

MS-AD: a Metal Stamping Dataset for Visual Anomaly Detection under Real-World Conditions

Original

MS-AD: a Metal Stamping Dataset for Visual Anomaly Detection under Real-World Conditions / Karacomak, C., Ponzio, F., Bigaran, S., Di Cataldo, S., Scaltrito, L.. - (2026). (2026 IEEE 31st International Conference on Emerging Technologies and Factory Automation (ETFA) Vasteras (SWE) 08-11 September 2026) [10.1109/ETFA68377.2026.11724790].

Availability:

This version is available at: 11583/3015911 since: 2026-10-01T09:31:55Z

Publisher:

IEEE

Published

DOI:10.1109/ETFA68377.2026.11724790

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

MS-AD: a Metal Stamping Dataset for Visual Anomaly Detection under Real-World Conditions

Can Karacomak*, Francesco Ponzio*, Simone Bigaran†, Santa Di Cataldo*, Luciano Scaltrito*

* Politecnico di Torino, Italy, name.surname@polito.it

† MISTA Spa, Asti, Italy, name.surname@mista.it

Abstract—Visual anomaly detection (VAD) has achieved strong results in automated quality control, yet most benchmarks are acquired under controlled laboratory conditions with static backgrounds and uniform illumination. This limits their ability to represent the visual complexity of active production lines. We introduce the Metal Stamping Anomaly Detection (MS-AD) dataset, collected directly from a high-speed metal stamping process under real operational conditions. MS-AD captures dynamic backgrounds, illumination changes caused by reflective surfaces, motion blur, vibration-induced artifacts, and distribution shifts between training and test data. The dataset contains 1,480 images of a metal strip component, with four pixel-wise annotated defect categories. We benchmark representative state-of-the-art anomaly detection methods and show that performance decreases under operational variability, highlighting the need for realistic evaluation settings for industrial visual inspection.

Index Terms—computer vision, visual anomaly detection, industrial inspection, unsupervised learning, dataset benchmarking

I. INTRODUCTION

Computer Vision (CV) has become a key technology for industrial quality control, where the objective is to identify defective products with minimal human intervention. This task, commonly referred to as Visual Anomaly Detection (VAD), has increasingly benefited from unsupervised approaches, in which models are trained exclusively on normal samples and are expected to detect previously unseen defects.

Deep learning-based methods have achieved remarkable results on widely used benchmarks such as MVTec AD and its successors [1]–[5]. In particular, feature-based approaches often reach near-ceiling performance under controlled benchmark conditions [6]–[8]. However, performance commonly decreases as acquisition conditions become more complex, revealing a gap between benchmark success and robustness in real industrial environments.

A major limitation of existing datasets is that they are largely acquired under controlled laboratory conditions, with static backgrounds, stable illumination, and limited acquisition variability. Although recent benchmarks introduce more challenging objects, multiple viewpoints, or controlled lighting variations, they only partially reflect the conditions en-

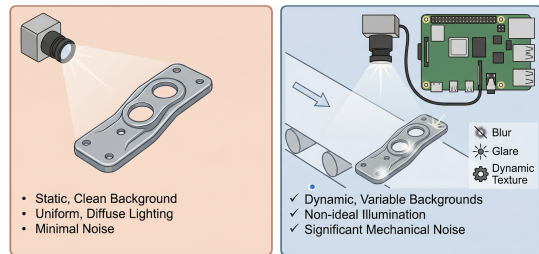


Fig. 1: Controlled laboratory conditions (left) vs real industrial environment represented in MS-AD (right).

countered in active production lines. In real manufacturing environments, inspection systems must operate under mechanical vibrations, motion blur, dynamic backgrounds, positional variability, and rapidly changing illumination, as illustrated in Fig. 1.

To address this limitation, we introduce the *Metal Stamping Anomaly Detection (MS-AD)* dataset, collected directly from a high-speed metal stamping process operating under real production conditions. Unlike existing benchmarks, MS-AD explicitly captures vibration-induced blur, dynamic backgrounds, illumination variability caused by reflective surfaces, and train–test distribution shifts. It also provides pixel-wise annotations for four defect categories on a new industrial object class.

