Blind Localization and Clustering of Anomalies in Textures

Andrei-Timotei Ardelean   Tim Weyrich

Friedrich-Alexander-Universität Erlangen-Nürnberg

{timotei.ardelean, tim.weyrich}@fau.de

Abstract

Anomaly detection and localization in images is a growing field in computer vision. In this area, a seemingly understudied problem is anomaly clustering, i.e., identifying and grouping different types of anomalies in a fully unsupervised manner. In this work, we propose a novel method for clustering anomalies in largely stationary images (textures) in a blind setting. That is, the input consists of normal and anomalous images without distinction and without labels. What contributes to the difficulty of the task is that anomalous regions are often small and may present only subtle changes in appearance, which can be easily overshadowed by the genuine variance in the texture. Moreover, each anomaly type may have a complex appearance distribution. We introduce a novel scheme for solving this task using a combination of blind anomaly localization and contrastive learning. By identifying the anomalous regions with high fidelity, we can restrict our focus to those regions of interest; then, contrastive learning is employed to increase the separability of different anomaly types and reduce the intra-class variation. Our experiments show that the proposed solution yields significantly better results compared to prior work, setting a new state of the art. Project page: reality.tf.fau.de/pub/ardelean2024blind.html.

1. Introduction

The detection of anomalies is an important problem, found in applications from widely different domains such as medicine, manufacturing, security, finance, etc. The problem is of great interest to the computer vision community as the anomalies are often observed in images or videos. In this work, we focus on identifying structural anomalies, generally found in textured surfaces, rather than generic object anomalies. Depending on the application, there are several tasks covered by the overarching title of anomaly detection, such as image-level detection, pixel-level detection (localization), and clustering of anomalies. These tasks can further be divided depending on the level of supervision that is available.
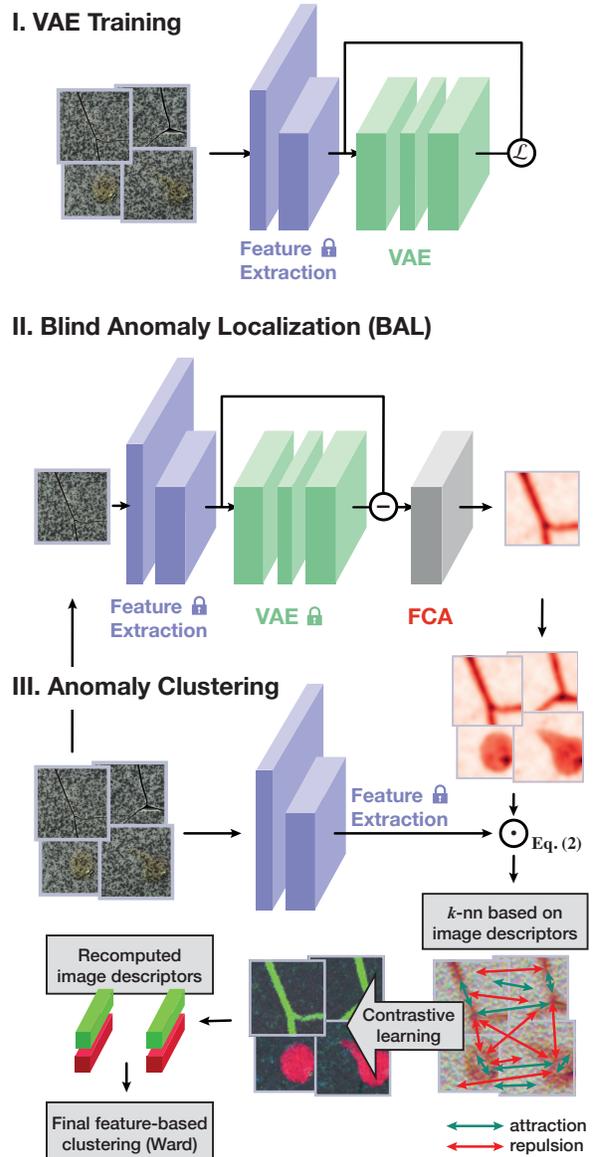

Figure 1. System overview. We train a VAE on features extracted from the input set, and apply FCA [5] on the residuals to obtain anomaly maps in a blind setting. These maps are used to mine positive and negative pairs for contrastive learning. The resulting improved image-level descriptors are then hierarchically clustered.

