Explainable Visual Anomaly Detection via Concept Bottleneck Models

Arianna Stropeni
University of Padova
Padova, Italy
arianna.stropeni@studenti.unipd.it

Valentina Zaccaria
University of Padova
Padova, Italy
valentina.zaccaria@phd.unipd.it

Francesco Borsatti
University of Padova
Padova, Italy
francesco.borsatti.1@phd.unipd.it

Davide Dalle Pezze
University of Padova
Padova, Italy
davide.dallepezze@unipd.it

Manuel Barusco
University of Padova
Padova, Italy
manuel.barusco@phd.unipd.it

Gian Antonio Susto
University of Padova
Padova, Italy
gianantonio.susto@unipd.it

Abstract

In recent years, Visual Anomaly Detection (VAD) has gained attention for its ability to detect defects using only normal images during training. While many methods provide visual explanations by highlighting anomalous regions, these lack clear semantic interpretation. To address this, we extend Concept Bottleneck Models (CBMs) to the VAD setting, enabling the model to learn meaningful concepts and generate human-interpretable descriptions of anomalies.

Our contributions are threefold: (i) we introduce a concept-based framework for anomaly explanation by adapting CBMs to VAD, and release the first concept-annotated dataset for industrial anomaly detection benchmark; (ii) we evaluate multiple supervision regimes, from fully supervised to synthetic-only settings, analyzing the trade-off between performance and labeling effort; and (iii) we propose a dual-branch architecture that combines concept-level explanations with pixel-level localization. Evaluated on three industrial benchmarks, our method, Concept-Aware Visual Anomaly Detection (CONVAD), achieves competitive detection performance while providing richer, concept-driven explanations that enhance interpretability and trust.

1. Introduction

Visual Anomaly Detection (VAD) aims to detect anomalies in images and localize them at the pixel level, with applications in domains such as manufacturing, medicine, and surveillance [2, 4, 5]. While existing VAD models provide anomaly segmentation masks, these visual outputs alone offer limited interpretability, as they lack human-understandable descriptions of the detected defects.

To address this limitation, we propose a Concept-Aware VAD paradigm based on Concept Bottleneck Models (CBMs) [6]. Unlike standard approaches, CBMs leverage supervised concept annotations to learn intermediate, human-interpretable representations that explain model predictions. This enables concept-level explanation, facilitating improved transparency and human-machine collaboration through possible user interventions on concept activations. However, standard CBMs lack pixel-level localization capabilities, which are essential in VAD to identify where defects occur. To bridge this gap, we propose CONVAD, a dual-branch architecture that combines concept-level explanations with pixel-level anomaly localization.
Although traditional VAD methods rely solely on normal samples, recent works show that incorporating limited supervision can significantly enhance performance [9, 14]. Motivated by this, we explore multiple supervision regimes, including settings with few real anomalies and purely synthetic defects, to analyze the trade-off between annotation effort, performance, and interpretability.
Our contributions can be summarized as follows: (i) we **adapt CBMs to the VAD setting** and release concept-annotated datasets for industrial benchmarks, created through a VLM-based pipeline that follows well-established approaches in CBMs literature [7, 13, 19];

(ii) we **systematically evaluate different supervision regimes**, including synthetic anomaly generation, providing a thorough analysis of the performance-labeling effort trade-off and allowing us to progressively move closer to the standard VAD paradigm; (iii) we **propose a dual-branch architecture** that combines concept-level explanations with pixel-level localization.

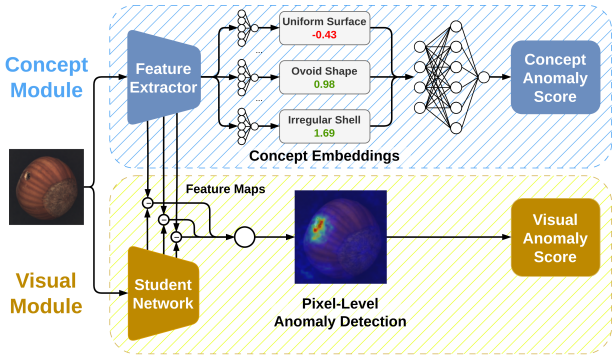


Figure 1. CONVAD Architecture with the i) CBM Module and the ii) Vision Module.

