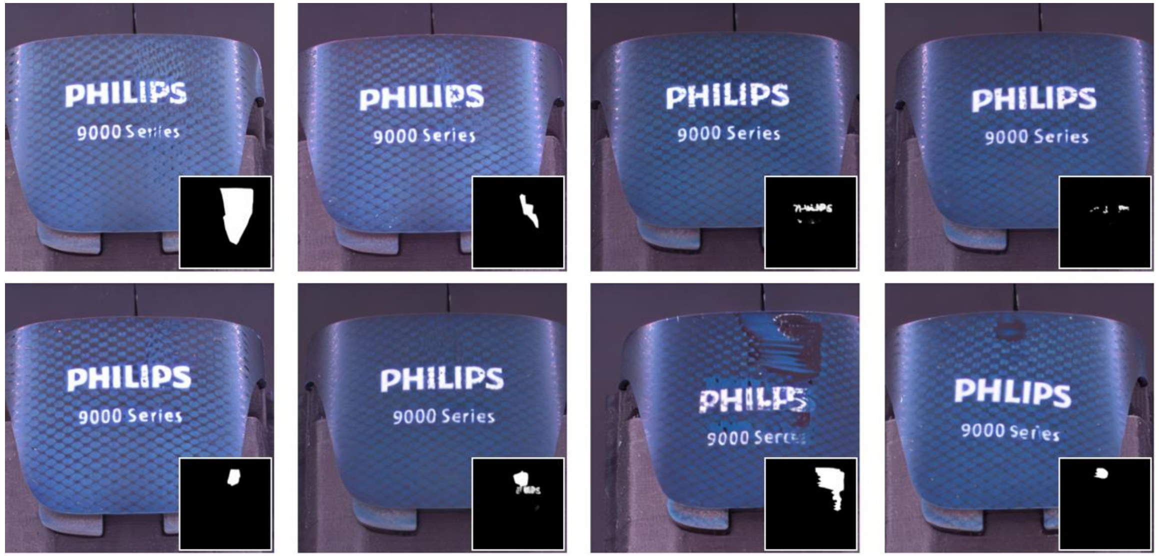
ANOMALY DETECTION · SYNTHETIC DATA · UNCERTAINTY QUANTIFICATION

Trustworthy Visual Quality Inspection under Data Scarcity in Manufacturing



Panagiotis Sapoutzoglou[1] · Jessy Ribaira[2] · Martin Kanounnikoff[2]
Bas Tijsma[3] · Christian Geiß[2,4] · Maria Pateraki[1]

[1] National Technical University of Athens · [2] German Aerospace Center (DLR)
[3] Philips · [4] University of Bonn

GECON 2026, 21st International Conference on the Economics of Grids, Clouds, Systems, and Services

Trustworthy and efficient AI for industrial AIoT systems

Industrial challenge

IoT systems generate large, distributed data streams. PANDORA develops methods that make AI pipelines more reliable, efficient and suitable for deployment across the edge–cloud continuum.

Two project goals

a) Prepare and deliver trustworthy datasets.
b) Improve the accuracy and sustainability of AI models operating in industrial AIoT environments.

Technical focus

Data-centric trustworthiness, uncertainty and explainability, efficient edge–cloud execution, federated and continual learning, and synthetic data generation.

PANDORA develops common methods and validates them through five industrial pilots.

TRUST IN PANDORA

Trust connects data quality, model behaviour and system operation

Data-centric trustworthiness

- Datasets are complete and consistent, with traceable provenance across the data lifecycle.

Transparent AI

- Explanations and uncertainty estimates expose why a model reached a decision and when confidence is limited.

Trust must hold across the complete AI pipeline

1

Efficient execution

AI pipelines manage energy, latency and scalability across the edge–cloud continuum.

2

Resilience and robustness

Models handle noise, incomplete data and heterogeneous industrial environments.

3

Operational trust

Data quality, model behaviour and system operation remain dependable together.

Two phases

Phase 1: Dataset preparation

1 Synthetic data generation
2 Uncertainty quantification
3 Explainable, domain-informed modelling

4 Semi-automated annotations
5 Data summarization

Phase 2: learning and inference strategies

7 Robust & resilient learning
8 Distributed and federated representation learning.

9 Continuous, explainable, domain-informed AI
10 Continual deep learning

11 Efficient inference. Split deep-learning inference across the computing continuum.

Five use cases

1

SIEMENS | Distributed AI for predictive edge applications
Support local adaptation while reducing data transfer and response time.