Traditionally, there are three main categories: supervised, semi-supervised, and unsupervised methods [12, 13]. Supervised methods assume the existence of a training dataset with labeled anomalies, a setting resembling supervised classification with unbalanced classes. Semi-supervised detection of anomalies only requires a curated set of normal samples (no anomalies) for training. Unsupervised methods do not require any labels and do not use a separate training set, but directly detect anomalies based on the data distribution. This setting has also been named **blind anomaly detection** [50] and fully unsupervised learning [19]. Notably, a large number of relevant works [21, 24, 31, 39], including surveys [35, 43], use the term of unsupervised anomaly detection while referring to the second category described above. To avoid the ambiguity regarding these terms, we refer to the training-set-free category as *blind*, following [50], and use the term *normality-supervised* for methods that use a curated set of anomaly-free images.

Anomaly clustering aims to not only detect outliers but also group the anomalies found among a set of elements into semantically coherent clusters. Such a system could greatly improve the interpretability of anomaly detection systems and allow more control over the results; for example, when only certain anomaly types are relevant for the end-application, the others can be filtered out. We address the problem of blind anomaly clustering in images, specifically textures, which are (mostly) stationary. We propose to solve anomaly clustering in two stages. Firstly, the anomalous regions are identified regardless of the type, essentially performing blind anomaly localization (BAL). Secondly, using contrastive learning, we map the original images to a new feature space where different anomaly types can more easily be separated.

In the BAL setting, the input consists of a set of images which may or may not contain anomalies, without any labels. Applying an existing normality-supervised method [33, 39] would be suboptimal due to the (possibly high level of) anomaly-contamination [19, 46]. At the same time, zero-shot (single-image) methods such as FCA [5] and Aota *et al*. [4] would not leverage the information from all images. Therefore, we propose a novel approach specifically for BAL. We harness the benefits of the zero-shot FCA and combine it with a variational autoencoder (VAE) trained on the entire input set, injecting global information in each image.

After localizing the anomalies, one can aggregate the anomalous features and apply a classic clustering algorithm on the pooled descriptors, such as k-means, hierarchical clustering, spectral clustering, etc. However, we find that due to the intrinsic variation of anomalous features, coupled with the imperfect localizations from BAL, the anomaly classes are hard to separate. We address this by performing contrastive learning, mining positive and negative pairs using the computed anomaly maps. This greatly improves the clustering performance and increases its stability.

To summarize, our contributions are as follows:
- We are the first to make the connection between blind anomaly localization (BAL) and anomaly clustering by formulating the problem of anomaly clustering as a two-stage process: BAL, followed by feature fine-tuning.
- We derive a novel solution for BAL, which applies the recent zero-shot method Feature Correspondence Analysis (FCA) [5] to residual maps obtained using a variational autoencoder.
- We propose a way to fine-tune features for anomaly clustering through contrastive learning, leveraging the anomaly maps predicted by BAL.

2. Related Work

Blind anomaly localization and clustering is a complex problem, with strong connections to several directions of research in computer vision. In this section, we briefly revise the tasks adjacent to our problem formulation and how the existing solutions relate to anomaly clustering.

Normality-supervised anomaly detection. Most of the prominent progress in anomaly detection in images consists of normality-supervised methods; *i.e*., methods which rely on a set of normal images, used as reference to compare against at test time. There are several classes of methods that follow this pattern. Reconstruction-based methods generally use autoencoders [8, 47], VAEs [3, 6, 51, 52], or GANs [2, 6, 40] which learn to reproduce normal images. During inference, comparing the reconstruction with the input image should yield higher errors in anomalous regions, since the network was not trained to reconstruct such elements. Another set of methods relies on pretrained feature extractors, modeling normality in different ways including feature banks [39], per-pixel gaussian distributions [21], k-nearest neighbors search [7, 18], etc. Modeling the features from the normal set allows assigning anomaly scores by measuring how well the model explains a new sample. Other approaches have also been proven competitive, such as self-supervised learning (synthesizing artificial anomalies) [33, 49], flow-based models [24], contrastive learning [31], etc. These methods show impressive results on the usual benchmarks; however, it has been shown that in general the performance degrades significantly when the normal set is contaminated (contains some anomalous images) [19, 46, 50].

