IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 May 2026, accepted 22 June 2026, date of publication 3 July 2026, date of current version 22 July 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3709944

RESEARCH ARTICLE

Improving Industrial Anomaly Clustering via Heatmap-Based Multi-Scale Views

OMAR DE MITRI[1,2], COSIMO DISTANTE[2,3], (Senior Member, IEEE), SILVIA ZOTTIN[4], AXEL DE NARDIN[4], GIAN LUCA FORESTI[3], (Senior Member, IEEE), AND MARCO F. HUBER[1,5], (Senior Member, IEEE)

[1]Research Unit Artificial Intelligence and Machine Vision, Fraunhofer IPA, 70569 Stuttgart, Germany
[2]Department of Innovation Engineering, University of Salento, 73100 Lecce, Italy
[3]Institute of Applied Sciences and Intelligent Systems—CNR, 73100 Lecce, Italy
[4]Department of Mathematics, Informatics and Physics, University of Udine, 33100 Udine, Italy
[5]Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, 70569 Stuttgart, Germany

Corresponding author: Omar De Mitri (omar.de.mitri@ipa.fraunhofer.de)

The work of Omar De Mitri and Marco F. Huber was supported by the Baden-Württemberg Ministry of Economic Affairs, Labour and Tourism (Künstliche Intelligenz (KI)-Fortschrittszentrum "Lernende Systeme und Kognitive Robotik"), under Grant 017-180036. The work of Silvia Zottin, Axel De Nardin, and Gian Luca Foresti was supported in part by the Friuli Venezia Giulia Region Project "Supporting the Diagnosis of Rare Diseases Through Artificial Intelligence" through the Project A under Grant CUP: F53C22001770002 and through the Project B under Grant CUP F53C22001780002, and in part by the MUR Italian Fund for Applied Sciences (FISA) "Predicter Artificial Intelligence for APIs (I-TERAP)" Project under Grant FISA-2023-00297 and Grant G23C25000440006.

ABSTRACT Anomaly detection in industrial contexts has become a critical task in computer vision due to its high relevance to industry applications. By grouping anomalies into coherent clusters based on their characteristics, which is commonly known as anomaly clustering, a better understanding of the anomalies and their causes is possible. We propose a novel anomaly clustering framework that leverages anomaly localization to create multi-scale views for contrastive learning. Unlike previous approaches, it integrates heatmap-guided multi-view feature extraction with self-supervised clustering, improving robustness and scalability in industrial defect analysis. Our approach utilizes spatial localization of anomalies by a state-of-the-art anomaly detector to generate multiple views of the image at different scales. These views are then used to train a convolutional neural network by means of contrastive learning, allowing the network to learn effective data representations. This technique enhances the network's ability to distinguish between the unique features of anomalies from different perspectives, improving representation learning for anomaly clustering. We demonstrate the effectiveness of our method on benchmark datasets, showing an improvement in NMI of 12.5% in clustering accuracy for objects and textures on the MVTec-AD dataset. Our approach sets a new state-of-the-art, advancing the capabilities of anomaly clustering in industrial applications.

INDEX TERMS Anomaly clustering, contrastive learning, industrial defect analysis, self-supervised learning.

I. INTRODUCTION

Anomaly detection techniques distinguish between normal (fault-free) and defective images using machine learning models trained only on normal images. Anomalies are identified as deviations from the normal distribution. This technique, often seen as a one-class classification, has become a subject of significant research interest in computer vision. Recently, several approaches based on teacher-student networks [1], [2], reconstruction [3], [4], or memory banks [5], [6] have been developed. Anomaly detection in industrial manufacturing is essential for automating defect identification and minimizing production failures. These systems, however, do not differentiate between different anomaly types, which is crucial for defect diagnosis and repair. This is where anomaly clustering becomes valuable.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

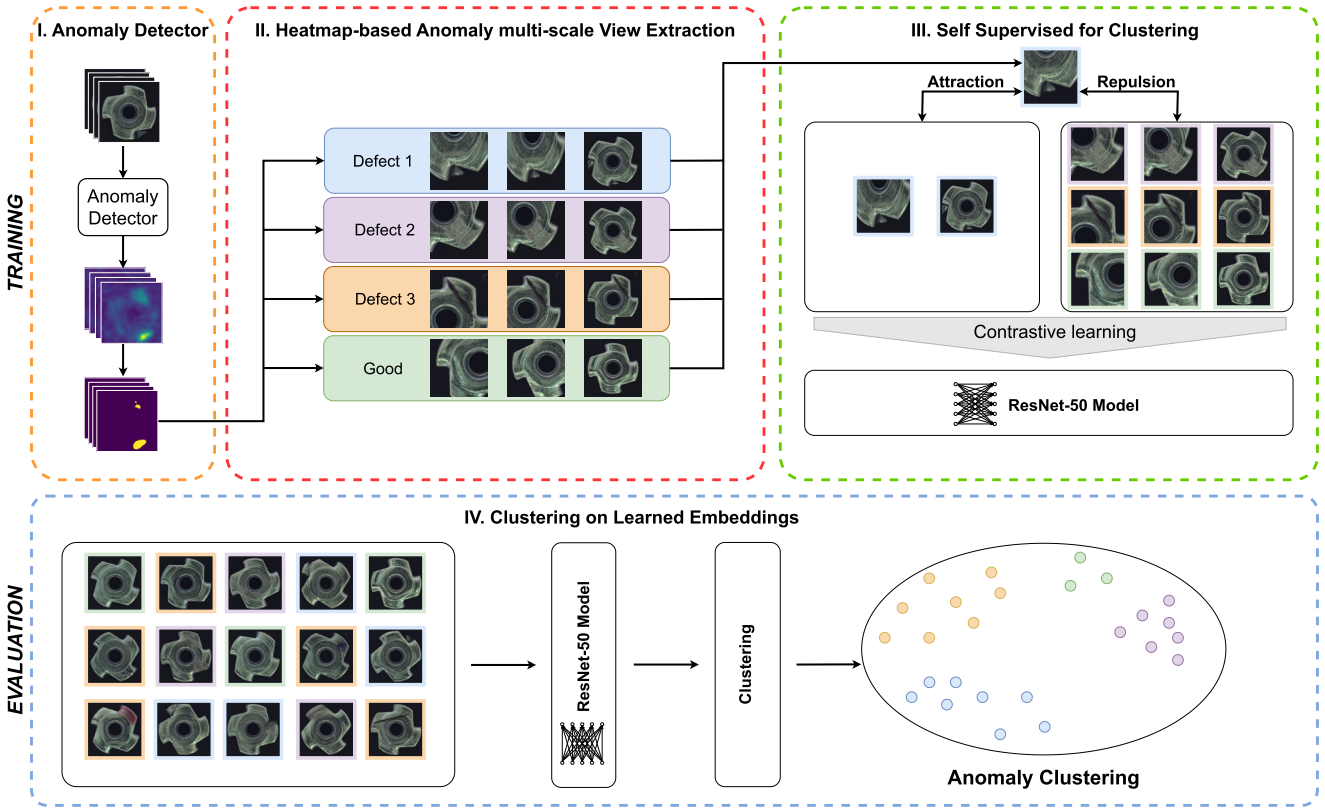

FIGURE 1. Visualization of the proposed framework. During the training phase, the anomaly detector module (orange box) first identifies defects by generating heatmaps and by calculating a statistical anomaly threshold to binarize the anomaly heatmap. Then, the Heatmap-based Anomaly multi-scale View Extraction (HAVE) module (red box) creates multi-scale views of the anomalies. These views are then used to train a self-supervised model in a contrastive manner (green box).During evaluation (blue box), the images are passed through the trained model to extract features. These features are then fed into a clustering algorithm, resulting in grouped representations that facilitate the categorization of different types of anomalies within the final embedding space.

Anomaly clustering groups the detected anomalies in images into clusters based on shared features or similarities, going beyond simple detection. Once an anomaly has been identified, this categorization into distinct defect types enables thorough statistical analyses, to uncover patterns and root causes of specific defects within the production cycle. Moreover, this structured information significantly facilitates the training of downstream defect classification systems, requiring reduced manual labeling efforts [7], [8]. However, anomaly clustering poses several challenges. First, the defect classes are inherently unbalanced towards the normal class. Furthermore, anomalies are often small and subtle, making it difficult to cluster images effectively when the images are analyzed holistically, since global representations may obscure defect-specific evidence with irrelevant background information. Finally, strong visual similarity between different defect types reduces the discriminative power of learned representations and makes anomaly clustering more difficult.

This paper introduces a novel method for clustering anomalies in images of industrial objects and textures, leveraging a self-supervised learning framework. A visual representation of the approach is shown in Figure 1. By utilizing spatial information from an anomaly detector, the system generates multi-scale views of anomalies, providing both local and global perspectives on individual defects. These multi-scale views are then used for contrastive training of a convolutional neural network, resulting in a clustered representation of the anomaly data. Our method surpasses current state-of-the-art (SOTA) techniques in industrial anomaly clustering. In summary, the contributions of this work are:

- We propose a novel method for clustering anomalies in object and texture images based on an efficient self-supervised learning approach.
- We introduce the Heatmap-based Anomaly multi-scale View Extraction (HAVE) module, which leverages heatmaps from the anomaly detection module to generate multi-scale defect views used to enhance the self-supervised learning process of the proposed framework.
- We conduct extensive experiments on two popular industrial anomaly detection datasets, demonstrating that our method consistently outperforms the SOTA in semi-supervised anomaly clustering.
- By reducing the level of supervision required and testing our approach even under weakly semi-supervised

conditions, we show that our method is comparable to the semi-supervised SOTA in anomaly clustering.

II. RELATED WORK

Over the years, anomaly detection in the industrial field has been extensively explored [9], [10]. Solutions have evolved across different paradigms, including supervised, unsupervised, and semi-supervised approaches, each addressing specific challenges in detecting anomalies in complex industrial environments.

Supervised methods [11], [12] rely on labeled data to train models to distinguish between normal and anomalous patterns. These methods typically perform well when a large amount of labeled data is available, but they require significant manual labeling, which can be costly and time-consuming in industrial settings. Compared to supervised methods, semi-supervised [1], [6], [13] and unsupervised [3], [14], [15], [16] approaches are more practical for real-world applications, as they require far less labeled data. To reduce the need for large amounts of labeled data, recent works have proposed one-shot and few-shot learning approaches [17], [18] as well as self-supervised methods [19], [20] that yield excellent results.