The main contributions of this work are: (i) the introduction of MS-AD, a new VAD dataset collected from an active metal stamping production line; (ii) the explicit modeling of operational variability, including motion blur, vibration, background changes, illumination variations, and distribution shifts; and (iii) a benchmark of representative state-of-the-art VAD methods under realistic operating conditions.

II. BACKGROUND AND RELATED WORK

VAD has been extensively studied in manufacturing, leading to the introduction of numerous benchmark datasets for evaluating state-of-the-art (SOTA) methods.

Commonly Known Benchmarks. MVTec AD [1] is the most widely adopted VAD benchmark, providing high-resolution images acquired under controlled laboratory conditions with static backgrounds and uniform illumination. While

this setup enables reproducible evaluation, it offers limited acquisition variability.

Several datasets extend MVTec AD along different dimensions of complexity. MVTec AD 2 [2] introduces transparent and overlapping objects, greater intra-class variability, smaller defects, and controlled lighting variations. VisA [3] focuses on cluttered scenes, complex object structures, and multiple object instances. BTAD [4] targets automotive manufacturing defects, while Real-IAD [5] provides diverse object categories and defect distributions.

Despite these advances, existing benchmarks mainly increase complexity through object characteristics, scene composition, or controlled acquisition variations. Disturbance-prone effects common in production environments, such as motion blur, background variability, and uncontrolled illumination changes, remain largely underrepresented. This motivates the development of complementary benchmarks that explicitly capture operational variability and enable a more comprehensive assessment of model robustness.

Visual Anomaly Detection Methods. VAD methods differ in how normality is modeled and deviations are identified. Early approaches relied on *image reconstruction*, typically using autoencoders trained on normal samples, where anomalies are detected through reconstruction errors. Although effective in controlled settings, these methods are often sensitive to acquisition-related variability [9].

More recent approaches operate on *feature-level representations*, measuring deviations in learned feature spaces rather than reconstructing inputs [6], [8], [9]. In this work, we focus on three widely adopted paradigms:

(i) *Knowledge distillation* methods compare feature representations produced by a fixed *teacher* network and a *student* network trained on normal data. RD [10] formulates anomaly detection as feature reconstruction, RD++ [7] introduces reverse distillation with a constrained bottleneck, and EfficientAD [8] combines a lightweight student-teacher architecture with an auxiliary autoencoder to improve efficiency.

(ii) *Feature-based discrimination* methods operate directly in embedding space, learning to separate normal and anomalous patterns. SimpleNet [11] generates synthetic anomalies and trains a lightweight classifier to distinguish them.

(iii) *Memory-based* methods compare test samples against stored normal representations. PatchCore [6] employs a coreset-based memory bank and nearest-neighbor distances to compute anomaly scores, making performance dependent on feature similarity and distribution shifts.

To evaluate MS-AD, we select EfficientAD and RD++ as representatives of knowledge distillation, SimpleNet for feature-based discrimination, and PatchCore for memory-based methods. These models were chosen for their strong performance, widespread adoption, and suitability for industrial deployment. Together, they provide complementary perspectives on robustness under realistic acquisition variability.

Evaluation Metrics. Evaluating anomaly localization is challenging because performance depends not only on detection accuracy but also on the ability to capture the spatial

structure and scale of defects. Since different metrics respond differently to class imbalance, defect size, and anomaly distribution, VAD evaluation commonly combines pixel-, region-, and image-level measures [9].

1) *Traditional Metrics:* Traditional VAD metrics are primarily based on pixel-level classification performance. Pixel-wise AUROC measures the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across thresholds and is therefore widely used. However, in highly imbalanced industrial settings, AUROC may overestimate performance despite poor localization of small defects.

AUPRC evaluates the precision-recall trade-off and is generally more suitable for imbalanced data, but it is strongly influenced by anomaly prevalence, complicating comparisons across datasets. Similarly, the F1-score depends on a fixed decision threshold, limiting its usefulness for comprehensive model ranking. These limitations also apply when such metrics are extended to image-level evaluation.

2) *Advanced Metrics:* To address these limitations, recent metrics focus on region- and image-level evaluation.

AU-PRO measures the overlap between predicted and ground-truth regions up to a predefined FPR limit (e.g., 30%), treating each defect region independently. This reduces the dominance of large anomalies and gives smaller defects a more balanced contribution to the evaluation [1].