2

ERICSSON | Network monitoring and anomaly detection in a 5G-enabled smart building.
Edge intelligence identifies abnormal network behaviour and supports timely intervention.

3

PHILIPS | Visual quality inspection
Synthetic defect generation, anomaly detection and uncertainty-aware classification under limited defect data.

4

FRAMATOME | Critical energy infrastructure
Trustworthy AI for monitoring and decisions in safety-critical settings.

5

HALCOR | Energy efficiency in production
AI analysis for efficient production under deployment constraints.

Contents

Main obstacles to automating visual quality inspection

Defects are scarce

- Production lines are optimized to make good parts.
- Real defective samples are rare and expensive to collect, starving learning-based inspectors of training data.

Decisions lack confidence

- Hard labels with no uncertainty estimate carry asymmetric costs: a false reject wastes a good product; a false accept lets a defect through.

Trust gates automation

- Economic value depends on trusting routine decisions enough to automate them; while reserving scarce human expertise for genuinely ambiguous cases.

***Case study**: print-defect inspection on Philips electric-shaver chassis, across four defect categories.*

Combining the strengths of two paradigms

Supervised

- Learns defect appearance from labeled examples
- Can detect AND classify defect type
- Needs labeled defect data — costly & scarce

Unsupervised

- Trains only on defect-free samples
- Flags deviations from normality
- Detects, but does not categorize defects

Our contributions

1

A staged inspection pipeline: ROI segmentation → unsupervised anomaly detection → uncertainty-aware defect classification.

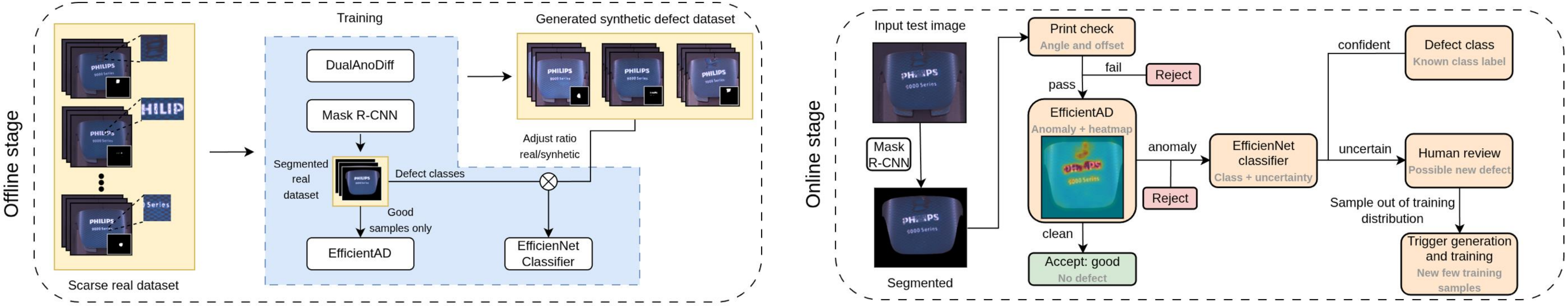
2

An empirical study of diffusion-based synthetic defect generation, evaluated on classification and localization of real defects.

3

Uncertainty-aware classification via Bayesian neural networks and deep ensembles, assessing performance and calibration under data scarcity.

Proposed staged inspection pipeline



Offline: DualAnoDiff generates synthetic defects to train the EfficientNet classifier; EfficientAD trains on defect-free samples only.
Online: Each unit passes a print check, EfficientAD anomaly detection, and EfficientNet classification — uncertain cases are deferred to human review.

Batzner, K. et al.: EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Com-puter Vision (WACV). pp. 128–138 (2024)
Jin, Y. et al.: Dualanodiff: Dual-interrelated diffusion model for few-shot anomalyimage generation. arXiv:2408.13509 (2024)