While traditional anomaly detection separates data into the two categories of normal and anomalous, we aim to group anomalies into multiple clusters, each representing a distinct type of anomalous behavior. Assuming that each image contains only one type of anomaly [21], [22], this task is similar to general image clustering, which has been extensively studied in the literature. For example, the SPICE framework [23] addresses clustering by dividing the neural network into a feature model for measuring instance-level similarity and a clustering head to identify discrepancies at the cluster level. Van Gansbeke et al. [24] introduced SCAN, a method that uses embedding features from a representation learning model to compute instance similarity, encouraging similar samples to share the same label and thus facilitating label feature learning. The GATCluster method [25] proposes an efficient two-step strategy for large images: first, it computes pseudo-targets for a large batch using a split-and-merge approach, then trains the model in mini-batches using these pseudo-targets. However, anomaly clustering presents several domain-specific challenges such as an inherent defect- class imbalance and a marked inter-class similarity between anomaly types that make it a complex task.

In the literature, only a few works address anomaly clustering within the industrial anomaly detection field. The first to introduce the anomaly clustering problem is Sohn et al. [22], who present a simple yet effective clustering framework that leverages patch-based pre-trained deep embeddings combined with off-the-shelf clustering methods. Ardelean et al. [21] propose a novel approach for clustering anomalies in largely stationary texture images in a fully unsupervised setting. Their approach combines blind anomaly localization with contrastive learning to solve this task. By accurately identifying anomalous regions, they focus on these areas of interest and then apply contrastive learning to improve the separability of different anomaly types while reducing intra-class variation. More recently, AnomalyNCD [26] addresses multi-class anomaly classification by explicitly incorporating anomaly-centered priors into the learning process. To overcome the issue of non-prominent anomalies, it introduces a main element binarization strategy to focus on anomaly-relevant regions, while a mask-guided representation learning module improves feature quality in the presence of weak semantic signals. Additionally, it supports both region-level and image-level classification through a region merging strategy. In contrast, the proposed framework uses anomaly heatmaps to extract structured multi-scale representations of anomalies, improving clustering robustness without requiring explicit supervision or pseudo-label refinement.

III. METHOD

Our approach introduces a new method for anomaly clustering, extending traditional anomaly detection into a comprehensive four-step process (cf. Figure 1). The first component involves training a SOTA anomaly detection backbone to identify potential defects (Sec. III-A). Next, spatial anomaly information detected from the detector is processed through a novel Heatmap-based Anomaly multi-scale View Extraction (HAVE) process, which uses the detected anomaly heatmaps to generate multi-scale views of the corresponding anomalies (Sec. III-B). In the third step, the generated views are used to train a network in a self-supervised manner via contrastive learning. In this process, views of the same image are attracted to each other, while views of different images are repulsed, aiming to create a clustered representation of the anomalies (Sec. III-C). After training of the network, in the fourth step, the learned embeddings are clustered (Sec. III-D). The full implementation of the proposed method is publicly available in our repository: https://gitlab.cc-asp.fraunhofer.de/omm/improving-industrial-anomaly-clustering

A. ANOMALY DETECTOR MODEL

The first step of the proposed approach for anomaly clustering involves the adoption of an anomaly detection module to extract the anomaly heatmaps corresponding to the input images. While our method is agnostic to the anomaly detection backbone, we use PatchCore [6] due to its superior performance on MVTec-AD [27]. However, as demonstrated in Section IV-G, HAVE can integrate seamlessly with alternative anomaly detectors such as MambaAD [28], as long as the selected model produces an anomaly heatmap at a pixel-level in the images.

PatchCore is a semi-supervised model that captures patch-level features from normal training images and stores them in a memory bank, creating a high-dimensional ''normality'' space. During inference, each patch in a new image is compared to the patches in this memory bank to generate

an anomaly score. Patches that deviate from the norm will have higher anomaly scores, allowing the model to identify unusual patterns accurately. The model returns an anomaly heatmap $A_i \in \mathbb{R}^{H\times W}$ for each image $I_i \in \mathbb{R}^{H\times W\times 3}$, $i = 1 \dots N$, and an anomaly threshold $T_l \in \mathbb{R}, T_l > 0$ for each object and texture category $l = 1 \dots L$.

Here, H and W are the height and width of an image in pixels, respectively. The threshold is computed without relying on ground truth annotations, using the k-sigma method. Specifically, the threshold T_l is defined as $T_l = \mu_l + k \cdot \sigma_l$, where μ_l and σ_l represent the mean and standard deviation of all anomaly scores for category l, respectively. The user-defined parameter k is a scaling factor that controls the sensitivity to outliers. We selected a k-sigma threshold such that, under the assumption of normally distributed anomaly scores, the 0.99 quantile is reached, as stated in [27]. The anomaly detector is not simply used as a pre-processing component to identify defects. Its role is to provide spatially precise cues about anomalies, which guide the subsequent generation of multi-scale views focused on defect positions. This process suppresses irrelevant background information and allows the contrastive learning phase to concentrate on the appearance and context of the anomalous region.

B. HEATMAP-BASED ANOMALY MULTI-SCALE VIEW EXTRACTION (HAVE)

Anomaly clustering requires localizing and differentiating anomalies, which can be challenging due to their small size and subtle variations. Instead of extracting random patches or relying solely on global representations, we introduce HAVE, which leverages anomaly heatmaps to generate structured multi-scale views for contrastive training. HAVE operates in three key steps.

The first step is represented by the processing of the anomaly heatmaps generated by the anomaly detection backbone. Given an input image I_i and its corresponding anomaly heatmap A_i, obtained from an anomaly detection model, we threshold the heatmap to extract regions of interest $M_{\text{bin}} \leftarrow 1(A_i > T_l)$ where T_l is the anomaly threshold introduced above. The resulting binary mask M_{bin} highlights anomalous regions while discarding irrelevant background information.

As a second step, we extract from the input images a set of multi-scale views centered around the detected anomalies. The anomalous regions represented in the binary mask M_{bin} are partitioned into components C_i using the Spaghetti connected-components labeling algorithm [29]. For each component $c_{i,z} \in C_i$ with centroid $p_{i,z}$ and for each crop-scale factor $s_j \in S$ from a user-defined set of scale factors $S \subset \mathbb{R}$, we generate square views $v^d_{i,z,j}$ centered at $p_{i,z}$, with size $(s_j \cdot t)$, where $t \in \mathbb{N}$ is the target size. Here, the superscript d indicates defective images. Each cropped view is then rescaled to $t \times t$, i.e., $v^d_{i,z,j} \leftarrow \texttt{scale}(\texttt{crop}(I_i, r_{i,z,j}, t))$ where $r_{i,z,j} = (i, p_{i,z}, s_j)$. The crop-scale factor s_j controls the field of view of the defect-centered crop. In this setting, $s_j > 1$ results in a larger crop enabling more global context, whereas $s_j < 1$ yields a narrow crop with an apparent zoom-in. This ensures that defects are examined from multiple perspectives, reinforcing robust feature learning by allowing the framework to capture both global context and fine-grained details.

Finally, the cropping values R, comprising the tuples $r_{i,z,j}$ of all defective images, are stored and later applied to "good" images I^{good} according to $v^g_{i,z,j} \leftarrow \texttt{scale}(\texttt{crop}(I_g, r_{i,z,j}, t))$, where $I_g \in I^{good}$ are randomly selected normal images, to generate corresponding views. This step encourage spatial alignment between defect-centered views and their undamaged counterparts, increasing the likelihood that views with and without defects are extracted from the same spatial region. This promotes contrastive learning, particularly the learning of local discriminative features. Algorithm 1 summarizes the entire process carried out for each image in the dataset.

Algorithm 1 HAVEás Main Algorithm in the Semi-Supervised Setting

Require:
1) Set of input images $\{I_i\}_{i=1}^N, I_i \in \mathbb{R}^{H\times W\times 3}$
2) Set of anomaly heatmaps $\{A_i\}_{i=1}^N, A_i \in \mathbb{R}^{H\times W}$
3) k-sigma-based anomaly threshold T_l
4) Set of crop-scale factors $S = \{s_j\}$
5) Target image size t

Output:
- Set of views $(V_i)_{i=1}^N$, where: $V_i = \{v^d_{i,z,j}, v^g_{i,z,j}\}^{z=1,\dots,|C_i|}_{j=1,\dots,|S|}$ with each $v \in \mathbb{R}^{t\times t\times 3}$

$R \leftarrow \emptyset$
$I^{\text{good}} \leftarrow \emptyset$
for each image $I_i \in I$ **do**
  $V_i \leftarrow \emptyset$
  $M_{\text{bin}} \leftarrow 1(A_i > T_l)$
  if $\texttt{sum}(M_{\text{bin}}) > 0 \wedge \texttt{class}(I_i) \neq$ "good" **then**
    $C_i \leftarrow \texttt{connected\_components}(M_{\text{bin}})$
    for each c_z in C_i **do**
      $p_{i,z} \leftarrow \texttt{centroid}(c_z)$
      for each s_j in S **do**
        $r_{i,z,j} \leftarrow (i, p_{i,z}, s_j)$
        $v^d_{i,z,j} \leftarrow \texttt{scale}(\texttt{crop}(I_i, r_{i,z,j}, t))$
        $R \leftarrow R \cup r_{i,z,j}$
        $V_i \leftarrow V_i \cup v^d_{i,z,j}$
      end for
    end for
  else
    $I^{\text{good}} \leftarrow I^{\text{good}} \cup I_i$
  end if
end for

for each $r_{i,z,j}$ in R **do**
  $I_g \leftarrow \texttt{random}(I^{\text{good}})$
  $v^g_{i,z,j} \leftarrow \texttt{scale}(\texttt{crop}(I_g, r_{i,z,j}, t))$
  $V_i \leftarrow V_i \cup v^g_{i,z,j}$
end for

C. SELF-SUPERVISED LEARNING FOR CLUSTERING

Inspired by the SimCLR framework [30], this work proposes an adapted approach specifically tailored for anomaly clustering. The proposed network architecture comprises a convolutional encoder based on the ResNet-50 model [31] and an MLP projector. The encoder utilizes convolutional layers to capture high-level structural and spatial information within input images, facilitating the identification of anomalous features. The MLP projector projects the extracted features into a compact latent space during the training. The network learns data representations by maximizing the agreement between scaled views of the same sample. Specifically, agreement is maximized between different views of the same image for anomalous images. For defect-free images, the available class information is exploited, and thus, the agreement is maximized among all views of such images. We evaluate the impact of the agreement strategy for defect-free images in Section IV-F4 of the ablation study.