AUPIMO further evaluates overlap on a per-image basis, reducing the influence of defect volume and improving comparability across varying anomaly distributions [12]. It also adopts the Shared False Positive Rate (S-FPR) framework, where thresholds are determined exclusively from normal samples, enforcing a consistent noise floor and penalizing excessive false positives. As a result, AUPIMO provides a stricter and often more realistic assessment.

Since each metric emphasizes different aspects of performance, no single measure fully characterizes anomaly detection and localization performance, and defining a comprehensive evaluation framework remains an open challenge.

III. DATASET

MS-AD was collected from an active metal stamping production line under realistic operating conditions. The target object is a metal strip acquired after cutting and punching operations and before bending, where defects may arise from the stamping process or raw material imperfections.

Defect Types. The dataset contains four defect categories: *scratch*, *hit*, *black spot*, and *natural*. Since natural anomalies are rare, scratches and hits were induced with sharp tools to simulate broken-punch damage, and black spots were added using paint to simulate the most common color change. Natural anomalies were annotated separately. Samples with multiple defects receive a *mixed* label. Representative examples and pixel-wise masks are shown in Fig. 2.

Acquisition and Statistics. Images were acquired using a Raspberry Pi 5 connected to a Raspberry Pi HQ camera (12.3 MP Sony IMX477 sensor) equipped with a 12 mm low-distortion lens and synchronized with the production line

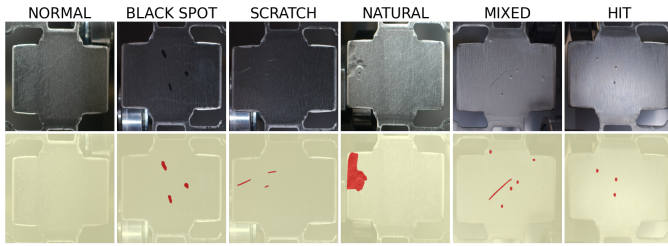


Fig. 2: Examples of normal and anomalous samples from MS-AD, along with corresponding pixel-wise segmentation masks.

through GPIO triggering. The dataset contains 1,480 images at a resolution of 640×640 pixels. The training set includes 561 defect-free samples, while the test set contains 919 images, comprising 445 normal and 474 anomalous samples. Pixel-wise annotations were generated using CVAT [13]. The distribution of samples across defect categories is moderately balanced, with all defect types represented in both training and testing data.

A distinctive characteristic of MS-AD is the small size of defects relative to the image. The average defect area is only 0.39% of the image, comparable to challenging categories such as Wall Plugs (0.27%) and Sheet Metal (0.58%) in MVTec AD 2 [2], and Screws (0.34%) in MVTec AD [1]. This makes MS-AD a fine-grained anomaly localization benchmark.

Operational Variability. Unlike laboratory datasets, MS-AD captures variability introduced by real production conditions. Small positional shifts (± 0.3 cm) alter the visible background, while the reflective metal surface produces substantial illumination changes. The dataset also includes normal samples acquired under different lighting conditions and slight rotations occurring during production. Furthermore, the high operating speed of the stamping process (2–3 presses per second) introduces vibration-induced blur and motion artifacts.

Preprocessing. To mitigate unavoidable motion blur from constant movement on low-cost edge equipment, 5–10 consecutive frames are acquired and the least blurred is selected. A reference point is then localized to crop the strip. Although necessary for automated acquisition, this procedure introduces positional and scale shifts of up to 20 pixels. These effects are present in both training and testing data, requiring models to learn operational variability while remaining sensitive to subtle defects.

The dataset is provided in the same format as MVTec AD, enabling direct integration with existing evaluation pipelines and reproducible benchmarking under realistic conditions.

IV. EXPERIMENTAL RESULTS

The selected models were evaluated on MS-AD as well as on established benchmark datasets. We first describe the experimental setup and then analyze the results with respect to evaluation metrics, model behavior, and dataset complexity.

Experimental Setup. As described in Section II, we evaluate four representative VAD models (EfficientAD, RD++, SimpleNet, and PatchCore) on MVTec AD, VisA, MVTec AD 2, BTAD, Real-IAD, and MS-AD. All models are trained

exclusively on normal samples following the standard unsupervised anomaly detection setting.