Blind anomaly detection. The main response to the problem of anomaly contamination for normality-supervised methods is represented by training-set refinement. SRR [46], SROC [19], and STKD [34] present different ways to detect outliers in the anomaly-contaminated training set and eliminate them. The refined set can then be used by a normality-supervised method to build the notion of normality. The results of these approaches have been shown for small amounts of contamination (up to 20%), with degrading

performance for higher factors. On the other hand, our task concerns a generic blind anomaly setting, where the number of anomalous images can be much higher (up to 75% in our experiments). Patel *et al.* [38] developed a solution for such cases, where the ratio of anomalous samples is high. Instead of eliminating the outlying images, only the anomalous regions are masked, allowing the method to leverage the normalcy information available in these images. Another approach explicitly designed for blind anomaly detection was recently proposed by Zhang *et al.* [50]. The method builds on PatchCore [39], addressing the challenge of creating an anomaly-robust feature bank. Our approach takes a different path, *i.e.*, combining variational feature reconstruction with a zero-shot detection method.

Zero-shot anomaly localization A distinct category of supervision in anomaly detection problems consists of zero-shot methods. This scenario assumes that each image is analyzed in isolation; therefore, it heavily relies on priors and/or presumed stationarity of the input image. Strong priors are leveraged by WinCLIP [29], SAA [11], and April-GAN [15] in the form of pretrained visual-language models, using text prompts to identify anomalous regions. On the other hand, the methods of Aota *et al.* [4] and Ardelean and Weyrich (FCA) [5] target textured images. Here, the anomalies are marked as the regions that break the overall homogeneity of the image. More concretely, FCA computes the anomaly score of a pixel by computing its contribution to the Wasserstein distance between the local (patch) and global (image) distributions. These methods can theoretically be directly applied to blind anomaly detection; however, they would not leverage all available information, suggesting they are suboptimal for this task.

Deep Image Clustering Consistently with prior work [32, 42], we assume that each image has only one type of anomaly. This makes the task similar to generic image clustering, tackled by various methods such as IIC [30], GATCluster [36], SCAN [44], and SPICE [37]. Nonetheless, anomaly clustering has two peculiarities which make it more challenging. Firstly, the classes are naturally unbalanced, and we assume no prior knowledge on how rare a specific anomaly type is compared to other types or the normal class. Secondly, the anomalies can be small and subtle, making it difficult to cluster the images when approached holistically [42].

Unsupervised semantic segmentation. The combination of clustering and localization of anomalies can also be seen as an instance of unsupervised semantic segmentation. For example, methods such as PiCIE [17], STEGO [26] and HP [41] identify semantically similar regions in different images. Our anomaly-targeted, margin-based contrastive learning algorithm bears similarities with the correspondence distillation formulation of Hamilton *et al.* [26]. However, the method we propose is adapted to the specifics of anomaly clustering, which gives it a performance edge.

3. Method

We propose a system for clustering anomalies as a two-stage process. The first component performs blind anomaly localization (BAL) by extending the zero-shot feature correspondence analysis (FCA) algorithm of Ardelean and Weyrich [5] to leverage the properties of reconstruction-based anomaly detection methods. The second component refines the texture descriptors used for clustering to make the anomalous classes more separable. Finally, the instances are clustered using classic feature clustering based on image-level embeddings. An overview of the pipeline is presented in Figure 1.