2. Methodology

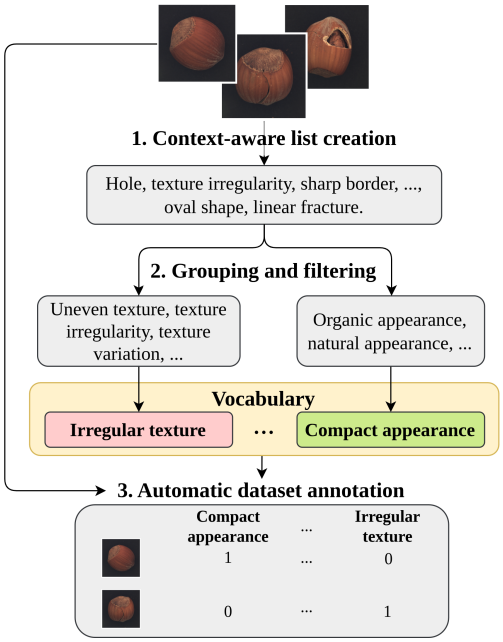


Figure 2. Pipeline for creating the Concept Dataset through concept annotation of a Vision Language Model (VLM).

2.1. Concept Dataset Pipeline

A key limitation of CBMs is the lack of concept-annotated datasets, particularly in domains such as VAD. Prior work addresses this using automated pipelines based on VLMs [7, 13, 19], but these approaches target natural image classification and do not transfer well to industrial anomaly detection, where concepts are domain-specific, fine-grained, and anomalies are underrepresented in pretraining data.

To overcome this, we design an anomaly-aware concept extraction pipeline with tailored prompts and validate it against manual annotations. Our method adapts existing approaches while introducing domain-specific refinements for VAD. The pipeline, as shown in Figure 2, consists of three steps:

(1) Context-aware concept extraction. We sample a small subset of images from each object category and use a VLM with a category- and defect-aware prompt to first generate visual descriptions and then extract a set of candidate concepts from it, reducing hallucinations and ensuring grounding on visual concepts.

(2) Grouping and filtering. Morphologically and semantically related concepts are merged through a VLM using few-shot examples to remove redundancy, followed by similarity-based filtering in the CLIP embedding space [8] to discard near-duplicate concepts.

(3) Automatic annotation. The final concept vocabulary is used to annotate all images with binary concept labels via VLM querying, using a concept-aware prompt similar to the first step.

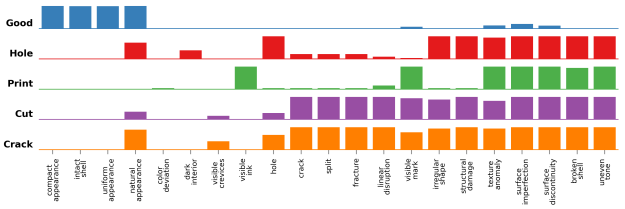


Figure 3. Example of concept vocabulary and distribution across defect types for the MVTec hazelnut category.

2.2. CBM Adaptation for Visual Explanations

Modern VAD approaches offer pixel-level anomaly localization, whereas standard CBMs lack this capability. We extend CBMs to provide visual localization of anomalies alongside concept-based explanations (a schematic illustration is shown in Fig. 1). The proposed architecture adopts a student–teacher paradigm [17], consisting of two identical feature extraction networks. The teacher network is pre-trained on ImageNet and fine-tuned using the CBM and downstream task objectives, while the student network is randomly initialized and trained to match the teacher feature maps only on normal samples. At inference, both networks process the input image, and an anomaly heatmap is computed from the discrepancy between their feature maps, as anomalous regions lie outside the student training distribu-

tion. This approach pairs the textual explanations produced by the CBM with a spatial visualization of anomalous regions.

2.3. Synthetic Anomaly Generation

In this study, synthetic anomalies are generated by editing normal images using a text-to-image and image-editing system. Let $\mathcal{G}_{\text{edit}}(\mathbf{x}, p)$ denote the editing function, where $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ is the input image and p a textual prompt. Given a normal image $\mathbf{x}_n$, an anomalous sample is obtained as $\mathbf{x} * a = \mathcal{G} * \text{edit}(\mathbf{x}_n, p_a)$, where p_a specifies the defect while preserving the original image context.

Prompts are designed to control anomaly generation and include: (i) defect type, (ii) object type, and (iii) pose and view details, ensuring consistency in geometry, viewpoint, and lighting.

3. Experimental Setting

3.1. CBM Scenarios

We provide the results for several scenarios differentiated by the data provided to the CBM model:
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.