In line with previous studies, performance is measured using AUROC (pixel- and image-level), AU-PRO, and AUPIMO (Section II), enabling evaluation of localization accuracy, image-level detection, and robustness to false positives.

To ensure reproducibility, EfficientAD [8] was trained for 150 epochs using a medium-sized teacher, RD++ [7] for 200 epochs with a WideResNet-50 backbone, SimpleNet [11] for 160 epochs with a ResNet-50 backbone, and PatchCore [6] with an input size of 366 pixels and a 10% coreset subsampling ratio. Data augmentation was limited to random horizontal flipping, as additional transformations degraded performance.

Results and Analysis. VAD evaluation comprises two tasks: *pixel-level anomaly localization* and *image-level classification*.

We first analyze *localization performance*. As defined in Section II, table I shows high AUROC-Px and AU-PRO values compared to a marked drop in AUPIMO, particularly on MS-AD. All models exhibit this behavior despite AU-PRO values above 67%.

TABLE I: Anomaly localization accuracy across datasets (%).

Model	Dataset	AUROC-Px	AU-PRO _{30%}	AUPIMO
EfficientAD	MVTec AD	97.96	93.50	66.08
	VisA	99.00	94.00	58.06
	MS-AD	91.33	84.18	0.003
RD++	MVTec AD	98.23	95.03	71.93
	VisA	99.00	89.30	44.30
	MS-AD	94.04	73.88	8.13
SimpleNet	MVTec AD	97.13	89.60	71.39
	VisA	91.17	68.90	34.66
	MS-AD	91.77	67.01	5.10
PatchCore	MVTec AD	98.01	92.70	67.21
	VisA	98.26	79.70	38.02
	MS-AD	95.11	76.08	1.70

The discrepancy stems from the stricter false-positive constraints imposed by AUPIMO. Unlike AUROC-Px and AU-PRO, which are influenced by large background regions, AUPIMO relies on thresholds derived from normal samples. Consequently, even small background variations can generate false positives, particularly in MS-AD, where the background is dynamic, and lighting variations introduce partial differences between the training and test sets, even for normal samples.

Fig. 3 shows anomaly score distributions for normal and anomalous samples, which are highly overlapped for EfficientAD and PatchCore. Low class separability primarily affects AUPIMO, due to threshold selection from normal samples.

We next consider *image-level classification*. Although state-of-the-art methods typically achieve near-ceiling image-wise AUROC (I-AUROC) on MVTec AD [1], performance on MS-AD drops to 82–88% (Table II), indicating that increased variability affects both localization and image-level discrimination.

Table II jointly reports I-AUROC and AU-PRO under different FPR constraints. While all models achieve high AU-PRO values without restrictions, performance decreases substantially as the FPR limit becomes stricter.

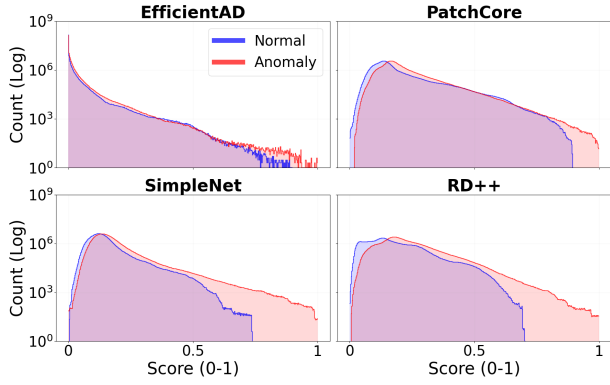


Fig. 3: Pixel-wise anomaly score distributions on MS-AD.

TABLE II: Image-wise AUROC and AU-PRO with different FPR limits on MS-AD (%).

Model	I-AUROC	AU-PRO _{100%}	AU-PRO _{30%}	AU-PRO _{5%}
EfficientAD	85.66	93.84	84.18	57.07
RD++	87.74	91.84	73.88	31.42
SimpleNet	85.71	87.99	67.01	28.25
PatchCore	82.57	92.65	76.08	22.24

These results indicate that increased variability in MS-AD reduces class separability and makes reliable detection under strict false-positive constraints substantially more challenging.