3.1. Blind Anomaly Localization

The task of blind anomaly localization can be formulated as a function that takes as input a set of images $\{I_i\}_{i=1}^{N}$, $I_i \in \mathbb{R}^{H \times W \times 3}$ and returns an anomaly map ($A_i \in \mathbb{R}^{H \times W}$) for each image. Note that the set of images has mixed normal and anomalous images, and no labels are available.

Applying a zero-shot method for each image in isolation may work well for simple, stationary textures, but fails when the normal class is more complex, so that information from multiple images is required to determine anomalies. In order to homogenize the distribution in the normal regions using the available data, we employ a Variational Autoencoder (VAE) trained to reconstruct the input. Instead of operating on images directly, we use a feature extractor F consisting of the first layers of a neural network pretrained on ImageNet [22]; the variational reconstruction is performed on the extracted features. We observe that using 1×1 convolutions is sufficient, as the neural features already capture the local information of the patch in the RGB image; this also allows the network to be trained very fast. At inference, we encode the input features (*i.e.*, $\mu_i, \sigma_i = VAE_E(F(I_i))$) and use the predicted embedding mean (μ_i) for decoding. The residual between the input and the reconstruction is then used as input to FCA using the default parameters ($\sigma_p = 3.0$ and $\sigma_s = 1.0$), yielding the final Equation (1):

$$A_i = FCA(F(I_i) - VAE_D(\mu_i)) . \qquad (1)$$

The union of the two elements is elegant and straightforward, and it works surprisingly well in practice. In the following, we give another intuition for the connection, and we empirically demonstrate this synergy in the experiments section.

A prominent class of methods for normality-supervised anomaly detection is represented by the reconstruction-based approaches, *e.g.*, a Variational Autoencoder trained to reconstruct normal images. At inference time, given an anomalous image, the network will mostly succeed in recovering normal regions. On the other hand, anomalies will not be correctly represented by the network since they were not observed during training. Therefore, the mismatch between the original image and the reconstructions can be used as

an anomaly score. One can simply use the L^1 or L^2 norm of the difference [6, 14], or more involved measures such as SSIM [8], or Rec-grad [51, 52] for obtaining the final anomaly map. As observed in [6, 38], VAE-based methods have a certain natural robustness to anomaly contamination and generally perform better than simpler autoencoders. This is because the *normal* pixels constitute the overwhelming majority, and the regularization of the latent space forces the network to focus on minimizing the error on this majority. While the magnitudes of the residuals between the input and reconstruction can be relatively noisy [8], the structure of the errors provides an additional discriminatory signal for localizing the anomalies. Concretely, we observe that identifying abnormal structures in the feature residuals is equivalent to a zero-shot anomaly localization task where the input consists of the difference between the original features and the reconstruction.

3.2. Clustering

From the previous step, we now have for each image an anomaly map $A_i \in \mathbb{R}^{H \times W}$ and the extracted feature map $F_i = F(I_i), F_i \in \mathbb{R}^{H \times W \times C}$. When performing clustering, we use the same assumption as Sohn *et al.* [42] and consider each image to subsume at most one type of anomaly. In this case, the clustering of anomalies can be made by clustering the respective images in the input set (with one extra cluster for the images that do not present any anomalies). To this end, Sohn *et al.* aggregate the features using the anomaly maps to obtain one descriptor per image and then use agglomerative clustering on the summative descriptors.

Due to the high variability in the appearance of anomalies of the same type, it is hard to separate clusters based on aggregate descriptors in the original space (*i.e.*, the feature space of the pretrained network). Therefore, we propose to fine-tune descriptors by adding an additional 1×1 convolutional head after the layers extracted from the pretrained network F, which are kept frozen. The added convolutions are trained using a contrastive learning regime as depicted in part III of Figure 1.

To obtain positive and negative pairs for contrastive learning, we first binarize the anomaly maps using a threshold t computed based on the distribution of the anomaly scores. In practice, this threshold can be selected using validation images or based on additional information regarding the amount of anomalies expected in the dataset. In our experiments, we used a generic heuristic described in more detail in the supplementary material, Section S4.