To enforce this contrastive learning process, we selected the Normalized Temperature-scaled Cross-Entropy Loss (NT-Xent) [32] as the objective loss function. NT-Xent promotes smaller distances between the embeddings of the positive pairs in the latent space while pushing apart the embeddings of the negative pairs and is defined as

$$\text{NT-Xent}(z_i, z_j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{M} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{1}$$

where $\text{sim}(z_i, z_j)$ represents the cosine similarity between the embeddings z_i and z_j, $\mathbb{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 if $k \neq i$, and τ represents a non-negative temperature parameter. Specifically, τ is a scaling factor of the exponential function to change the degree of contrast between positive and negative pairs. A low τ amplifies the differences between similar embeddings. Consequently, the exponential function will penalize negative pairs with slight similarity to a greater extent. On the other hand, with a high τ, the difference between positive and negative pairs is reduced.

D. CLUSTERING ON LEARNED EMBEDDINGS

Once training is complete, clustering is performed. In this phase the original images, without crops, are passed through the trained network, which is used as features extractor. The extracted features are then clustered using K-means [33]. We assume that the number of clusters K is known for performance evaluation, similar to [21], [22], and [26].

IV. EXPERIMENTS

To evaluate the effectiveness of our contributions, we conducted a comparative analysis between the anomaly clustering approach introduced in this work and existing SOTA approaches on three distinct datasets. In addition, we conducted an extensive ablation study to examine the impact of each individual component.

A. DATASETS

To test our approach, we selected three widely used datasets for anomaly detection in industrial image settings: MVTec Anomaly Detection (MVTec-AD) [27], Magnetic Tile Defect (MTD) [34] and the Metal Parts Defect Detection Dataset (MPDD) [35]. For all the datasets, the training set exclusively comprises defect-free images, while the test set includes images featuring various types of defects and defect-free images.

The MVTec-AD dataset consists of 5,354 high-resolution color images and is a widely adopted benchmark in recent studies focusing on anomaly detection and localization. This dataset comprises 3,629 training images and 1,725 test images distributed across 15 categories. Among these categories, five represent various textures, while the rest encompass a range of products with diverse attributes. As in Sohn et al. [22], the ''combined'' defect-classes containing multiple types of anomalies were removed from the dataset.

The MTD dataset is designed for saliency detection of surface defects and consists of 1,076 images for training and 660 images for testing. The MTD dataset contains six types of defects: blowhole, break, crack, fray, uneven, and free.

The MPDD dataset consists of metal components exhibiting significant variability in object positioning, orientation, lighting conditions, and background complexity. The dataset comprises 888 training and 458 test images, categorized into six categories.

B. METRICS

Similar to previous works on anomaly clustering [21], [22], the metrics Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F1 score are used for performance evaluation. NMI is a metric based on information theory and is calculated as the ratio of normalized mutual information to the arithmetic mean between the entropies of the predicted cluster and label ground truth. NMI takes values from 0 to 1, where 1 represents the perfect match. ARI is based on the Rand index [36] and assesses how each pair of elements is evaluated in the partitions produced by the clustering algorithm and the reference partition with the ground truth labels. ARI ranges from -1 to 1, where 1 represents a perfect match and 0 random similarity. Negative ARI values indicate a worse match than a randomly expected match. Finally, the F1 score was used to account for the imbalance between defect classes.

The optimal matching between the ground truth and the predicted cluster is performed by the Jonker-Volgenant algorithm [37] provided by the SciPy library [38]. Compared to the Hungarian algorithm [39] used in Sohn et al. [22], the Jonker-Volgenant algorithm [37] addresses the same problem with faster execution time.

C. TRAINING SETUP

Using the official implementation, we trained the PatchCore [6] model with a WideResNet-50 [40] backbone

TABLE 1. NMI, ARI and F1 scores for semi-supervised and weakly semi-supervised anomaly clustering methods on the MVTec-AD and MTD datasets. We compare the current SOTA with our method built upon two different anomaly detection backbones: PatchCore and MambaAD. For all metrics a higher value is better ↑. The best NMI metrics are reported in bold, while the second-best results are underlined.

	Model	Metric	MVTec-AD			MTD
			Average Objects	Average Textures	Average All	Average
Semi-supervised	Sohn et al. [22]	NMI↑	0.577	0.669	0.608	0.390
		ARI↑	0.449	0.570	0.489	0.314
		F1↑	0.628	0.698	0.651	0.490
	AnomalyNCD + CPR [26]	NMI↑	0.687	**0.833**	<u>0.736</u>	-
		ARI↑	0.622	0.777	0.674	-
		F1↑	0.775	0.865	0.805	-
	AnomalyNCD + PatchCore [26]	NMI↑	0.627	0.755	0.670	0.380
		ARI↑	0.557	0.688	0.601	0.390
		F1↑	0.740	0.826	0.769	0.617
	Our-PatchCore	NMI↑	<u>0.705</u>	0.788	0.733	**0.491**
		ARI↑	0.644	0.738	0.675	0.451
		F1↑	0.758	0.824	0.780	0.421
	Our-MambaAD	NMI↑	**0.744**	<u>0.826</u>	**0.771**	<u>0.447</u>
		ARI↑	0.680	0.799	0.719	0.338
		F1↑	0.752	0.866	0.790	0.323
Weakly semi-sup	Our-PatchCore	NMI↑	0.645	0.777	0.689	0.398
		ARI↑	0.540	0.678	0.586	0.486
		F1↑	0.706	0.816	0.743	0.474
	Our-MambaAD	NMI↑	0.602	0.751	0.652	0.424
		ARI↑	0.521	0.702	0.582	0.526
		F1↑	0.691	0.818	0.734	0.442

network, pre-trained on the ImageNet [41] dataset, at a resolution of 224 × 224 pixels. The heatmaps are scaled and processed according to the procedure outlined in Section III-B. The crop-scale factors S used were $\{0.8, 1.0, 4.0\}$. The extracted views as well as the original images are all rescaled to the target image size t. In our experiments, we set t to 512 pixels for MVTec-AD and MPDD datasets, and to 256 pixels for the MTD dataset, due to the lower resolution of the original images. The ResNet-50 [31] contrastive network was trained with an input resolution of $t \times t$ pixels, using the NT-Xent loss function Eq. (1) with temperatures $\tau = 0.2$. The network is fine-tuned using pre-trained ImageNet [41] weights available in the PyTorch library. We explored different combinations of learning rates and batch sizes; given the modest size of the dataset, we used batch sizes of $\{16, 32\}$ and learning rates of $\{10^{-5}, 10^{-6}\}$. During training, we used the AdamW optimizer [42] with a weight decay of 10^{-6}. A data augmentation strategy is also applied, incorporating affine transformations and random color distortions. The affine transformation includes random rotations up to 90 degrees, slight translations, and scale variations to preserve local anomaly features within each view. Our method requires approximately 120 minutes for multi-scale view generation and self-supervised training on the MVTec dataset using a single NVIDIA Tesla A100 GPU.

D. RESULTS

We compared the performance of our model to the current state-of-the-art approach by Sohn et al. [22] in the semi-supervised setting for MVTec-AD and MTD datasets. Following the same protocol as Sohn et al. [22], we assume access to ground truth labels only for normal data in the semi-supervised scenario. Table 1 presents aggregated results obtained by averaging the performance over three runs. Detailed results by category, including standard deviation, are shown in Table 8 of the Appendix. As we can see, our method achieves the best performance and significantly outperforms the previous state-of-the-art on all the individual categories of the MVTec-AD datasets, except for one object category and one texture category. On both the average of texture and object categories in the MVTec-AD dataset, our framework outperforms all others across all metrics. Overall, our approach improves by 0.125, 0.186, and 0.129 on the NMI, ARI and F1 metrics, respectively, for the MVTec-AD dataset compared to the previous SOTA. These results highlight the effectiveness of our framework in accurately clustering and identifying anomalies, with consistent advantages across both object and texture categories.

On the MTD dataset as well, our approach surpasses the SOTA method by Sohn et al. [22], with an improvement of 0.101 for NMI and 0.137 for ARI. However, the

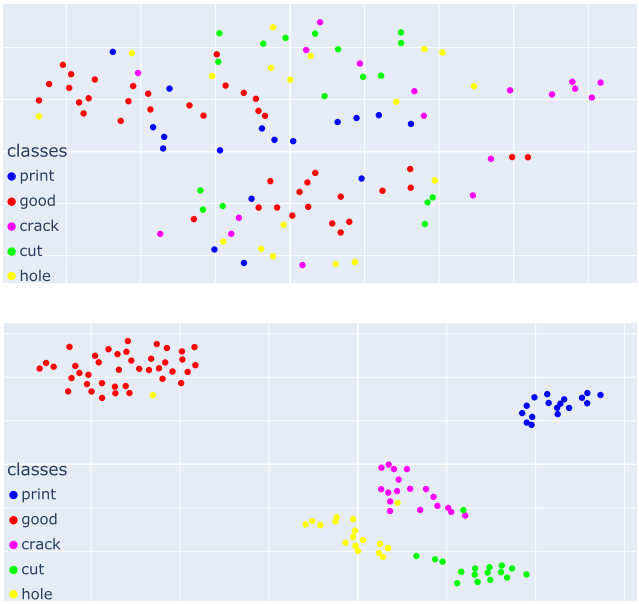

FIGURE 2. Visualization of network features for the MVTec-AD Hazelnut category before (top) and after (bottom) contrastive learning. Feature vectors are projected into two dimensions using t-SNE. The color of each instance corresponds to the respective type of anomaly in the ground truth.

F1 score remains higher for the approach by Sohn et al. Although our approach achieves good cluster separation, the lower F1 score indicates issues with the model's accuracy, particularly in correctly classifying anomalies. We believe that this discrepancy is due to inconsistencies in image format and resolution, resulting from a non-uniform acquisition process in the MTD dataset. These variations in image properties across different instances could contribute to the challenges the model faces in making precise classifications, especially for anomalies in certain defect classes.