3.2. Metrics

VAD algorithms can be assessed using multiple criteria, corresponding to different levels of analysis:
Image-level: To evaluate the model ability to detect abnormalities at the image level (each image is a different sample), we report the AUROC (Area under the Receiver Operating Characteristic curve) score, referred to as I-AUC.
Pixel-level: To assess performance at the pixel level, we consider the commonly used AUROC score, named P-AUC, computed by considering each pixel of each image as an independent sample.
Concept-level: For CBMs, we evaluate the accuracy of concept predictions using average AUROC score, referred to as C-AUC. Since concepts are binary, we compute the AUROC score for each concept separately and then combine them by averaging over all concepts.

4. Results

4.1. CBM Module

A comparison of the various CBM scenarios with detailed per-category results for MVTec-AD is available in Table 1.
The **Fully** supervised setting achieves strong anomaly detection and concept prediction performance, with notable gains on categories characterized by small and localized defects (e.g., capsule, grid, screw) with respect to unsupervised alternatives such as STFPM, highlighting the benefit of full supervision.
In the **Weakly** scenario, performance varies across categories: some remain robust, while the harder ones degrade significantly. Augmenting limited real anomalies with synthetic data (**Weakly+SAG**) consistently improves results, particularly for challenging categories with subtle defects. This suggests that synthetic anomalies, despite being imperfect, provide useful variability that enhances discrimination.
When relying solely on synthetic data (**SAG**), performance is strongly category-dependent: in some cases it approaches fully supervised results, while in others it suffers due to distribution mismatch with real anomalies.
Overall, results show a clear trade-off between supervision and performance: full supervision yields the best results, while combining few real anomalies with synthetic data offers an effective compromise, significantly reducing annotation effort.

4.2. Visual Module

Table 2 displays image- and pixel-level performance, highlighting the contribution of the Visual Branch for anomaly localization, and provides a comparison with unsupervised methods for VAD. Note that, to the best of our knowledge, no existing VAD methods provide explicit concept-level predictions, and related interpretability approaches rely on large VLMs at inference time, thus having a scale that makes a direct and practical comparison infeasible.
Our visual branch improves over Student-Teacher Feature Pyramid Matching (STFPM) by fine-tuning the teacher network on the concept prediction task, injecting domain knowledge, and yielding more informative features. Performance is comparable to PatchCore [10], while additionally providing concept-level explanations through the CBM branch. Importantly, the visual branch enhances overall robustness: it offers reliable pixel-level cues when the CBM branch fails and, following the student–teacher paradigm, improves generalization to previously unseen anomalies.

4.3. Intervention

We evaluate the impact of concept-level interventions to assess how manual correction of a few predicted concepts affects model performance (more details about the inter-

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

Table 1. Results obtained across the different training scenarios over the categories of MVTec-AD. Note that C-AUC (Concept AUC) is not reported for STFPM as it does not provide concept-based explanations.

Condition	Branch	I-AUC	P-AUC
CBM Fully	CBM	**0.97**	–
	Visual	0.96	**0.97**
PatchCore	–	0.96	0.95
STFPM	–	0.88	0.95

Table 2. Comparison of image-level (I-AUC) and pixel-level (P-AUC) performance. The branch column indicates concept-level or visual-level operation.

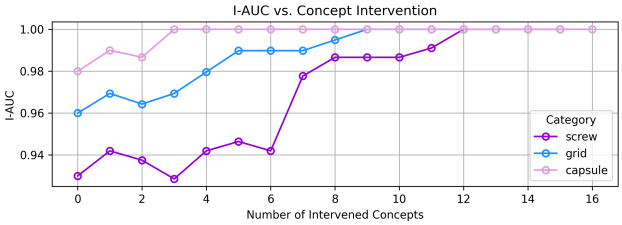


Figure 4. Performance gain in three MVTec categories for increasing number of intervened concepts. Fully supervised setting.

vention procedure in the Supplementary Material). Results highlighted in Fig. 4 show that even fixing a small subset of concepts leads to significant improvements in anomaly detection accuracy, confirming the effectiveness of the Concept Bottleneck structure in facilitating human-guided corrections. The obtained results prove the usefulness of CBM in the VAD setting to improve the human-machine collaboration and also high gains in terms of final performance.

5. Conclusion

We introduced CONVAD, a framework that extends CBMs to VAD by combining concept-level reasoning with pixel-level localization. This dual-branch design enables interpretable predictions while maintaining competitive anomaly detection performance across multiple benchmarks. Moreover, concept-level interventions provide a practical mechanism for human-machine collaboration, improving reliability with minimal effort.

