Explainable Visual Anomaly Detection via Concept Bottleneck Models

Arianna Stropeni, Valentina Zaccaria, Francesco Borsatti, Davide Dalle Pezze , **Manuel Barusco** , and Gian Antonio Susto
University of Padova, Italy

CVPR JUNE 3-7, 2026 DENVER COLORADO

INTRODUCTION

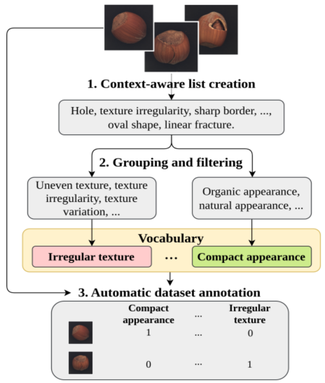
- Visual Anomaly Detection (VAD) identifies defects by training exclusively on normal images.
- While standard methods provide pixel-level heatmaps, these visual outputs lack human-understandable semantic descriptions.
- Concept Bottleneck Models (CBMs) offer a solution by learning intermediate, human-interpretable concepts to explain model predictions.
- These concepts facilitate transparency and human-machine collaboration through direct interventions on concept activations.

CONTRIBUTIONS

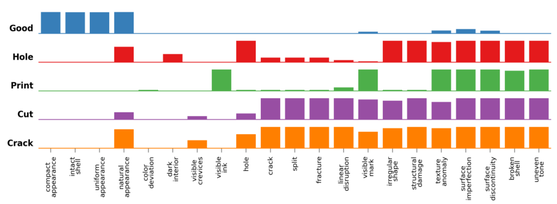
- Adaptation of CBMs to the VAD domain.
- Release of the first concept-annotated dataset for industrial visual anomaly detection.
- CONVAD Architecture: a novel dual-branch architecture that successfully combines concept-level semantic explanations with pixel-level anomaly localization.

FROM IMAGES TO CONCEPTS

Concept Pipeline: Utilizes a Vision Language Model (VLM) to extract context-aware concepts, filters redundant terms via CLIP embeddings and annotate the given dataset.



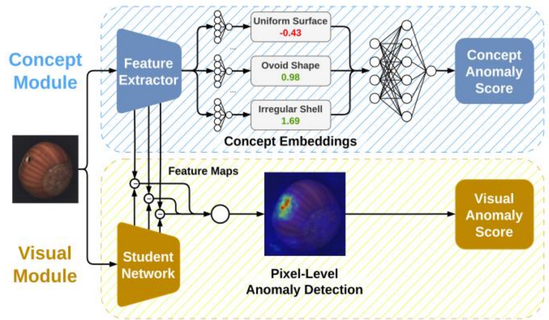
FROM IMAGES TO CONCEPTS



Example of concept vocabulary and distribution across defect types for the MVTec hazelnut category.

CONVAD ARCHITECTURE

- **Student-Teacher architecture:**
  - The teacher network is trained on concept predictions (Concept Module).
  - The student is a randomly initialized network that learns to match the teacher's features on normal samples (Visual Module).
- **Anomaly Heatmaps:** Anomalous regions are localized at the pixel level by computing the discrepancy between student and teacher feature maps.

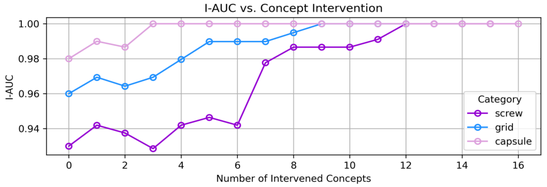


EVALUATION SETTINGS

- **Fully:** trained on both normal and anomalous images from the real-world dataset, with 80% of samples assigned to the training set.
- **Weakly:** the model is trained in a one-shot per defect type setting, considering a single anomalous image.
- **Weakly(3):** contrary to Weakly, it uses three real anomalous images for each defect type, thus considering a few-shot per defect type setting.
- **Synthetic Anomaly Generation (SAG)**: it assumes access only to the normal images from the dataset, while anomalous images are generated by a generative model.
- **Weakly(3)+SAG** and **Weakly+SAG** respectively correspond to the Weakly(3) and Weakly scenario augmented using SAG-generated samples.