In Figure 2, we provide the t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations of the Hazelnut class features of MVTec-AD dataset. Specifically, the figure illustrates a comparison between the distribution of test set instances in the latent space at the beginning (Fig. 2, top) and the end (Fig. 2, bottom) of the self-supervised contrastive-based training process. As we can see, the initial feature distribution is more scattered, with limited structure distinguishing different anomaly types. After training, however, we observed a marked improvement in the separability of anomaly features, resulting in distinct clusters corresponding to specific anomaly types. This clear clustering indicates that the contrastive learning approach enables the model to capture more discriminative features, effectively grouping similar anomalies together while maintaining separation between distinct types. Further qualitative comparisons with Sohn et al. [22] and AnomalyNCD [26] based on t-SNE visualizations are reported in the Appendix.

TABLE 2. Model performance evaluation under different crop-scale factors *s*. The results from the overall best-performing combination are reported in bold while the scores of the best-performing zoom level for each crop type are underlined. Higher *s* corresponds to wider crops (more context), while smaller *s* corresponds to narrower crops (more details).

Narrow Crop ($s<1$)	Original ($s=1$)	Wide Crop ($s>1$)	NMI↑	ARI↑	F1↑
-	1.0	-	0.302	0.182	0.447
-	1.0	2.5	0.641	0.559	0.689
-	1.0	3.0	<u>0.674</u>	0.593	0.715
-	1.0	3.5	0.670	<u>0.608</u>	0.702
-	1.0	4.0	0.671	0.605	<u>0.720</u>
-	1.0	4.5	0.641	0.567	0.686
0.4	1.0	-	<u>0.469</u>	0.374	0.534
0.6	1.0	-	0.467	<u>0.376</u>	<u>0.551</u>
0.8	1.0	-	0.462	0.326	0.531
0.8	1.0	3.0	0.667	0.594	0.712
0.4	1.0	4.0	0.678	0.611	0.734
0.8	1.0	4.0	**0.687**	**0.627**	**0.740**
0.6	1.0	4.0	0.675	0.596	0.716

E. WEAKLY SEMI-SUPERVISED

We analyse the weakly semi-supervised setting on datasets where the classification of images as "good" or "anomalous" relies solely on predictions from the anomaly detector. In the semi-supervised setting, ideal anomaly detection is assumed, while the weakly semi-supervised case is susceptible to false negatives, which may lead to erroneously including "good" images in the batch. This contamination can compromise clustering quality and degrade overall performance. The generation of views in the HAVE module was adapted to the weakly semi-supervised setting. For this purpose Algorithm 1, which describes the process in the semi-supervised case, is slightly modified. In Line 6 the condition is changed to $\texttt{sum}(M_{\text{bin}}) > 0$, i.e., the additional condition $\texttt{class}(I_i) \neq$ "good" is dropped.

As shown in Table 1, the weakly semi-supervised approach leads to lower performance compared to the semi-supervised setting. This is evidently due to the information provided to the detector model and its performance. Nevertheless, our weakly semi-supervised results surpass the semi-supervised approach by Sohn et al. [22] for MVTec-AD and are comparable for MTD. On the MTD dataset, the weakly semi-supervised method exhibits notably low NMI, ARI, and F1 scores, which correlate with suboptimal anomaly detection performance of PatchCore, as detailed in the Appendix.

F. ABLATION STUDY

In this section, we present the details of the ablation study conducted to validate the importance of each component of the proposed framework.

1) EFFECTIVENESS OF MULTI-SCALE AUGMENTATION

To assess the effectiveness of our multi-scale augmentation strategy within HAVE, we analyze the model's performance

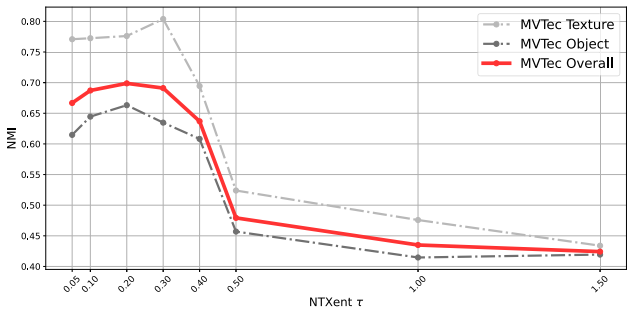

FIGURE 3. Evaluation of clustering performance in MVTec-AD dataset across different τ using NMI.

under different combinations of crop-scale factors s on the MVTec-AD dataset. Smaller values of s produce narrow crops that emphasize fine-grained details, whereas larger values yield wide crops that include more contextual information. The results, obtained with fixed hyperparameters and τ set to 0.1, are averaged over three runs and reported in Table 2. As a baseline, we evaluate the clustering performance by applying the K-means algorithm [33] directly to the features extracted from the ResNet-50 [31], pre-trained on ImageNet [41].

We then conduct a two-view analysis, combining the original image with either narrow crops ($s<1$) or wide crops ($s>1$). The results show that wide crops contribute more significantly to performance improvements than narrow ones, as reported in Table 2, particularly for crop factors between 3.0 and 4.0, which effectively capture broader contextual cues while maintaining sufficient detail. Narrow crops still improve the performance compared to the single-view baseline but with a smaller gain, suggesting that emphasizing local details alone is less beneficial than incorporating additional context.

Extending the analysis to three views, we integrate both narrow and wide crops, further enhancing performance compared to the best two-view model. Among the tested configurations, the combination $\{0.8, 1.0, 4.0\}$ achieves the highest performance, suggesting that multi-scale augmentation exploits complementary information from different zoom levels, improving the model's ability to capture both global and local anomaly features.

2) SENSITIVITY ANALYSIS

Selecting the appropriate value for parameter τ in the NTXent loss plays a crucial role in the success of the task as it controls the degree of separation between the embeddings in contrastive learning. To evaluate its impact, we analyzed the performance of the model on the MVTec-AD dataset when selecting different temperature values τ. The results, summarized in Figure 3, indicate a notable decline in the network's ability to learn effective data representations as τ increases. The optimal value of NMI is achieved when τ is set to 0.2, which, on average, yields the best performance across the dataset.

3) ROTATION AS DATA AUGMENTATION

We investigated the relationship between clustering quality and the application of varying degrees of image rotation in the training data augmentation pipeline. Networks were trained with maximum rotation augmentations set to $\{30, 45, 90\}$ degrees, and six models were developed per scenario, each employing different combinations of hyperparameters. All the models were trained with τ set to 0.1. Results are reported in Table 3. As observed, our work surpasses the current SOTA across all metrics, even without image rotations. Nonetheless, rotations help the model achieve better generalization in learning data representations. While 90-degree rotations can, on average, enhance the NMI metric, their impact is not equal between textures and objects. In particular, texture categories benefit more from augmentation, with a total increase of 0.115 in NMI, compared to objects, where performance remains more or less unchanged. This, in our opinion, can be explained by the property of texture categories to be more invariant to rotations than object categories. The increase in variety provided by rotation for these categories provides additional examples that are more consistent with the natural variations of textures (e.g., grids).

4) GOOD INSTANCES AGGREGATION

In this section, we investigate the effect of maximizing the agreement across all views of all "good" images during contrastive training. We compare the case where the networks were trained using the same criterion for both anomalous and good classes with those where all views of all good images are aggregated together. We trained six models per scenario, with different combinations of hyperparameters applying a 90-degree rotation. All the models were trained with τ set to 0.1. Table 4 shows the results for each class of the the MVTec-AD dataset. Without the use of good instances aggregation, our work outperform already the current SOTA. However, using good class aggregation during contrastive training significantly improves performance in all the observed metrics.

G. GENERALITY ON DIFFERENT AD BACKBONES

To further demonstrate the backbone-agnostic nature of our work, we evaluated its performance when using MambaAD [28] a model for multi-class anomaly detection based on Mamba [43], as the anomaly detection backbone. Similar to PatchCore [6], MambaAD [28] produces pixel-level anomaly heatmaps for each image, allowing seamless integration in our pipeline. We trained the backbone with its official implementation within the ADer framework [44]. Table 1 presents results on MVTec-AD and MTD, averaged over three runs. In the semi-supervised setting, MambaAD surpasses Sohn et al. [22] and our results with PatchCore on the MVTec dataset. This improvement may be attributed to the richer multi-scale information contained in the anomaly heatmaps, which align better with our approach.

TABLE 3. Variation in clustering results for selected MVTec-AD categories when adjusting the degree of rotation used as data augmentation. For all metrics a higher value is better ↑. The best NMI metrics are reported in bold.

MVTec-AD	w/o Rotation			Rotation 30			Rotation 45			Rotation 90		
Categories	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑
bottle	0.761	0.710	0.879	0.697	0.647	0.854	0.688	0.622	0.840	**0.789**	0.748	0.898
cable	**0.769**	0.782	0.824	0.675	0.737	0.631	0.708	0.831	0.775	0.748	0.767	0.773
capsule	0.486	0.317	0.586	0.531	0.346	0.609	0.467	0.322	0.567	**0.569**	0.431	0.647
hazelnut	**0.892**	0.912	0.947	0.836	0.850	0.877	0.871	0.875	0.907	0.844	0.829	0.864
metal nut	0.748	0.659	0.823	0.705	0.583	0.711	**0.846**	0.790	0.899	0.738	0.564	0.679
pill	**0.573**	0.446	0.607	0.505	0.329	0.596	0.553	0.372	0.649	0.521	0.359	0.591
screw	0.392	0.388	0.478	0.417	0.425	0.421	**0.586**	0.449	0.443	0.466	0.447	0.499
toothbrush	**0.508**	0.498	0.844	0.430	0.370	0.798	0.467	0.432	0.821	0.438	0.378	0.798
transistor	0.640	0.568	0.588	0.655	0.470	0.620	0.618	0.748	0.648	**0.670**	0.536	0.579
zipper	0.780	0.744	0.794	0.709	0.640	0.654	0.811	0.748	0.733	**0.821**	0.756	0.754
carpet	0.800	0.742	0.835	0.758	0.671	0.685	0.743	0.674	0.800	**0.805**	0.780	0.885
grid	0.375	0.366	0.428	0.397	0.453	0.470	0.507	0.468	0.479	**0.555**	0.503	0.638
leather	0.739	0.679	0.721	0.858	0.866	0.927	0.894	0.878	0.927	**0.937**	0.936	0.963
tile	0.942	0.943	0.972	**0.979**	0.975	0.992	0.949	0.946	0.970	**0.979**	0.975	0.992
wood	0.548	0.440	0.619	0.746	0.660	0.799	**0.755**	0.681	0.801	0.702	0.613	0.706
Average object	0.655	0.602	0.737	0.616	0.540	0.677	0.662	0.619	0.728	**0.661**	0.605	0.669
Average texture	0.681	0.634	0.715	0.748	0.725	0.769	0.770	0.729	0.795	**0.796**	0.761	0.837
Average	0.663	0.601	0.723	0.652	0.581	0.709	0.697	0.643	0.749	**0.706**	0.706	0.751

TABLE 4. Semi-supervised results on MVTec-AD dataset, with and without good instance aggregation processing. All categories and the averages are reported. The best NMI metrics are reported in bold.