For an image i, let $S_i \in \mathbb{R}^{K_i \times C}$ denote the set of anomalous features (where the anomaly value is larger than the threshold), *i.e.*, $S_i = \{H(F_i)^{xy}) \mid A_i^{xy} > t\}$, and $\bar{S}_i \in \mathbb{R}^{(H \cdot W - K_i) \times C}$ the set of normal features. H represents the added trainable layers. One can now form positive pairs from $S_i \times S_i$ and $\bar{S}_i \times \bar{S}_i$, as well as negative pairs from $S_i \times \bar{S}_i$. However, this setup

only ensures consistency between normal and anomalous features at the level of each image and will not preserve the relationship between different types of anomalies. To find positive and negative pairs between different images we first find the k-nearest neighbors of each image based on a cumulative feature, weighted by the anomaly map, calculated as

$$D_i = \sum_{xy} \left(F_i^{xy} \frac{e^{A_i^{xy} \tau^{-1}}}{\sum_{xy} e^{A_i^{xy} \tau^{-1}}} \right). \tag{2}$$

Using the image descriptors $\{D_i\}_{i=1}^N$, we find the k-nearest neighbors $\mathcal{N}(i) \in \mathbb{N}^k$, and we sample k non-neighbors from the bottom 0.5-quantile of the distances: $C(i) \in \mathbb{N}^k$. From these image-level relations, we pool the anomalous and normal features to obtain collective sets of neighboring anomalies $(P_i = S_i \cup \bigcup_{j \in \mathcal{N}(i)} S_j)$, neighboring normals $(\bar{P}_i = \bar{S}_i \cup \bigcup_{j \in \mathcal{N}(i)} \bar{S}_j)$, and non-neighboring anomalies $(C_i = \bigcup_{j \in C(i)} S_j)$. We can now form more suitable positive pairs from $S_i \times P_i$ and $\bar{S}_i \times \bar{P}_i$ and negative pairs from $S_i \times \bar{P}_i$ and $P_i \times C_i$. Finally, the described pairs are processed using the contrastive loss formulation of Hadsell *et al.* [25] to optimize H.

After training H, we obtain clusters by computing (again) image-level descriptors as in Equation 2, but using $H(F_i)$ instead of F_i directly. The descriptors are then processed by an off-the-shelf feature-clustering method; we use agglomerative clustering with Ward linkage in our experiments.

3.3. Implementation Details

The variational autoencoder uses 4 convolutional 1×1 layers to compute μ_i and σ_i in the encoder and contains 3 layers in the decoder. All convolutions are followed by ReLu activations. The number of feature channels in each intermediate layer matches the input (*i.e.*, 512) except for the latent bottleneck where dimension is 128. For each instance, the input features are rescaled between 0 and 1 as in [5]. The VAE is trained for 10k iterations, using AdamW as an optimizer with learning rate of 10^{-4} and 0.1 weight decay; the optimization takes about one minute.

For simplicity, all images are resized to 512×512 before feature extraction. Following [4, 5], we use the output of the second convolutional block of a pretrained WideResnet-50 [48]. The anomaly maps predicted by FCA are smoothed using a Gaussian kernel with $\sigma = 1.0$, and the features are smoothed with $\sigma = 2.0$ and centered to have the mean equal 0 before computing Equation (2). Similar in effect with the average pooling in Sohn *et al.* [42], the blurring reduces high-frequency noise and facilitates clustering. Following [5], we discard the borders of the images as the anomaly scores are not reliable near the edges.

The network H trained using contrastive learning consists of two 1×1 convolutions with 512 channels, followed by L^2 normalization. The layers are trained using AdamW($lr =$