Our analysis of supervision regimes shows that synthetic anomalies can effectively complement limited real data, offering a favorable trade-off between performance and annotation cost, although fully synthetic training remains less effective.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019. 1

[2] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection, 2024. 1

[3] Manuel Barusco, Francesco Borsatti, Davide Dalle Pezze, Francesco Paissan, Elisabetta Farella, and Gian Antonio

Susto. Paste: Improving the efficiency of visual anomaly detection at the edge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4026–4035, 2025. 1

[4] 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)*, pages 9584–9592, 2019. 1

[5] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation, 2021. 1

[6] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 1

[7] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023. 1, 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 1

[9] Blaž Rolih, Matic Fučka, and Danijel Skočaj. No label left behind: a unified surface defect detection model for all supervision regimes. *Journal of Intelligent Manufacturing*, pages 1–21, 2025. 1

[10] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2022. 3

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[12] Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 2023. 3

[13] Divyansh Srivastava, Ge Yan, and Lily Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37:79057–79094, 2024. 1, 2

[14] Han Sun, Yunkang Cao, Hao Dong, and Olga Fink. Unseen visual anomaly generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25508–25517, 2025. 1

[15] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 1

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. 2

[17] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection, 2021. 2

[18] Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*, 2023. 1

[19] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19187–19197, 2023. 1, 2

[20] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022. 1

Explainable Visual Anomaly Detection via Concept Bottleneck Models

Supplementary Material

5.1. Implementation Details

Models In the Concept Dataset Annotation Pipeline, we mainly leverage Gemma3 [15] as a VLM, except in the Concept Filtering phase, where we use the CLIP ViT-B/32 text encoder [8]. For the text-to-image and image-editing system in the synthetic anomaly generation pipeline, we use the capabilities of Nano Banana, Google's latest image generation model. We train the CBM using MobileNet-v2 as a feature extractor [11] to guarantee efficiency and make our pipeline suitable for resource-constrained settings. To ensure consistency, we report the comparison with PatchCore and STFPM using MobileNet-v2, as reported in a recent study for VAD on edge devices [3].

Dataset Our main experimental results are obtained using the MVTecdataset. Further results on additional datasets, including Visa, can be found in the Supplementary Material.

Hyperparameter Optimization To select the best hyperparameter configuration, we employ Bayesian optimization, implemented through the Optuna library [1] using the Tree-structured Parzen Estimator (TPE) algorithm [18] (see the Supplementary Material for more details).

Training Details During training, the CBM model is optimized using the Joint Training paradigm. This approach was selected since it consistently outperforms the other paradigms (ablation study in the Supplementary Material). CBM results are reported as the average performance over three runs with different random seeds. Additional details about training and data augmentation are described in the Supplementary Material.

6. Additional Experiments

6.1. Weakly(3) Scenario

In Table 3 we show the supervised scenario using only 3 real anomalous examples per defect type, compared with the same setting, but augmented with SAG data. We can draw similar considerations as those reported in Section 3.1 for the data augmentation effect on the weakly supervised setting.

Category	Weakly(3)		Weakly(3)+SAG	
	C-AUC	I-AUC	C-AUC	I-AUC
Bottle	0.99	1.00	0.78	0.99
Cable	0.82	0.92	0.78	0.91
Capsule	0.70	0.80	0.57	0.62
Carpet	0.60	0.97	0.82	0.96
Grid	0.57	0.58	0.74	0.90
Hazelnut	0.86	1.00	0.83	0.97
Leather	0.77	0.97	0.83	1.00
Metal Nut	0.80	0.70	0.76	0.89
Pill	0.67	0.75	0.72	0.84
Screw	0.62	0.65	0.60	0.81
Tile	0.90	0.99	0.95	0.97
Toothbrush	0.69	0.56	0.82	0.84
Transistor	0.60	0.77	0.64	0.91
Wood	0.89	1.00	0.84	0.99
Zipper	0.90	1.00	0.84	0.97
Average	0.76	0.84	0.76	0.90

Table 3. Results by category obtained in the Weakly(3) and Weakly(3)+SAG settings over the MVTec-AD categories.

6.2. VisA dataset

In Table 4 we report the experiments in the Fully Supervised setting, for all categories of the VisA dataset [20]. The original dataset paper reports PatchCore with a ResNet-50 backbone as the top performer. Despite having an order of magnitude more parameters, it achieves unsupervised performance comparable to CONVAD trained in the Fully Supervised setting with a MobileNet-v2 backbone, illustrating the trade-off between model efficiency and annotation cost.

