as mean $\pm$ std across 3 seeds; best values in **bold**. † marks significant improvements over the baseline ($p < 0.01$, Welch's t-test, Holm-Bonferroni corrected).

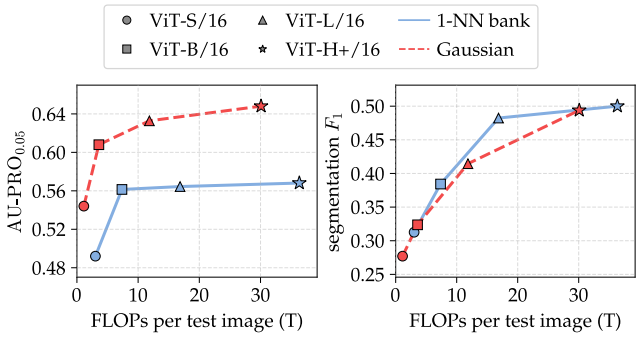


Figure 2: MVTec AD 2 test set accuracy over FLOPs. The two plots dissociate: the Mahalanobis scorer leads on monitoring (AU-PRO$_{0.05}$) while using less compute, but on control (segmentation F_1 at t_c) its advantage disappears.

Monitoring quality. We report the threshold-free metrics in Table 2. On MVTec AD 2 the Mahalanobis scorer improves the mean AU-PRO$_{0.05}$ from 56.81 to 64.81 over the memory bank (+8.00 , $p < 0.001$). On VisA the Mahalanobis scorer leads the mean AU-PRO$_{0.3}$ with 95.27 vs. 93.60 of the NN (+1.67, $p < 0.001$), and on MVTec AD the NN leads with 95.88 versus 95.51 of the Mahalanobis (+0.36, $p < 0.001$). Image AUROC is close on all 3 datasets, always in favor of the memory bank, but this difference is only statistically significant on MVTec AD ($p < 0.001$). Overall, across the six threshold-free comparisons, the Mahalanobis scorer proves that its monitoring signal is competitive with the memory bank's, improving ranking on two out of three datasets for AU-PRO, at a fraction of the scoring cost. Figure 2 shows that the Mahalanobis scorer has better AU-PRO$_{0.05}$ per FLOP at every backbone size. In the Fleming and Lau (2014) the metacognitive sensitivity of the Mahalanobis scorer is higher on VisA and MVTec AD 2, and slightly lower on MVTec AD.

Decision calibration. Table 1 reports the segmentation F_1 at different operating points. The $\mu{+}3\sigma$ rule would deliver a per-pixel FPR of 1.35×10^{-3} if the scores were Gaussian. On MVTec AD 2, the true FPR on held-out normal images is 1.9×10^{-3} for NN and 4.6×10^{-3} for Mahalanobis, because the mean and variance are set by the center of the distribution, not the tail. This difference is reflected in the segmentation F_1 score: on MVTec AD 2 $\mu{+}3\sigma$ costs the Mahalanobis scorer 13.46 more segmentation F_1 score points than the NN one (19.88 vs 33.34, $p < 0.001$) despite the higher AU-PRO$_{0.05}$. This is the sensitivity-bias dissociation of Fleming and Lau (2014): the ranking improved, but the threshold did not. The EVT threshold improves the Mahalanobis scorer by 29.51 points of segmentation F_1 ($p < 0.001$) and the NN one by 16.66 ($p = 0.128$). Under the EVT threshold the gap between the scorers narrows from 13.46 to 0.61 ($p = 0.169$). At the oracle threshold the NN still leads on F_1 by 6.73 (63.15 vs. 56.42), so about half the deficit was threshold placement, and the other half remains even at the oracle threshold. On MVTec AD and VisA the Mahalanobis scorer holds a small but significant lead under $\mu{+}3\sigma$ ($p < 0.001$ for both), and the EVT threshold makes both methods better. The NN scorer gains 21.48 points on AD and 32.68 on VisA, and the Mahalanobis one 29.57 and 39.95, with all four significant ($p < 0.001$). Unlike MVTec AD 2, the EVT rule separates the scorers on the other two datasets: the Mahalanobis scorer is 10.78 points better than the NN on MVTec AD ($p = 0.003$) and 8.74 points better on VisA ($p < 0.001$); at the oracle threshold the gap on MVTec AD narrows to only 1.24.

Dataset	AU-PRO		Image AUROC	
	NN	Mahalanobis	NN	Mahalanobis
VisA	$93.60_{\pm 0.08}$	$\mathbf{95.27}^{\dagger}_{<0.01}$	$\mathbf{96.80}_{\pm 0.16}$	$96.53_{<0.01}$
MVTec AD	$\mathbf{95.88}^{\dagger}_{<0.01}$	$95.51_{<0.01}$	$\mathbf{99.21}^{\dagger}_{<0.01}$	$97.94_{<0.01}$
MVTec AD 2	$56.81_{\pm 0.23}$	$\mathbf{64.81}^{\dagger}_{<0.01}$	$\mathbf{85.03}_{\pm 0.42}$	$83.68_{\pm 0.03}$

Table 2: Threshold-free monitoring, mean $\pm$ std across 3 seeds; best in **bold**. We report AU-PRO$_{0.05}$ for MVTec AD 2, AU-PRO$_{0.3}$ for MVTec AD and VisA. † marks a significant NN–Mahalanobis difference ($p < 0.01$, Welch's t-test).