Synthetic defect generation with DualAnoDiff

Dual-branch diffusion model

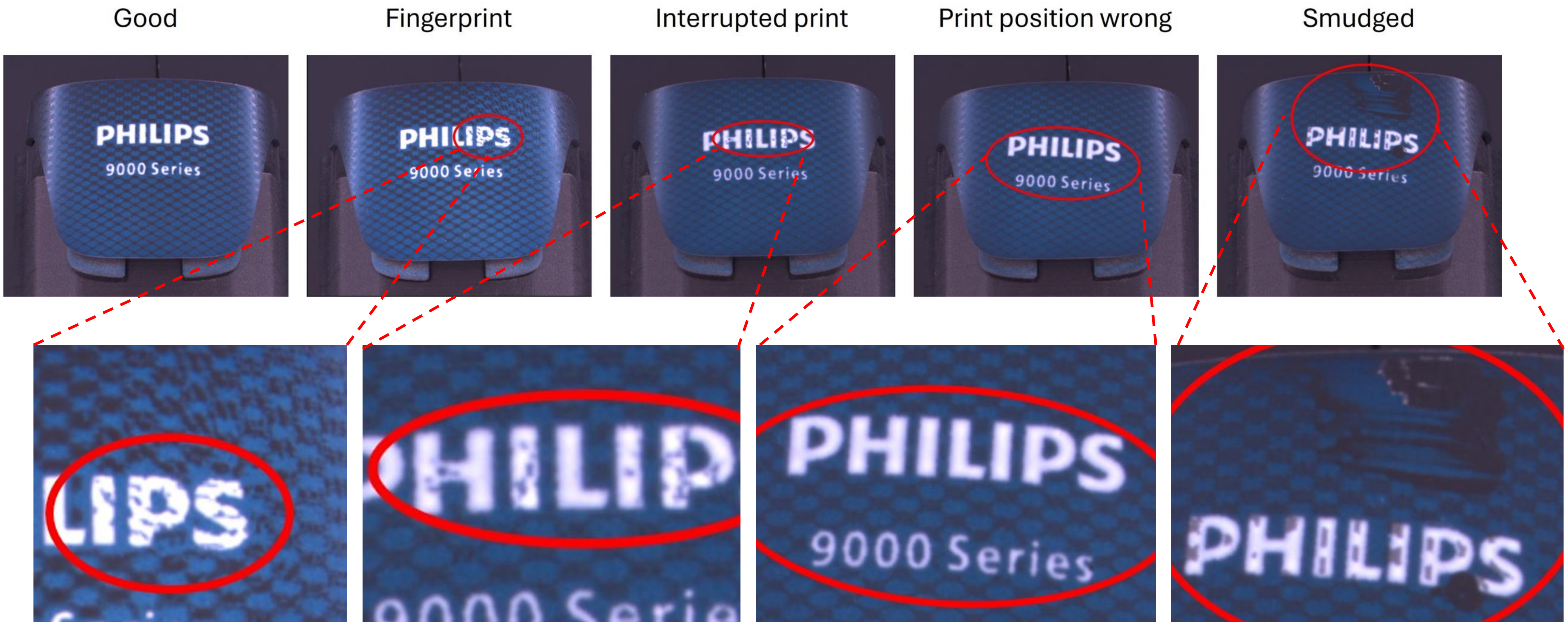
- Built on Stable Diffusion v1.5.
- A global branch renders the full anomalous object
- A foreground branch renders the isolated defect region with shared attention keeping the image and mask consistent.

50 real seed images (10 fingerprint · 30 interrupted print · 10 smudged)