Dataset	w/o good instances aggregation			With good instances aggregation		
MVTec-AD	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑
bottle	0.701	0.620	0.827	**0.789**	0.748	0.898
cable	0.703	0.691	0.699	**0.748**	0.767	0.773
capsule	0.344	0.182	0.435	**0.569**	0.431	0.647
hazelnut	0.737	0.712	0.840	**0.844**	0.829	0.864
metal nut	0.684	0.580	0.747	**0.738**	0.564	0.679
pill	0.478	0.326	0.621	**0.521**	0.359	0.591
screw	**0.535**	0.349	0.535	0.466	0.447	0.499
toothbrush	0.231	0.019	0.618	**0.438**	0.378	0.798
transistor	0.557	0.507	0.507	**0.670**	0.536	0.579
zipper	0.756	0.706	0.812	**0.821**	0.756	0.754
carpet	0.700	0.678	0.812	**0.805**	0.780	0.885
grid	0.447	0.282	0.395	**0.555**	0.503	0.638
leather	0.889	0.881	0.923	**0.937**	0.936	0.963
tile	0.959	0.951	0.985	**0.979**	0.975	0.992
wood	**0.713**	0.603	0.677	0.702	0.613	0.706
Average Object	0.572	0.469	0.657	**0.661**	0.582	0.708
Average Texture	0.742	0.679	0.758	**0.796**	0.761	0.837
Average	0.629	0.539	0.691	**0.706**	0.642	0.751

In contrast, in the weakly semi-supervised case, the performance of the MambaAD-based framework degrades more significantly than the one of the PatchCore-based one, though it still surpasses Sohn et al. [22], even though it works in a semi-supervised setting. Consistent with the observation in Section IV-E, this degradation correlates with the anomaly detector's performance. Specifically, the categories capsule, pill and screw, which exhibit the lowest clustering performance, also show the lowest AUROC scores. On the MTD dataset, the results with the MambaAD backbone show improvements in both the semi-supervised and weakly semi-supervised settings compared to the semi-supervised approach of Sohn et al. [22], but remain lower than those achieved with PatchCore [6]. This outcome may be influenced by MambaAD's [28] increased sensitivity to the dataset's limited resolution.

H. FAILURE ANALYSIS AND INTER-CLASS SIMILARITY

We analysed the structure of the learned embedding space and the distribution of clustering errors across the different categories to evaluate the performance of the proposed framework.

In particular, we quantified intra-class compactness and inter-class separability using the cosine distance. Specifically, inter-class separability was measured by calculating the cosine distance between the centroids of the different defect classes within each category, whilst intra-class compactness was calculated as the cosine distance between each embedding and the centroid of its reference defect class. Figure 5 shows, for each category in the MVTec-AD dataset, the relationship between the average intra-class distance and the average inter-class distance.

As expected, the best-clustered categories tend to exhibit low intra-class distance and high inter-class distance, indicating more compact and better-separated embeddings. Conversely, categories showing lower clustering performance are associated with lower separation between centroids and, in some cases, greater intra-class dispersion. This confirms that inter-class similarity represents one of the main sources of error in the method.

Further evidence emerges from the analysis of the confusion matrices presented in the Appendix, which show

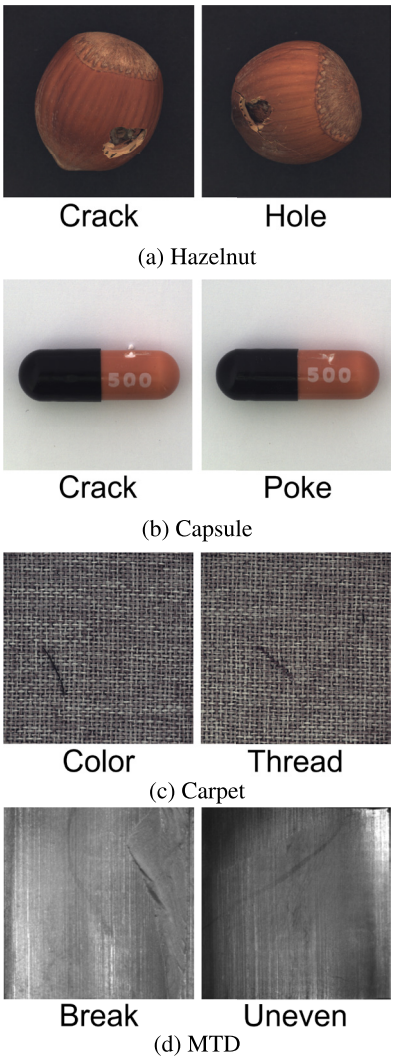

FIGURE 4. Examples of inter-class similarity among the Hazelnut, Capsule, Carpet, and Magnetic Tile anomaly types in the MVTec-AD and MTD datasets. Each class presents two visually similar images from different anomaly types that our system clusters together, even though these images belong to distinct anomaly categories.

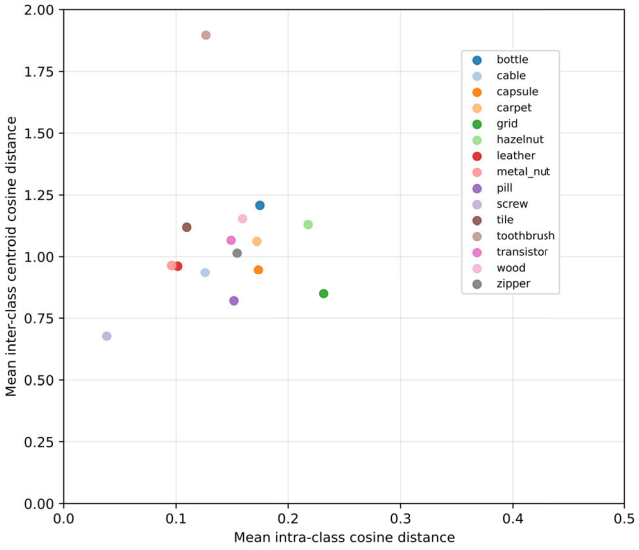

FIGURE 5. Relationship between average intra-class cosine distance and average inter-class centroid cosine distance for each MVTec-AD category. Categories with better clustering performance tend to exhibit lower intra-class distance and higher inter-class distance.

that the incorrectly clustered classes often concentrate on a limited number of pairs of visually distinguishable defect categories, rather than being distributed uniformly across all classes.

I. IMPLEMENTATION DETAILS AND RUNTIME ANALYSIS

In real-world deployment, the framework is inherently divided into an offline stage and an online stage. The offline stage includes running the anomaly detection backbone on the available data, generating the multi-scale views through HAVE, training the contrastive encoder, extracting defect embeddings, and fitting the K-means model on the learned latent representation. The online stage is considerably lighter, as it only requires extracting the embedding of the incoming sample with the trained encoder and assigning it to the nearest cluster centroid. When the method is deployed in an end-to-end inspection pipeline, an upstream anomaly detector can be used as a gating stage, so that clustering is performed only for samples identified as defective. Therefore the online computation cost is substantially lower than the offline one since it does not involve multi-view generation, contrastive optimization, and K-means fitting.

The offline complexity mainly depends on the contrastive learning of the multi-scale views generated by HAVE. Let N^d be the number of defective images and $\bar{C}$ the average number of anomalous connected components per defective image, the total number of generated views is approximately $V \approx 2N^d\bar{C}S$. Therefore, the dominant offline cost for a fixed encoder architecture and input resolution can be expressed as $\mathcal{O}(EV) = \mathcal{O}(EN^d\bar{C}S)$, where E is the number of training epochs. When dealing with large industrial datasets, the proposed pipeline can be further optimized by parallelizing the HAVE module, thereby reducing the preprocessing bottleneck. Furthermore, larger datasets may benefit from the use of larger batch sizes during contrastive training.

The view generation and K-means fitting stages are comparatively lighter, with costs $\mathcal{O}(N + V)$ and $\mathcal{O}(NKT)$, respectively, where T is the number of iterations and assuming fixed crop resolution and embedding dimension. The online stage complexity depends on the nearest-centroid assignment and scales linearly with the number of clusters K, i.e., $\mathcal{O}(K)$ assuming fixed embedding dimensionality.

We evaluated the inference time on a workstation equipped with an NVIDIA RTX 4070 GPU, an AMD Ryzen Threadripper 7960s processor, and 256 GB RAM memory.

TABLE 5. Results obtained on MPDD dataset, comparing AnomalyNCD-PC method with our semi-supervised and weakly semi-supervised anomaly clustering. For all metrics a higher value is better ↑. The best NMI metrics are reported in bold.