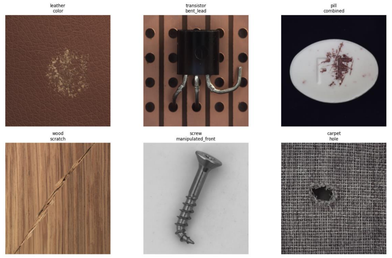
Category	STFPM	Fully		Weakly		Weakly+SAG		SAG	
	I-AUC	C-AUC	I-AUC	C-AUC	I-AUC	C-AUC	I-AUC	C-AUC	I-AUC
Bottle	1.00	0.99	1.00	0.95	1.00	0.79	0.97	0.80	0.94
Cable	0.91	0.87	1.00	0.69	0.79	0.67	0.76	0.80	0.88
Capsule	0.71	0.92	0.98	0.55	0.61	0.56	0.63	0.52	0.52
Carpet	0.96	0.72	1.00	0.61	0.97	0.81	0.90	0.75	0.72
Grid	0.77	0.96	0.81	0.55	0.41	0.71	0.88	0.72	0.64
Hazelnut	0.93	0.95	1.00	0.85	0.99	0.85	0.99	0.89	0.99
Leather	0.97	0.85	1.00	0.72	0.94	0.72	0.98	0.80	0.90
Metal Nut	0.92	0.92	1.00	0.73	0.82	0.70	0.84	0.59	0.66
Pill	0.81	0.81	0.97	0.63	0.76	0.63	0.75	0.62	0.60
Screw	0.55	0.82	0.93	0.59	0.53	0.55	0.70	0.48	0.49
Tile	0.99	0.9	1.00	0.76	0.94	0.84	0.96	0.76	0.86
Toothbrush	0.84	0.92	0.80	0.82	0.47	0.75	0.77	0.61	0.56
Transistor	0.96	0.55	0.85	0.35	0.73	0.59	0.73	0.66	0.66
Wood	0.99	0.81	1.00	0.71	0.90	0.80	0.98	0.82	0.93
Zipper	0.91	0.92	1.00	0.70	0.69	0.82	0.95	0.70	0.68
Average	0.88	0.86	0.97	0.68	0.77	0.72	0.85	0.73	0.77

Results obtained across the different training scenarios over the categories of MVTec-AD.

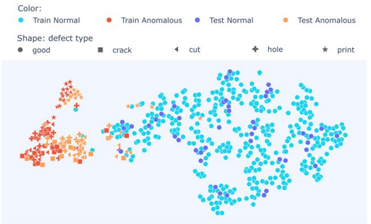


Performance gain in three MVTec categories for increasing number of intervened concepts. Fully supervised setting.

ANOMALOUS IMAGES



Examples of well-generated synthetic anomalous images.



Hazelnut t-SNE embeddings for the SAG scenario.

RESULTS

- **Visual Localization Performance:** The Visual Branch achieves highly competitive performance, recording an Image-level AUROC (I-AUC) of 0.96 and a Pixel-level AUROC (P-AUC) of 0.97 in the fully supervised setting. This matches PatchCore while uniquely adding concept-level explanations.
- **Synthetic Augmentation:** While fully supervised models perform best, augmenting a few real anomalies with synthetic data (Weakly+SAG) provides a highly effective compromise, significantly reducing annotation effort while maintaining robust detection.
- **Human-in-the-Loop Intervention:** Concept-level intervention allows human operators to manually correct a few mispredicted concepts. Fixing even a small subset of concepts drastically improves overall anomaly detection accuracy.

Contact

Manuel Barusco, manuel.barusco@phd.unipd.it, University of Padova
Valentina Zaccaria, valentia.zaccaria.1@phd.unipd.it, University of Padova
Arianna Stropeni, arianna.stropeni@studenti.unipd.it. University of Padova
Franesco Borsatti, franecsco.borsatti.1@phd.unipd.it , University of Padova
Davide Dalle Pezze, davide.dallepezze@unipd.it, University of Padova
Gian Antonio Susto, gianantonio.susto@unipd.it, University of Padova

Paper

MoViAD Library