500 synthetic images + masks generated per defect category

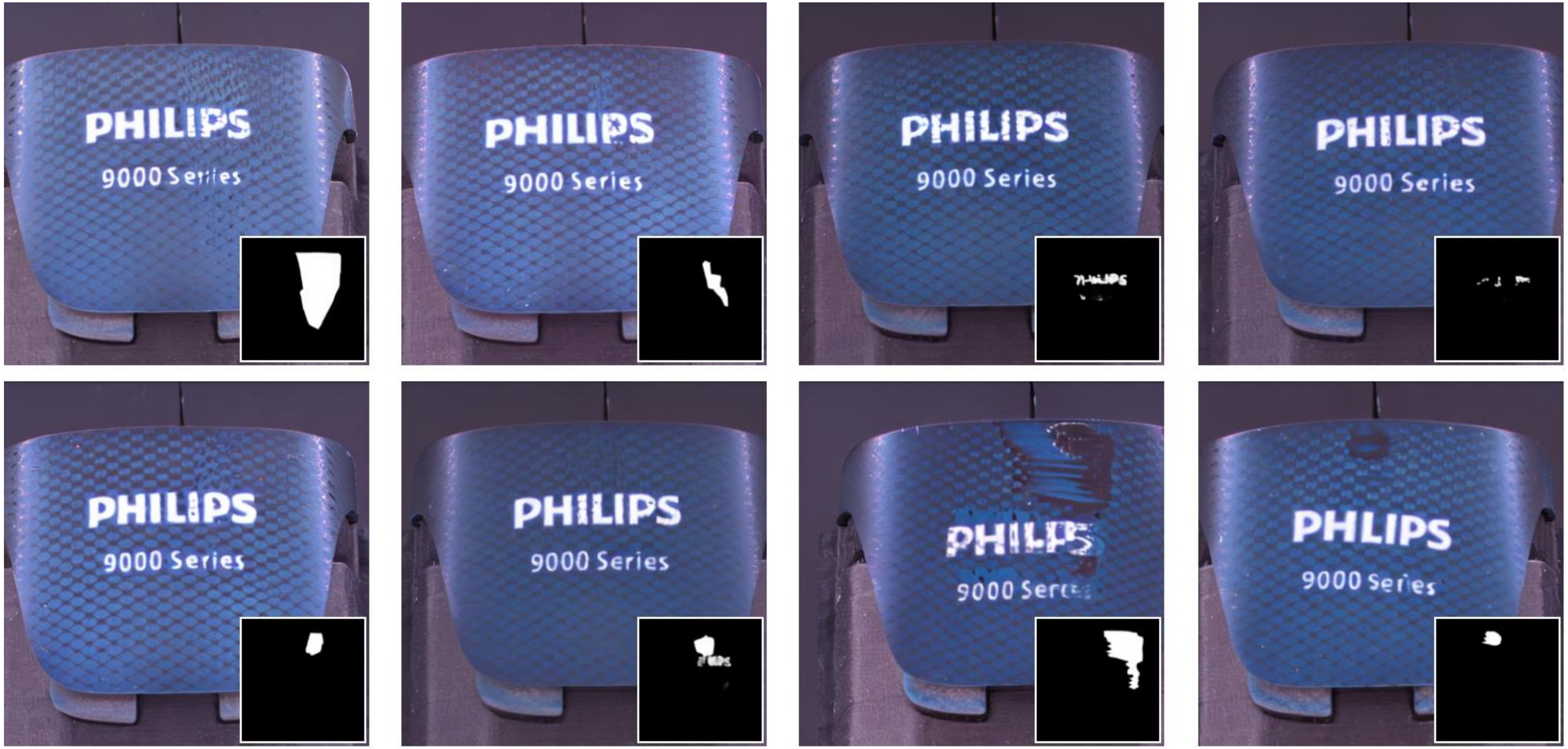
Synthetic defect generation with DualAnoDiff

Real examples of defect classes



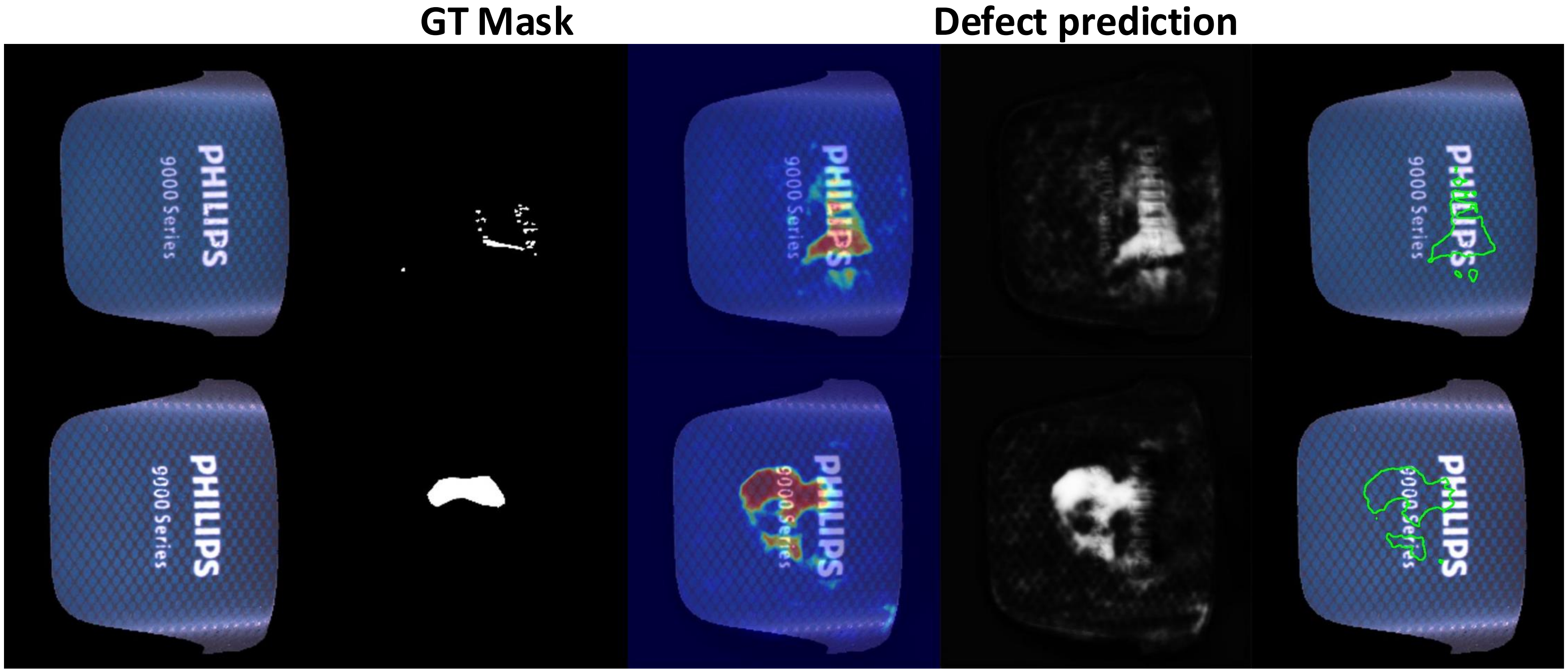
Synthetic defect generation with DualAnoDiff