Supervision	Semi-supervised						Weakly semi-supervised		
MPDD	AnomalyNCD-PC [26]			Our-PC			Our-PC		
Metric	NMI↑	ARI↑	F1↑	NMI	ARI↑	F1↑	NMI	ARI↑	F1↑
Br. black	0.616	0.603	0.785	**0.740**	0.676	0.797	0.341	0.266	0.649
Br. brown	0.355	0.281	0.675	**0.721**	0.681	0.815	0.567	0.536	0.767
Br. white	**1.00**	1.00	1.00	0.896	0.922	0.948	0.776	0.775	0.934
Connector	0.628	0.661	0.909	**0.852**	0.908	0.973	0.496	0.522	0.855
Metal Pl.	0.695	0.614	0.794	0.787	0.727	0.778	**0.807**	0.778	0.895
Tubes	0.280	0.103	0.673	**1.000**	1.000	1.000	0.430	0.429	0.817
Average	0.698	0.544	0.806	**0.833**	0.838	0.907	**0.570**	0.551	0.820

On preprocessed inputs with batch size 1, the encoder requires 6.21 ms median latency (p50), while nearest-centroid assignment requires only 0.11 ms. The combined clustering-stage latency is therefore 6.33 ms (p50), showing that the online cost is dominated by feature extraction rather than by cluster assignment. To assess scalability with respect to the number of clusters, we refitted K-means with $K \in \{2, 4, 8, 16, 32, 64\}$ on the learned embeddings and measured online nearest-centroid assignment latency using the corresponding real centroids. Assignment latency remained nearly constant in the tested range (about 0.11–0.13 ms p50), while end-to-end clustering-stage latency remained around 6.34–6.38 ms p50. This indicates that, for realistic industrial settings, the online clustering stage scales well and is dominated by encoder inference rather than by the number of cluster centroids.

J. RESULTS ON MPDD DATASET

To further evaluate the robustness of our approach, we assess its performance on the MPDD dataset [35]. We compare our method with AnomalyNCD [26] using a PatchCore backbone. The results, averaged over three runs, are reported in Table 5. In the semi-supervised settings, our method achieves strong clustering performance, outperforming the AnomalyNCD-PC method and demonstrating how multi-scale information allows the model to learn defect characteristics even with images with high variability in pose and lighting, like those from MPDD.

In the weakly semi-supervised settings, the average performance is consistently lower than in the semi-supervised case, with varying degrees of impact across different categories, and also in this case we can see a correlation with the AD performance as observed in Sections IV-E and IV-G.

K. LIMITATIONS

While our approach demonstrated improved performance compared to previous-state-of-the art, there are still some limitations that emerged during its development, highlighting suggesting potential further refinements. First, a primary limitation of our approach is that the proposed architecture is highly dependent on the performance of the anomaly detection module. Specifically, the detection and segmentation of anomalies in the input image are driven by an anomaly detection model. Since the quality of the detected anomalies directly influences the clustering results, any weaknesses in the anomaly detection model can negatively impact the overall performance of our framework. While this dependency on the anomaly detection module represents a limitation, we believe it also serves as a strength of our approach. The flexibility of our framework allows for seamless integration with various anomaly detection backbones beyond PatchCore. This adaptability enables the framework to be easily applied to different industrial environments and tasks, where the choice of anomaly detection model might vary based on specific requirements or constraints.

Furthemore, as also confirmed by the failure analysis in Section IV-H, a drop in performance can be observed due to the very prominent visual similarity between different defect classes in the selected datasets. In Figure 4, we show

TABLE 6. AUROC at image-level and pixel-level for PatchCore [6] and MambaAD [28] on MVTec-AD and MTD datasets. For all metrics, a higher value is better ↑.

AD	PatchCore [6]		MambaAD [28]	
MVTec-AD	Image-level AUROC↑	Pixel-level AUROC↑	Image-level AUROC↑	Pixel-level AUROC↑
bottle	1.0	0.987	1.00	0.990
cable	0.988	0.985	0.984	0.964
capsule	0.996	0.992	0.936	0.984
hazelnut	1.0	0.989	1.00	0.990
metal nut	0.999	0.987	0.997	0.971
pill	0.966	0.983	0.952	0.971
screw	0.982	0.996	0.948	0.994
toothbrush	0.905	0.988	0.955	0.989
transistor	0.991	0.942	1.00	0.967
zipper	0.992	0.989	0.989	0.982
carpet	0.976	0.991	0.999	0.992
grid	0.998	0.987	1.00	0.992
leather	1.0	0.994	1.00	0.993
tile	1.0	0.963	0.986	0.936
wood	0.987	0.949	0.987	0.945
Average Object	0.982	0.984	0.976	0.980
Average Texture	0.992	0.977	0.994	0.972
Average	0.985	0.982	0.982	0.977
MTD	Image-level AUROC↑	Pixel-level AUROC↑	Image-level AUROC↑	Pixel-level AUROC↑
Average	0.941	0.808	0.958	0.846

TABLE 7. AUROC at image-level and pixel-level for PatchCore [6] on MPDD dataset. For all metrics, a higher value is better ↑.

Dataset	Metrics	
MPDD	Image-level AUROC↑	Pixel-level AUROC↑
Bracket Black	0.957	0.994
Bracket Brown	0.983	0.976
Bracket White	0.984	0.997
Connector	0.974	0.996
Metal Plate	1.0	0.995
Tubes	0.929	0.991
Average	0.971	0.991

TABLE 8. NMI, ARI and F1 scores with standard deviation for semi-supervised and weakly semi-supervised anomaly clustering methods on the MVTec-AD and MTD datasets. For all metrics a higher value is better ↑.

Supervision	Semi-supervised						Weakly semi-supervised					
Dataset	Our-PatchCore			Our-MambaAD			Our-PatchCore			Our-MambaAD		
MVTec-AD	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑
bottle	0.757±0.184	0.688±0.164	0.864±0.112	0.777±0.275	0.722±0.296	0.885±0.185	0.760±0.137	0.673±0.127	0.856±0.086	0.830±0.183	0.785±0.193	0.906±0.137
cable	0.800±0.135	0.819±0.110	0.730±0.089	0.864±0.066	0.906±0.096	0.792±0.219	0.786±0.066	0.761±0.121	0.751±0.143	0.752±0.133	0.672±0.137	0.800±0.156
capsule	0.635±0.227	0.523±0.216	0.692±0.212	0.563±0.160	0.420±0.143	0.572±0.158	0.551±0.199	0.438±0.237	0.621±0.286	0.345±0.091	0.196±0.124	0.494±0.073
hazelnut	0.888±0.305	0.905±0.284	0.925±0.261	0.833±0.194	0.832±0.208	0.855±0.269	0.957±0.137	0.968±0.123	0.977±0.107	0.748±0.130	0.639±0.088	0.762±0.305
metal nut	0.808±0.317	0.724±0.384	0.862±0.333	0.872±0.192	0.806±0.275	0.922±0.185	0.596±0.075	0.370±0.073	0.690±0.080	0.786±0.277	0.736±0.316	0.880±0.224
pill	0.506±0.128	0.316±0.114	0.554±0.268	0.527±0.100	0.365±0.074	0.563±0.171	0.386±0.129	0.196±0.095	0.477±0.232	0.328±0.125	0.142±0.114	0.418±0.094
screw	0.506±0.064	0.514±0.047	0.548±0.150	0.486±0.149	0.472±0.161	0.535±0.145	0.487±0.102	0.346±0.223	0.485±0.221	0.216±0.198	0.131±0.154	0.341±0.127
toothbrush	0.562±0.407	0.560±0.476	0.862±0.273	1.00±0.000	1.00±0.000	1.00±0.000	0.398±0.000	0.313±0.000	0.775±0.000	0.642±0.000	0.714±0.000	0.905±0.000
transistor	0.802±0.170	0.669±0.431	0.758±0.347	0.714±0.116	0.539±0.132	0.633±0.128	0.777±0.169	0.663±0.468	0.713±0.229	0.658±0.255	0.616±0.142	0.695±0.319
zipper	0.781±0.073	0.721±0.158	0.779±0.229	0.801±0.157	0.737±0.177	0.761±0.076	0.750±0.185	0.675±0.178	0.719±0.168	0.716±0.172	0.583±0.186	0.714±0.132
carpet	0.799±0.152	0.766±0.192	0.844±0.219	0.825±0.086	0.813±0.164	0.887±0.183	0.753±0.101	0.557±0.407	0.834±0.389	0.834±0.136	0.816±0.175	0.889±0.172
grid	0.614±0.175	0.523±0.179	0.659±0.135	0.659±0.356	0.630±0.323	0.700±0.349	0.688±0.189	0.641±0.201	0.732±0.167	0.564±0.219	0.533±0.208	0.617±0.220
leather	0.840±0.232	0.818±0.273	0.883±0.248	0.880±0.243	0.857±0.274	0.904±0.258	0.862±0.090	0.843±0.133	0.922±0.091	0.850±0.037	0.833±0.030	0.899±0.022
tile	0.936±0.109	0.932±0.123	0.955±0.062	0.954±0.117	0.953±0.114	0.969±0.108	0.802±0.097	0.681±0.258	0.796±0.265	0.829±0.133	0.764±0.221	0.874±0.213
wood	0.754±0.293	0.651±0.351	0.779±0.341	0.811±0.094	0.739±0.184	0.872±0.167	0.781±0.128	0.668±0.153	0.884±0.153	0.680±0.292	0.566±0.378	0.810±0.246
MTD	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑	NMI↑	ARI↑	F1↑
Average	0.491±0.004	0.451±0.005	0.421±0.024	0.447±0.008	0.338±0.016	0.323±0.019	0.398±0.009	0.486±0.094	0.474±0.027	0.424±0.030	0.526±0.079	0.442±0.023