We next analyze generalization across datasets. Table III reports AU-PRO_{30%} across all benchmarks using values from the original publications when available.

TABLE III: Cross-dataset AU-PRO_{30%} (%).

Model	MVTec	VisA	MVTec2	BTAD	Real-IAD	MS-AD
EfficientAD	93.5	94.0	58.7	79.9	86.0	84.1
RD++	95.0	89.3	54.3	77.3	–	73.8
SimpleNet	89.6	68.9	46.4	69.6	89.9	67.0
PatchCore	92.7	79.7	53.8	76.0	90.6	76.1

Across all models, performance is near-ceiling on simpler datasets such as MVTec AD but drops substantially on MVTec AD 2 and MS-AD. EfficientAD shows the most consistent behavior, while the relatively high scores on Real-IAD suggest that different sources of variability affect models differently.

Discussion. The results show that both evaluation metrics and dataset characteristics strongly affect performance assessment. Models ranking highly under AUROC or AU-PRO may exhibit substantially lower performance under stricter metrics such as AUPIMO. Moreover, the consistent degradation observed on more complex datasets suggests that controlled benchmarks do not fully reflect real-world operating conditions, highlighting the need for evaluation protocols that jointly consider multiple metrics and realistic acquisition variability.

V. CONCLUSION

This work introduces MS-AD, a visual anomaly detection dataset designed to bridge the gap between controlled benchmarks and real industrial environments. Although focused

on a single object category, MS-AD captures key sources of operational variability, including vibration-induced motion blur, spatial misalignment, and significant lighting changes.

Experimental results show that SOTA methods experience substantial performance degradation under these conditions, with model rankings varying across evaluation metrics.

By emphasizing robustness to spatial and photometric variability, MS-AD provides a complementary benchmark for assessing the readiness of VAD systems for industrial deployment. Future work will extend the dataset to additional object categories and further refine evaluation protocols for realistic operating conditions. The dataset and source code are publicly available at <https://gitlab.com/vad-iad/MS-AD>.

REFERENCES

- [1] P. Bergmann, M. Fauser, D. Sattlegger, and C. 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*, 2019, pp. 9592–9600.
- [2] L. Heckler-Kram, J.-H. Neudeck, U. Scheler, R. König, and C. Steger, “The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection,” *International Journal of Computer Vision*, vol. 134, no. 4, p. 175, 2026.
- [3] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, “Spot-the-difference self-supervised pre-training for anomaly detection and segmentation,” in *European conference on computer vision*. Springer, 2022, pp. 392–408.
- [4] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, “Vt-adl: A vision transformer network for image anomaly detection and localization,” *arXiv preprint arXiv:2104.10036*, 2021.
- [5] C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. 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*, 2024, pp. 22 883–22 892.
- [6] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, “Towards total recall in industrial anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 318–14 328.
- [7] T. D. Tien, A. T. Nguyen, N. H. Tran, T. D. Huy, S. Duong, C. D. T. Nguyen, and S. Q. Truong, “Revisiting reverse distillation for anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 24 511–24 520.
- [8] K. Batzner, L. Heckler, and R. König, “Efficientad: Accurate visual anomaly detection at millisecond-level latencies,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 128–138.
- [9] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, “A unifying review of deep and shallow anomaly detection,” *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.
- [10] H. Deng and X. Li, “Anomaly detection via reverse distillation from one-class embedding,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9737–9746.
- [11] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, “SimpleNet: A simple network for image anomaly detection and localization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 402–20 411.
- [12] J. P. Bertoldo, D. Ameln, A. Vaidya, and S. Akçay, “Aupimo: Redefining visual anomaly detection benchmarks with high speed and low tolerance,” *arXiv preprint arXiv:2401.01984*, 2024.
- [13] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, T. Osmanov, D. Kruchinin, A. Zankevich, DmitriySidnev, M. Markelov, Johannes222, M. Chenuet, a andre, telenachos, A. Melnikov, J. Kim, L. Ilouz, N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, vugia truong, zliang7, lizhming, and T. Truong, “opencv/cvat: v1.1.0,” Aug. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.4009388>