Category	C-AUC	I-AUC
Candle	0.89	0.97
Capsules	0.66	0.77
Cashew	0.82	0.97
Chewing Gum	0.91	0.99
Fryum	0.79	1.00
Macaroni 1	0.62	0.96
Macaroni 2	0.71	0.95
PCB1	0.76	0.97
PCB2	0.69	0.81
PCB3	0.70	0.93
PCB4	0.72	1.00
Pipe Fryum	0.83	0.99
Average	0.76	0.94
Average (PatchCore)	-	0.93

Table 4. Results obtained on the **Visa Dataset** in the Fully Supervised setting.

6.3. Real-IAD dataset

In Table 5 we report the experiments in the Fully Supervised setting, for all categories of the Real-IAD dataset [16]. The I-AUC SimpleNet column reports the best-performing model results from the original dataset paper, which confirms the ability of CONVAD to outperform unsupervised methods that require larger model sizes.

Category	C-AUC	I-AUC	I-AUC SimpleNet
Audiojack	0.82	0.90	0.88
Bottle Cap	0.85	0.96	0.95
Button Battery	0.89	0.95	0.88
End Cap	0.88	0.92	0.81
Eraser	0.94	0.97	0.93
Fire Hood	0.78	0.87	0.95
Mint	0.68	0.85	0.69
Mounts	0.89	0.96	0.95
PCB	0.87	0.94	0.92
Phone Battery	0.80	0.86	0.93
Plastic Nut	0.92	0.94	0.86
Plastic Plug	0.81	0.89	0.94
Porcelain Doll	0.90	0.98	0.87
Regulator	0.87	0.89	0.98
Rolled Strip Base	0.95	1.00	1.00
Sim Card Set	0.90	1.00	0.99
Switch	0.90	0.96	0.97
Tape	0.93	0.98	0.99
Terminal Block	0.93	0.97	0.99
Toothbrush	0.90	0.98	0.91
Toy	0.93	0.96	0.91
Toy Brick	0.90	0.93	0.84
Transistor 1	0.88	0.99	0.98
U-Block	0.87	0.93	0.93
USB	0.81	0.95	0.97
USB Adaptor	0.80	0.94	0.87
VC Pill	0.96	1.00	0.92
Wooden Beads	0.92	0.93	0.85
Woodstick	0.92	0.96	0.86
Zipper	0.97	1.00	1.00
Average	0.88	0.94	0.92

Table 5. Results obtained on the **Real-IAD Dataset** in the Fully Supervised setting. We considered single-view Real-IAD, keeping only the view with the highest number of anomalies in each category.

7. Prompts Details

7.1. Concept Extraction Prompt

We construct the prompt by adding context cues that can guide the model in extracting meaningful concepts: we specify which object is present in the pictures, whether it is anomalous or not and, if so, which defect is present. Next, we employ a Chain-of-Thought prompting strategy and formulate two consecutive prompts:

1. *"Provide a description of the image that includes information about all the relevant features that are visible. Focus only on what can be seen, avoiding speculations or assumptions.*
2. *"Provided the following description, extract the five most meaningful concepts. Concepts should be defined in such a way that, observing the picture, it is possible to clearly answer with yes or no about its presence.*

It is worth mentioning that this step is applied to a small subset of the dataset, corresponding to 5% of it, so it requires the availability of only a few labeled anomalous images. When labeled anomalous samples are not available, the same procedure can be applied to synthetically generated pictures, keeping in mind that the quality of the extracted concepts is bounded by the quality of the generation process.

7.2. Dataset Annotation Prompt

A very similar prompt for dataset annotation as we did for concept extraction is employed: for each image, we include information about the object and, if present, the type of anomaly.

"I provide an image of a {category}. The image has been classified as {label}, [which implies that it shows a visible defect or anomaly, specifically {defect type}.] Knowing this, choose among the following list of concepts which ones can be clearly seen in the picture. Output the result as a JSON object of the following form: {Concept 1: True, Concept 2: False, ...}.

Through the previous prompt, we ensure to focus only on binary concepts.

7.3. Anomaly Generation Prompt

We generate anomalous images leveraging the capabilities of Google's Nano Banana. We assume we have a set of normal pictures and a list of defects that we want to include in the synthetic images, provided by a domain expert. Next, we query a Large Language Model (LLM), in our case GPT-5, to provide ten synonyms of the defect we want to add, to ensure richer variability, and build prompts according to this structure:

"Modify this image by adding a {type of defect} to the {name of the object}, keeping the same angle, view, and pose."