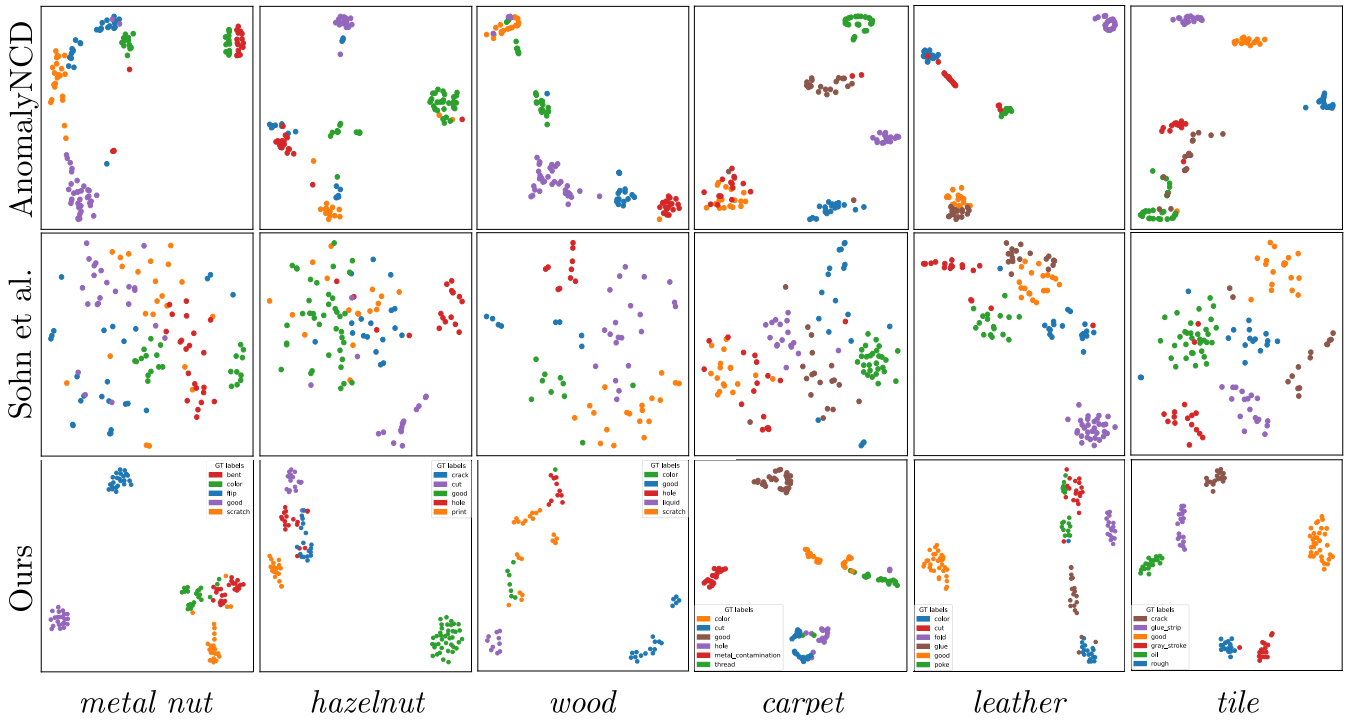

FIGURE 6. Qualitative comparison of different anomaly clustering methods using t-SNE visualizations on the MVTec-AD dataset.

a few examples from the MVTec-AD and MTD datasets, highlighting instances visually similar defect classes in the Hazelnut, Caspule, Carpet, and Magnetic Tile categories. In these case, the model may cluster together samples belonging to disting different defect classes because their visual differences are subtle and difficult to separate based solely on appearance. This makes the formation of well-separated anomaly clusters considerably more challenging. For example, in the Capsule category of MVTec-AD, the anomaly types "Crack" and "Poke" appear visually similar, and our approach may group them into the same cluster.

Finally, from a trustworthiness and ethical perspective, the proposed framework is intended as a decision-support tool for industrial inspection rather than as a fully autonomous replacement for human operators. Its practical reliability may be affected by inaccurate anomaly localization, visually similar defect categories and performance degradation caused by domain shift. For this reason, the method could be applied in a human-in-the-loop deployment, where ambiguous cases can be reviewed by an operator and periodic monitoring can be used to detect drift and trigger recalibration or centroid updates. Demographic fairness is not directly relevant in this setting. Since industrial inspection images may contain

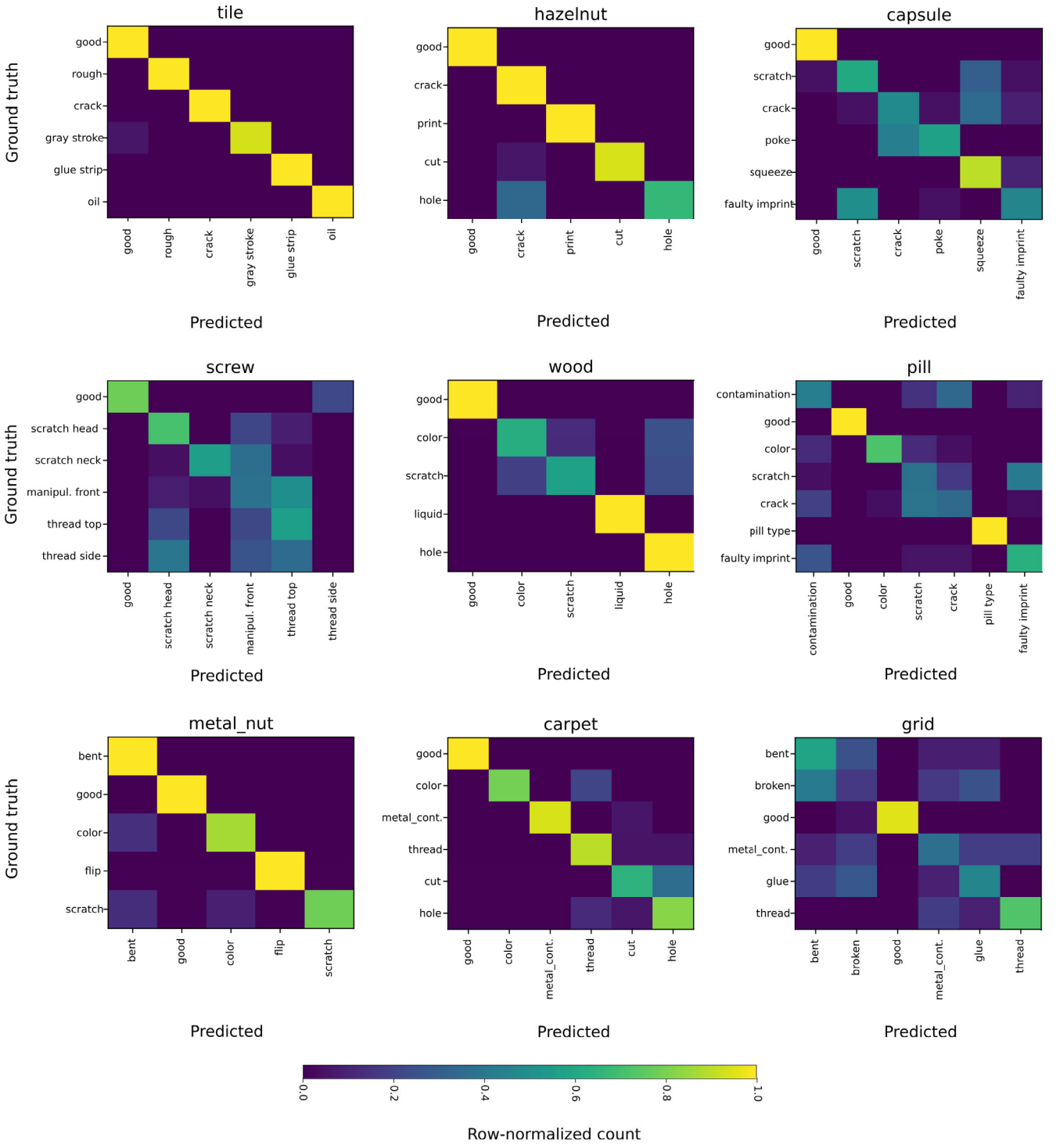

FIGURE 7. Row-normalized confusion matrices for selected MVTec-AD categories after optimal label matching between predicted clusters and ground-truth classes using the Jonker–Volgenant algorithm.

proprietary information, deployment should comply with the applicable access-control and data-protection requirements.

V. CONCLUSION

In this work, we introduced a novel framework for industrial anomaly clustering, leveraging spatial localization and contrastive learning. Our approach achieves state-of-the-art performance on MVTec-AD and MTD datasets, demonstrating its robustness across diverse industrial defects. Our contributions are multi-faceted. First, we propose an efficient method for extracting rich, multi-scale view anomaly representations, enabling robust self-supervised learning

without the need for extensive labeled data. This contribution addresses a critical gap in the industrial anomaly detection landscape, where the scarcity of labeled data often impedes the deployment of supervised models. Second, we present comprehensive experiments on benchmark datasets (MVTec-AD and Magnetic Tile Defect), showing consistent and substantial gains in clustering performance metrics (NMI, ARI, and F1 score) compared to existing methods. Notably, our model maintained high performance even under weakly semi-supervised conditions, underscoring its robustness and applicability in real-world industrial scenarios. Moreover, the flexibility of our framework allows for seamless integration with various anomaly detection backbones beyond PatchCore, expanding its potential use cases in diverse industrial environments.

Our approach not only advances the understanding of anomaly clustering in computer vision but also provides practical insights for improving automated defect analysis in manufacturing contexts.

Future work may explore the adaptation of our framework to different types of anomalies and environments, as well as the application of more sophisticated data augmentation strategies to further enhance model generalization.

APPENDIX
ANOMALY DETECTOR RESULTS

Tables 6 and 7 report the performance results of the anomaly detection models used during the first phase of our proposed work. Specifically, Table 6 presents the AUROC scores at both the image-level and pixel-level for PatchCore [6] and MambaAD [28] on the MVTec-AD and MTD datasets. Table 7, on the other hand, provides the corresponding AUROC results for PatchCore on the MPDD dataset.

APPENDIX
EXTENDED RESULTS

In Table 8, the results of our approach on the MVTec-AD dataset are reported per class, along with the MTD and standard deviation computed over three runs. We compare our method built upon two different anomaly detection backbones: PatchCore [6] and MambaAD [28].

APPENDIX
ADDITIONAL T-SNE VISUALIZATION

In this section, we provide additional qualitative results to further assess the effectiveness of our method. Specifically Figure 6 presents a t-SNE visualization comparison between our approach, Sohn et al. [22] and AnomalyNCD [26] on selected classes on the MVTec-AD dataset. Compared with other methods, our approach exhibits clearer separation and fewer clustering errors between the different defect categories.

APPENDIX
CONFUSION MATRICES

In Figure 7 we provide the confusion matrices for selected MVTec-AD categories. The matrices are row-normalized and highlight, for each ground-truth defect class, how samples are distributed across the matched predicted classes. Strong diagonal responses indicate well-separated clusters, whereas off-diagonal activations reveal the most frequent misclassifications between different defect classes.

REFERENCES

[1] T. D. Tien, A. T. Nguyen, N. H. Tran, T. D. Huy, S. T. M. Duong, C. D. T. Nguyen, and S. Q. H. Truong, "Revisiting reverse distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24511–24520.

[2] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9727–9736.

[3] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, and S. Pan, "Omni-frequency channel-selection representations for unsupervised anomaly detection," *IEEE Trans. Image Process.*, vol. 32, pp. 4327–4340, 2023.

[4] X. Zhang, M. Xu, and X. Zhou, "RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16699–16708.