Generated samples



Synthetic defect generation with DualAnoDiff

- Train DualAnoDiff with limited training samples (5-8 per category)
- Generate synthetic images
- Train detection/segmentation module using real good samples and the synthetic defects



Synthetic defect generation with DualAnoDiff

Classification

Relative improvement over real-only baseline (%)

Segmentation

Cell	AUROC-I	AP-I	f1_max-I	Acc-I	AUROC-P	AP-P	f1_max-P	PRO-P
Synthetic only (syn=500)	+19.2%	+4.7%	+4.4%	+8.2%	−5.9%	−18.4%	−12.1%	−16.1%
Real 25% + syn=500	+16.6%	+4.1%	+3.2%	+5.9%	−4.1%	−8.9%	−3.7%	−13.2%
Real 50% + syn=500	+18.3%	+4.4%	+5.1%	+9.6%	−4.4%	−12.1%	−5.9%	−7.8%
Real 75% + syn=500	+19.1%	+4.7%	+4.9%	+9.0%	−4.3%	−5.6%	−2.1%	−9.1%
Real 100% + syn=500	**+20.0%**	+4.8%	+5.3%	+9.9%	−4.8%	−5.9%	−1.6%	−6.9%
Real 100% + syn=125	+17.9%	+4.4%	+4.6%	+8.5%	−0.5%	**+6.4%**	**+5.8%**	−1.6%
Real 100% + syn=250	+18.5%	+4.6%	+4.7%	+8.8%	−1.8%	−3.9%	−0.3%	−3.0%

Unsupervised anomaly detection with EfficientAD

Mask R-CNN segmentation → **EfficientAD student–teacher + autoencoder** → Anomaly score + pixel heatmap

Why EfficientAD

- Trains exclusively on defect-free samples — sidesteps defect scarcity entirely.
- Millisecond-level latency, suited to in-line production deployment.
- Detection stays agnostic to the defect taxonomy: it decides whether an anomaly is present and localizes it; classification of defect type is handled downstream.

Clean → accepted as **good**
Anomaly → passed to classifier for defect type + uncertainty

ONLINE STAGE: STEP 3

Uncertainty-aware defect classification

Both approaches use EfficientNet-B3, fine-tuned on defect dataset, and decompose uncertainty into aleatoric (data noise) and epistemic (model knowledge gaps) components.

Deep Ensemble

- 5 independently trained EfficientNet-B3 models, distinct random seeds
- All ImageNet pre-trained, fine-tuned on defect dataset
- Uncertainty from disagreement across ensemble members
- Lower latency, higher throughput

Bayesian (Flipout) CNN

- Early MBConv blocks (1–5) kept deterministic for stable features
- Final stages use Flipout layers instead of standard Conv2d
- Pseudo-independent weight perturbations during forward passes
- Better calibration, higher compute cost

High epistemic uncertainty → likely out-of-distribution / unseen defect → routed to human review.

Classification performance (500 real+synthetic images/class)

Binary task (fingerprint vs. smudged) – 8 test images per class

- Both Deep Ensemble and Bayesian classifiers achieve 1.0 accuracy across every training-set size tested.
- Interpret with caution — only 16 real test images.