	All	Anomaly Concepts	Normal Concepts
Acc	0.93 ± 0.08	0.92 ± 0.08	0.98 ± 0.01
Prec	0.76 ± 0.24	0.71 ± 0.23	0.99 ± 0.00
Rec	0.77 ± 0.28	0.72 ± 0.28	0.99 ± 0.01

Table 6. Average value and standard deviation of accuracy, precision and recall of predicted concepts over the *hazelnut* category.

Model type	C-AUC	C-F1	I-AUC	I-F1
Joint	0.86	0.76	0.97	0.86
Sequential	0.85	0.75	0.95	0.85
Independent	0.87	0.72	0.96	0.85

Table 7. Comparison of the results of the three CBM learning paradigms over the MVTec-AD dataset, in terms of concept prediction performance and anomaly detection.

Some categories and defect types, which proved to be particularly challenging for the VLM, required manually tuning the previous prompt and discarding some of the generated images. Some examples of the synthetic pictures obtained can be found in Fig. 5.

8. Evaluation of the Concept Annotation Pipeline

To safely guarantee that the proposed pipeline for concept extraction and annotation can be used to produce high-quality results, we evaluate the automatically annotated dataset of the *hazelnut* category against a ground-truth version of it. Specifically, we compute accuracy, precision and recall of the predicted concepts. Table 6 displays a summary of our findings. Overall, predicted concepts can be considered of good quality, especially those that appear more often in normal images; as for anomalous concepts, we observe a very high variability in terms of precision and recall, which is mainly related to those concepts that describe a diffused feature of the image (e.g., *surface discontinuity, uneven tone, etc.*), which are more difficult to capture for a VLM and can be arbitrary for a domain expert as well.

9. Additional Training Details

Each model is trained for a maximum of 100 epochs, with an early stopping mechanism triggered after ten epochs. When performance reaches a plateau after five epochs, we also allow the learning rate to be reduced by a factor of 10.

Transformation-based Data Augmentation. We apply transformations to enhance the model generalization ability, carefully choosing those that preserve the presence of each concept, ensuring the integrity of attribute labels. The transformations applied are the following: horizontal and vertical flips with probability 0.5, random rotation by a 25-degree angle, brightness jitter by a factor of 0.2, and contrast jitter by a factor of 0.2.

CBM Training. We attach k parallel linear layers to the CNN feature extractor that act as concept predictors, while a simple Feed-Forward Neural Network with eight neurons is employed for anomaly detection. The feature extractor undergoes a pre-training phase directly on the dataset of interest, in which it is optimized according to a multi-class classification objective to learn which object is depicted in the image, while its last layers are fine-tuned during training for the concept prediction task.

10. CBM training Modalities

Table 7 displays a comparison of the results obtained by the three CBM paradigms on MVTec-AD. In consideration of these findings, the CBM model is trained using the Joint paradigm, unless stated otherwise in the paper, since it performs the best on three of the four evaluated metrics.

11. Intervention Procedure

In Section 4.3, we demonstrate the impact on performance of manually modifying predicted concepts with their ground-truth value during inference. However, this procedure can be costly, as it requires the supervision of a human expert, so several strategies have been devised to minimize such costs by focusing first on the concepts that should be more important to provide an increase in performance. We followed the UCP (Uncertain Concept Prediction) heuristic proposed in [12] and computed the entropy-based uncertainty related to each concept prediction:

$$\mathcal{H}(c_i) = -(p_i \cdot \log(p_i) + (1 - p_i) \cdot \log(1 - p_i)). \quad (1)$$

We then sorted the entropy scores in descending order, following the idea that concepts predicted with more uncertainty might confuse the model and lead to incorrect predictions. Since the joint model uses the predicted concept logits to perform the main task, we substitute them with either the 5th or the 95th percentile of the training distribution.