[5] J. Bae, J.-H. Lee, and S. Kim, "PNI: Industrial anomaly detection using position and neighborhood information," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6350–6360.

[6] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14298–14308.

[7] A. D. Nardin, S. Zottin, E. Colombi, C. Piciarelli, and G. L. Foresti, "Is ImageNet always the best option? An overview on transfer learning strategies for document layout analysis," in *Proc. Int. Conf. Image Anal. Process.*, Jan. 2024, pp. 489–499.

[8] S. Zottin, A. D. Nardin, G. L. Foresti, E. Colombi, and C. Piciarelli, "ICDAR 2024 competition on few-shot and many-shot layout segmentation of ancient manuscripts (SAM)," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2024, pp. 315–331.

[9] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Mach. Intell. Res.*, vol. 21, no. 1, pp. 104–135, Feb. 2024.

[10] Y. Cui, Z. Liu, and S. Lian, "A survey on unsupervised anomaly detection algorithms for industrial images," *IEEE Access*, vol. 11, pp. 55297–55315, 2023.

[11] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang, "Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24490–24499.

[12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, Jul. 2009.

[13] W. Cai and J. Gao, "SSCL: Semi-supervised contrastive learning for industrial anomaly detection," in *Pattern Recognition and Computer Vision*. Singapore: Springer, 2024, pp. 100–112.

[14] Y. Zhao, "LogicAL: Towards logical anomaly synthesis for unsupervised anomaly localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 4022–4031.

[15] Y. Zhao, "OmniAL: A unified CNN framework for unsupervised anomaly localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3924–3933.

[16] C. Lv, Z. Zhang, F. Shen, and F. Zhang, "Unsupervised automatic defect inspection based on image matching and local one-class classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4435–4444.

[17] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19606–19616.

[18] X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma, "PromptAD: Learning prompts with only normal samples for few-shot anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16848–16858.

[19] A. De Nardin, P. Mishra, G. L. Foresti, and C. Piciarelli, "Masked transformer for image anomaly localization," *Int. J. Neural Syst.*, vol. 32, no. 7, Jul. 2022, Art. no. 2250030.

[20] L.-L. Chiu and S.-H. Lai, "Self-supervised normalizing flows for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2927–2936.
[21] A.-T. Ardelean and T. Weyrich, "Blind localization and clustering of anomalies in textures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 3900–3909.
[22] K. Sohn, J. Yoon, C.-L. Li, C.-Y. Lee, and T. Pfister, "Anomaly clustering: Grouping images into coherent clusters of anomaly types," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5479–5490.
[23] C. Niu, H. Shan, and G. Wang, "SPICE: Semantic pseudo-labeling for image clustering," *IEEE Trans. Image Process.*, vol. 31, pp. 7264–7278, 2022.
[24] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "SCAN: Learning to classify images without labels," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 268–285.
[25] C. Niu, J. Zhang, G. Wang, and J. Liang, "GATCluster: Self-supervised Gaussian-attention network for image clustering," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 735–751.
[26] Z. Huang, X. Li, H. Liu, F. Xue, Y. Wang, and Y. Zhou, "AnomalyNCD: Towards novel anomaly class discovery in industrial scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 4755–4765.
[27] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1038–1059, Apr. 2021.
[28] H. He, Y. Bai, J. Zhang, Q. He, H. Chen, Z. Gan, C. Wang, X. Li, G. Tian, and L. Xie, "MambaAD: Exploring state space models for multi-class unsupervised anomaly detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 71162–71187.
[29] F. Bolelli, S. Allegretti, L. Baraldi, and C. Grana, "Spaghetti labeling: Directed acyclic graphs for block-based connected components labeling," *IEEE Trans. Image Process.*, vol. 29, pp. 1999–2012, 2020.
[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[32] K. Sohn, "Improved deep metric learning with multi-class n–pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1849–1857.
[33] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Feb. 1982.
[34] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *Vis. Comput.*, vol. 36, no. 1, pp. 85–96, Jan. 2020.
[35] S. Jezek, M. Jonak, R. Burget, P. Dvorak, and M. Skotak, "Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions," in *Proc. 13th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops (ICUMT)*, Oct. 2021, pp. 66–71.
[36] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971.
[37] D. F. Crouse, "On implementing 2D rectangular assignment algorithms," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 4, pp. 1679–1696, Aug. 2016.
[38] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020.
[39] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.
[40] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2016, pp. 87.1–87.12.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization" *7th International Conference on Learning Representations, ICLR*. New Orleans, LA, USA: OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7
[43] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
[44] J. Zhang, H. He, Z. Gan, Q. He, Y. Cai, Z. Xue, Y. Wang, C. Wang, L. Xie, and Y. Liu, "A comprehensive library for benchmarking multi-class visual anomaly detection," 2024, *arXiv:2406.03262*.

OMAR DE MITRI received the master's degree in computer engineering from the University of Salento, Italy, in 2019, where he is currently pursuing the Ph.D. degree in complex systems engineering. Since 2019, he has been a Research Associate at the Artificial Intelligence and Machine Vision Department, Fraunhofer IPA, Stuttgart, Germany, where he is mainly involved in the research and development of AI-based solutions for optical quality control and automation in industrial contexts. He is a Researcher with expertise in artificial intelligence and computer vision applied to quality control and industrial automation. Over the years, he has collaborated on and coordinated several industrial research projects funded by public and private entities, promoting technology transfer between academic research and companies. He is active in dissemination through seminars and scientific publications in national and international journals and conferences. His research interests include pattern recognition, image synthesis, and the detection of optical anomalies in industrial settings, with a focus on limited data conditions.

COSIMO DISTANTE (Senior Member, IEEE) received the master's degree in computer science from the University of Bari, in 1997, and the Ph.D. degree in materials engineering from the University of Salento, Italy, in 2001. He has been a Visiting Researcher at the Computer Science Department, University of Massachusetts Amherst, MA, USA, where he conducted research in robotics, artificial neural networks, and vision. He joined the University of Massachusetts as a Teaching Assistant of the "artificial intelligence" course for the master's degree in science program, in 1998. In 2001, he joined Italian National Research Council (CNR). Since 2003, he has been a Contract Professor of computer vision, pattern recognition, and image processing within computer engineering at the University of Salento. He is currently a Research Director of CNR and the Unit Manager of the Institute of Applied Sciences and Intelligent Systems. He is also an Expert Member of the Ministry of University and Research (MUR), the Ministry of Enterprises and Made in Italy (MIMIT), and several regional agencies focused on technology and innovation. His research interests include artificial intelligence, specifically in computer vision and pattern recognition, image processing, machine learning, and robotics. He places particular emphasis on developing deep learning algorithms. In 2011, he received the National Innovation Prize from PNI-Cube Telecom Italia. He served as the General Chair for ACIVS2016; and IEEE Advanced Video and Signal-based Surveillance (AVSS), in 2017, 2021, and 2026; and ICIAP2022. He was the Program Chair of the International Conference on Image Processing and Vision Engineering (IMPROVE).

SILVIA ZOTTIN received the B.Sc. degree in computer science from the Ca' Foscari University of Venice, in 2018, and the M.Sc. degree (summa cum laude) in multimedia communication and information technologies (specialization in artificial intelligence and industrial automation) and the Ph.D. degree in computer science and artificial intelligence from the University of Udine, in 2021 and 2026, respectively. She is currently a Postdoctoral Researcher with the Artificial Vision and Machine Learning (AVML) Laboratory, University of Udine. She also collaborates with the Artificial Intelligence for Cultural Heritage (AI4CH) Research Group, University of Udine, particularly focusing on document analysis. She has published several papers in national and international journals and conferences. Her research interests include computer vision, anomaly detection, image segmentation, and machine learning in low-data regimes.

AXEL DE NARDIN received the M.Sc. (summa cum laude) and Ph.D. degrees in computer science, in 2020 and 2024, respectively. Since 2022, he has been a member of the Artificial Intelligence for Cultural Heritage (AI4CH) Research Group, which fosters interdisciplinary collaboration between humanities and computer science experts to develop novel techniques that leverage the use of AI systems for the analysis and preservation of cultural heritage. He is currently an Assistant Professor with the Artificial Vision and Machine Learning (AVML) Laboratory, University of Udine. Since 2021, he has published several papers in international venues, touching a wide range of application fields, such as industrial quality inspection, medical imaging, and cultural heritage. His research interests include semantic segmentation, anomaly detection, and pattern recognition, with a particular focus on low data settings.

GIAN LUCA FORESTI (Senior Member, IEEE) is currently a Full Professor of computer science with the Department of Mathematics, Computer Science, and Physics, University of Udine. He is also the Director of the Artificial Intelligence for Cultural Heritage (AI4CH) Center and the Artificial Vision and Machine Learning Laboratory (AVML). He is the PI of several research projects in the field of AI (machine learning for segmentation of ancient manuscripts), cybersecurity (anomaly detection in computer networks), and computer vision (autonomous and cooperative aerial and underwater systems). He has an H-index of more than 50 with over 10 000 citations. He is an IAPR and AAAI Fellow Member and the National Delegate of the Ministry of Defence in the NATO-STO Information System Technology (IST) Panel. He has been the Finance Chair of the 11th IEEE Conference on Image Processing (ICIP05) and the General Chair of the Eighth IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS11), the 16th International Conference on Image Analysis and Processing (ICIAP11), and the 22nd International Conference on Image Analysis and Processing (ICIAP23).

MARCO F. HUBER (Senior Member, IEEE) received the Diploma, Ph.D., and Habilitation degrees in computer science from Karlsruhe Institute of Technology (KIT), Germany, in 2006, 2009, and 2015, respectively. From June 2009 to May 2011, he was a Group Leader with the Fraunhofer IOSB, Karlsruhe, Germany. Subsequently, he was a Senior Researcher with AGT International, Darmstadt, Germany, until March 2015. From April 2015 to September 2018, he was responsible for product development and data science services of the research division at USU Software AG, Karlsruhe, Germany. At the same time, he was an Adjunct Professor of computer science with KIT. Since October 2018, he has been a Full Professor with the University of Stuttgart. He is currently the Scientific Director of digitalization and artificial intelligence with Fraunhofer IPA, Stuttgart, Germany. His research interests include machine learning, planning and decision making, image processing, and robotics.