Three-class task (fingerprint · smudged · interrupted print – 24 test images)

Metric (500 img/class)	Deep Ensemble	Bayesian
Accuracy	0.750	0.833
Precision	0.857	0.853
Recall	0.750	0.833
F1-score	0.709	0.830
AUROC	0.930	0.977

Bayesian classifier overtakes Deep Ensemble as training data grows; better accuracy, F1, and AUROC at 500 images/class, though at added inference cost.

Calibration and uncertainty (3-class, 500 img/class)

Metric	Deep Ensemble	Bayesian
Brier score ($\times 10^3$)	110.0	94.5
ECE ($\times 10^2$)	19.7	24.5
Epistemic unc. ($\times 10^2$)	2.75	0.161
Aleatoric unc. ($\times 10$)	3.93	2.25

What this enables

- A human-in-the-loop pipeline that stays robust to novel defect classes
- Only genuinely ambiguous or out-of-distribution units are reserved for human review.

Key takeaways

- Bayesian classifier is far better calibrated on well-represented classes.
- Epistemic uncertainty is a usable inspection signal.
- Aleatoric uncertainty rises again at 500 synthetic images/class in the 3-class task — more borderline, overlapping examples appear as classes grow.

Can the uncertainty fail ?

A classifier trained only on defect classes A and B is forced to map every input to those classes.
A novel defect C may still receive a confident prediction such as:

$$P(A \mid x_C) = 0.97, \qquad P(B \mid x_C) = 0.03$$

- If all ensemble members or stochastic forward passes make approximately the same prediction, the epistemic variance can remain low.
- This means: the model is consistent in its decision, but not necessarily correct about whether the input belongs to the training distribution.

low model disagreement ≠ in-distribution sample

Solution: Keep uncertainty estimation, but add a separate OOD score

Feature-space Mahalanobis distance
For each known defect class k, estimate the feature mean μ_k and covariance Σ_k from the training set. For a new sample, compute its minimum class-conditional Mahalanobis distance:

$$D(x) = \min_k (z - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (z - \mu_k)$$

Interpretation: a large value indicates that the feature representation is far from all known defect classes.
Mark the sample as unknown when: D(x) > T ⇒ unknown / OOD defect

RESULTS

Label – Double Print	**Label - Fingerprint**	**Label - Double Print**	**Label – Interrupted Print**
Pred Fingerprint	Pred Fingerprint	Pred Smudged	Pred Interrupted Print
Epistemic : 0.000028 Aleatoric : 1.097747 Total : 1.097775	Epistemic : 0.000022 Aleatoric : 1.098371 Total : 1.098393	Epistemic : 0.000027 Aleatoric : 1.098569 Total : 1.098596	Epistemic : 0.000028 Aleatoric : 1.098297 Total : 1.098325
Mahalanobis Distance : 1895.4497	Mahalanobis Distance : 654.1458	Mahalanobis Distance : 843.3183	Mahalanobis Distance : 612.9233
Unseen class: True	Unseen class: False	Unseen class: True	Unseen class: False

Pipeline latency and throughput

Stage	Throughput (img/s)	Latency (s)
Segmentation	5.26	0.191
Print check	25.96	0.039
Anomaly detection (EfficientAD)	3.99	0.250
Classification — Deep Ensemble	15.90	0.063
Classification — Bayesian	1.97	0.507
Total pipeline — Deep Ensemble	**1.88**	**0.532**
Total pipeline — Bayesian	**1.00**	**0.999**

< 1 second

end-to-end latency for the full pipeline — even with the costlier Bayesian classifier — well within range for production-line deployment.

- *Deep Ensemble trades comprehensive uncertainty for speed*
- *The Bayesian model spends more time on Monte Carlo sampling in exchange for richer, better-calibrated uncertainty estimates.*

Toward deployable, trustworthy inspection

- **Data scarcity, addressed**

  Diffusion-generated synthetic defects relieve scarcity at near-zero marginal cost versus collecting rare real defects — and improve classifier accuracy and calibration.

- **Trust, built in**

  Pairing classification with uncertainty estimation lets the pipeline defer only genuinely uncertain decisions to human experts, preserving automation's labor savings.

- **Next: close the loop**

  Evaluate the full end-to-end online pipeline; route out-of-distribution units into few-shot generation and retraining for new defect types.

Supported by the Horizon Europe project PANDORA (Grant Agreement No 101135775).

Thank you.