12. Concept Logits Analysis

In the SAG column of Table 1, where the model is trained exclusively on generated anomalies and evaluated on real ones, some categories achieve high concept-level AUC (C-AUC) and image-level AUC (I-AUC), while others fail to transfer effectively to real data. Two extreme examples illustrating this contrast are hazelnut, which transfers well, and metal nut, which does not. We examine the concept-logit embeddings extracted from the training and test examples to investigate this discrepancy. The test set contains

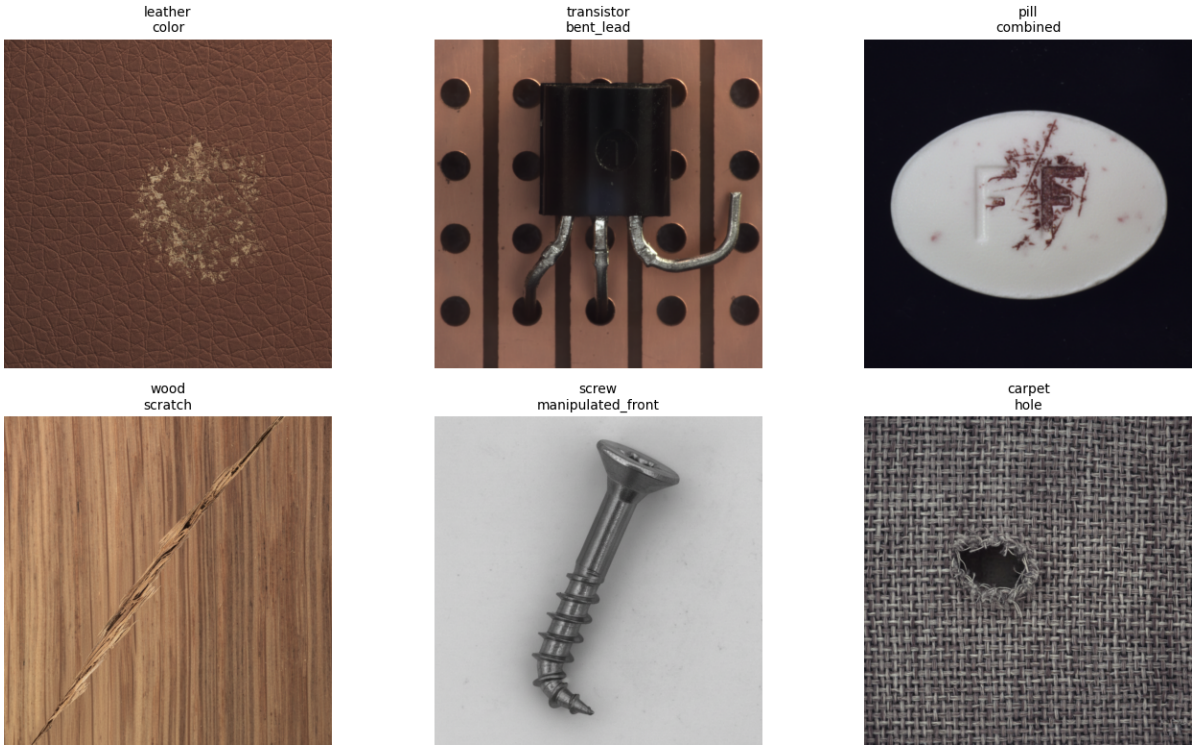


Figure 5. Examples of well-generated synthetic anomalous images.

only real normal and real anomalous examples, while the training set may include synthetic images in the SAG scenarios. All models compared within each category use the same random seed, training data, and evaluation set, with the latter containing only real anomalies that were not used for weak supervision in any of the models.

A well-learned bottleneck should ensure (i) a separation between normal and anomalous samples, provided that the concept set is sufficiently predictive, and (ii) tight clustering of images sharing the same or similar defect type (e.g., ”hole” or ”crack” in hazelnut). Moreover, if the synthetic anomalies are sufficiently well aligned with the real anomaly domain, their concept-logit representations should be consistent with those of real anomalies, indicating an effective transfer. To obtain qualitative insights about these aspects, we plot 2-dimensional t-SNE visualizations of the concept logits.

In the hazelnut SAG scenario, we obtain an average C-AUC of 0.89 and an I-AUC of 0.99. Figure 6, shows a clear separation between `Train Normal` samples `Train Anomalous` (synthetic). `Test Normal` samples appear in the same region as `Train Normal`, which is expected since their domains match, as both correspond to real images. In addition, `Test Anomalous` (real images) are mapped to the same broad regions as synthetic anomalies (`Train Anomalous`), which is ideal. Within this anomalous region, the ”print” defect forms a