Discussion and Conclusion

Analyzing AD through the monitoring–control framework of Nelson and Narens (1990) shows that on the same signal monitoring can improve while control degrades, and that control can be independently recovered. With the encoder fixed, a Mahalanobis scorer raises AU-PRO$_{0.05}$ by 8.00 points on MVTec AD 2 at a fraction of the scoring cost, and an EVT tail threshold recovers its thresholded F_1 on all three datasets without anomalous data. Threshold calibration should therefore be evaluated and reported separately from ranking quality.

Acknowledgements

This work was supported in part by NSF under Grants CNS-2312875 and OAC-2530896, DARPA under Cooperative Agreement D25AC00374-00, ONR under Award N00014-23-1-2221, and AFOSR under Award FA9550-23-1-0261.

References

Ackerman, R.; and Thompson, V. A. 2017. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21(8): 607–617.

Akcay, S.; Ameln, D.; Vaidya, A.; Lakshmanan, B.; Ahuja, N.; and Genc, U. 2022. Anomalib: A deep learning library for anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, 1706–1710. IEEE.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9584–9592.

Coles, S.; Bawa, J.; Trenner, L.; and Dorazio, P. 2001. *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Damm, S.; Laszkiewicz, M.; Lederer, J.; and Fischer, A. 2025. AnomalyDINO: Boosting Patch-based Few-shot Anomaly Detection with DINOv2. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1319–1329.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Pattern Recognition. ICPR International Workshops and Challenges (ICPR 2020 Workshops), Part IV*, volume 12664 of *Lecture Notes in Computer Science*, 475–489. Cham: Springer.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10): 906–911.

Fleming, S. M.; and Daw, N. D. 2017. Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation. *Psychological Review*, 124(1): 91–114.

Fleming, S. M.; and Lau, H. C. 2014. How to measure metacognition. *Frontiers in Human Neuroscience*, 8: 443.

Heckler-Kram, L.; Neudeck, J.-H.; Scheler, U.; König, R.; and Steger, C. 2026. The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection. *International Journal of Computer Vision (IJCV)*, 134(4).

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Huang, Z.; Li, X.; Liu, H.; Xue, F.; Wang, Y.; and Zhou, Y. 2025. AnomalyNCD: Towards Novel Anomaly Class Discovery in Industrial Scenarios. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 4755–4765.

Koriat, A. 1997. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4): 349–370.

Ledoit, O.; and Wolf, M. 2004. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 88(2): 365–411.

Lendering, C.; Akdag, E.; and Bondarau, E. 2026. SubspaceAD: Training-Free Few-Shot Anomaly Detection via Subspace Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28557–28566.

Nelson, T. O.; and Narens, L. 1990. Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation*, volume 26, 125–173. Elsevier. ISBN 978-0-12-543326-6.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*.

Rippel, O.; Mertens, P.; König, E.; and Merhof, D. 2021. Gaussian Anomaly Detection by Modeling the Distribution of Normal Data in Pretrained Deep Features. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–13.

Roming, L.; Lehnerer, F.; Funk, J. V.; Michel, A.; Maier, G.; Längle, T.; and Beyerer, J. 2026. SuperADD: Training-free Class-agnostic Anomaly Segmentation. arXiv:2605.14808.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14298–14308.

Sayyed, A. Q. M. S.; Rifat, S.; Zhang, M.; Swami, A.; De Lucia, M.; Bastian, N. D.; and Restuccia, F. 2026. Out-of-Distribution Detection in Computer Vision: A Comprehensive Survey and Research Challenges. *ACM Computing Surveys*.

Siméoni, O.; Vo, H. V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; Massa, F.; Haziza, D.; Wehrstedt, L.; Wang, J.; Darcet, T.; Moutakanni, T.; Sentana, L.; Roberts, C.; Vedaldi, A.; Tolan, J.; Brandt, J.; Couprie, C.; Mairal, J.; Jégou, H.; Labatut, P.; and Bojanowski, P. 2026. DINOv3. *Transactions on Machine Learning Research (TMLR)*.

Tailanian, M.; Pardo, A.; and Musé, P. 2024. U-Flow: A U-Shaped Normalizing Flow for Anomaly Detection with Unsupervised Threshold. *Journal of Mathematical Imaging and Vision*, 66(4): 678–696.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In Richard C. Wilson, E. R. H.; and Smith, W. A. P., eds., *Proceedings of the British Machine Vision Conference (BMVC)*, 87.1–87.12. BMVA Press. ISBN 1-901725-59-6.

Zhang, H.; Chen, H.; Cheng, Y.; Wu, S.; Sun, L.; Han, L.; Shi, Z.; and Qi, L. 2025. SuperAD: A Training-free Anomaly Classification and Segmentation Method for CVPR 2025 VAND 3.0 Workshop Challenge Track 1: Adapt & Detect. arXiv:2505.19750.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 392–408.