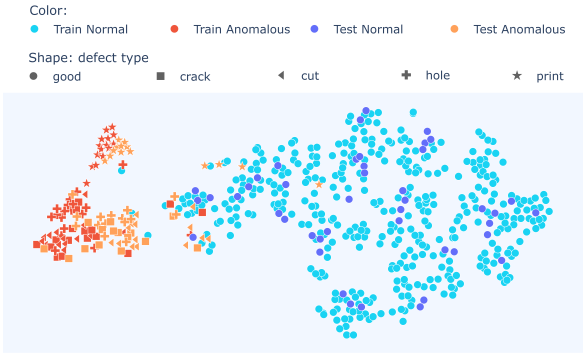
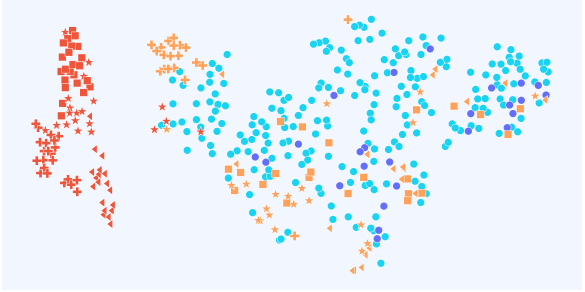


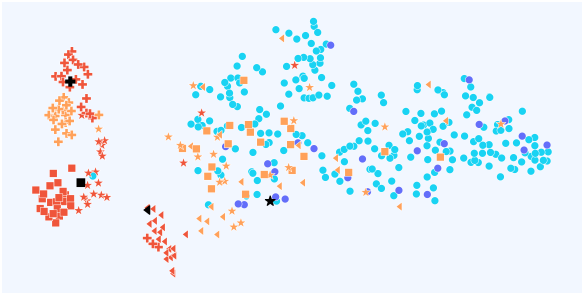
Figure 6. Hazelnut t-SNE embeddings for the SAG scenario.

distinct sub-cluster, while ”crack”, ”cut” and ”hole” share more overlapping embeddings, reflecting their visual similarity. These patterns are consistent with the high AUC scores and indicate that the *synthetic anomalies generated for the hazelnut category effectively capture relevant characteristics of real anomalies*.

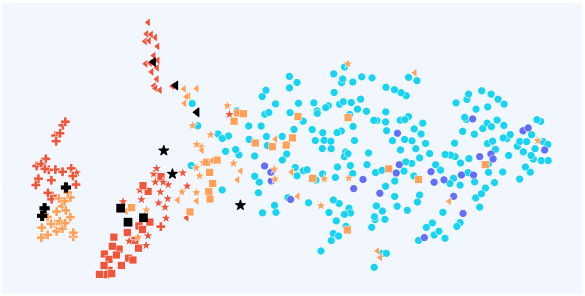
However, for the metal nut in the SAG scenario, the C-AUC drops to an average of 0.59, and the I-AUC drops to 0.66. In this case, in Figure 6(a), most `Test Anomalous` (real) samples are located within regions corresponding to `Train Normal` and `Test Normal` data, rather than being mapped near the synthetic `Train Anomalous` sam-



(a) SAG scenario



(b) Weakly supervised (1 sample per defect) + SAG scenario.



(c) Weakly supervised (3 samples per defect) + SAG scenario.

Figure 6. t-SNE representations of concept logits embeddings using `metal_nut`. Black markers highlight the few real samples used for Weakly+SAG scenarios.

ples. Only the ”flip” defect forms a slightly separated cluster (due to its visual dissimilarity to other images), while other defects are scattered among normal samples. This qualitative mismatch between synthetic and real anomaly embeddings explains the degraded AUCs and suggests poorly generated anomalies.

Exploring the embeddings in the Weakly+SAG scenarios provides a deeper understanding of why introducing even a small number of real anomalies, combined with synthetic anomalies, reduces this discrepancy in performance. For the metal nut in the Weakly+SAG scenario (adding one real anomaly per defect type in training), the C-AUC improves to an average of 0.70 and the I-AUC to 0.84, and Figure 6(b) shows a closer alignment between synthetic and real anomalies. With Weakly(3)+SAG (three real anomalies per type), the C-AUC increases to 0.76 and the I-AUC to 0.89. Simultaneously, the representations of real anomalies appear closer to those of synthetic ones, and clusters associated with different defect types become more clearly separated (Figure 6(c)). These results indicate that when using synthetic anomalies for training, even minimal real-data supervision can mitigate the domain shift and encourage the concept space to encode a more consistent defect-specific structure, eliminating the need to collect a large dataset of rare and difficult-to-acquire anomalies.