Method	MVTec AD textures			MTD			Leaves		
	NMI	ARI	F_1	NMI	ARI	F_1	NMI	ARI	F_1
SelFormaly [32]	0.743	0.675	0.795	–	–	–	–	–	–
SPICE [37]	0.268	0.079	0.174	0.028	0.014	0.309	0.351	0.272	0.429
SCAN [44]	0.277	0.153	0.335	0.071	0.029	0.282	0.394	0.356	0.526
STEGO [26]	0.389	0.271	0.459	0.056	0.111	0.586	0.055	0.021	0.344
Average [42]	0.273	0.123	0.402	0.065	0.024	0.289	0.341	0.359	0.519
Max Hausdorff [42]	0.625	0.534	0.708	0.193	0.112	0.381	0.546	0.605	0.780
Weighted Avg [42]	0.674	0.601	0.707	0.179	0.120	0.346	0.459	0.462	0.630
Ours	**0.790**	**0.716**	**0.806**	**0.221**	**0.392**	**0.670**	**0.712**	**0.736**	**0.829**

Table 1. Comparison between different methods applied to anomaly clustering. For all metrics a higher value is better.

$5 \cdot 10^{-4}$, *weight decay* = 0.01) for 10 epochs for each texture class. We take $k = 3$ neighbors and use a margin of 0.5.

The algorithm is not particularly sensitive to the aforementioned arguments. The most important parameter of our approach is the temperature τ (see Equation (2)) which is set to 0.002 in our experiments unless noted otherwise. We study the behaviour of this parameter in Section 4.4.

4. Experiments

To validate the utility of our contributions we compare the proposed anomaly clustering system to the current state of the art. We additionally assess the importance of each component through an extensive ablation study.

4.1. Datasets

Since we are targeting anomaly clustering in textures, we use some of the few datasets that contain largely homogeneous materials and present various (labeled) types of anomalies. Note that the anomaly labels are needed for evaluation but never seen during processing. The datasets used are MVTec AD [9, 10], magnetic tile defect (MTD) [27], and coffee leaves dataset (Leaves) [23]. As mentioned before, this work focuses on (mostly) textured images, therefore, for MVTec AD we use the 5 texture classes, each with ∼ 5 anomaly types. MTD contains 1344 images with 5 possible defects, and Leaves includes 422 images covering 4 different kinds of biotic stress. Similarly with Sohn *et al.* [42], we discard the images that contain more than one type of anomaly. For the Leaves dataset we work on full images and do not perform the specific preprocessing outlined by Esgario *et al.* [23].

4.2. Metrics

Normalized mutual information (NMI) is a widely used metric for evaluating clustering results. It is calculated as the mutual information between the predicted clustering and the ground truth labels divided by the arithmetic average of their entropies. We also evaluate the clusters using the adjusted Rand index (ARI) [28] and the F_1 score [16]. Since

the F_1 score requires direct correspondence, we compute the optimal assignment between clusters and labels with the Jonker-Volgenant algorithm [20] from SciPy [45]. For evaluation, we assume the number of clusters to be known; for an extended discussion see Section S5 in the supplementary.

4.3. Results

We compare our system with the state-of-the-art anomaly clustering method of Sohn *et al.* [42], including the introduced variants: Average, Maximum Hausdorff, and Weighted Average. We additionally consider deep image clustering methods (SCAN [44], SPICE [37]) and unsupervised semantic segmentation (STEGO [26]) in the comparison. Recently, SelFormaly [32] introduced a novel approach for solving anomaly detection tasks in a unified manner. While the authors also present results for anomaly clustering, we note that their method works in a normality-supervised setting, having a significant advantage compared to our method and the other baselines.

Table 1 presents our main results on anomaly clustering; the detailed, per-class scores can be found in the supplementary material (Section S2). We obtain a significant improvement in the quality of the clustering over previous methods, agreed by all metrics. Notably, we outperform SelFormaly [32] on MVTec textures despite using a completely unsupervised approach. Please also see Figure 2 for a qualitative assessment. The improvement brought by our system comes from two sources: high-fidelity blind anomaly localization (BAL) and fine-tuned features obtained from contrastive learning. We will analyze these contributions in turn in the following ablative study.

BAL. We compare the proposed combination of FCA and VAE residuals with plain zero-shot FCA [5]. The method of Sohn *et al.* [42] represents the main baseline, using the alpha values before the exponential activation (cf. Equation 5) as anomaly scores. On MVTec AD textures we also compare to a recent dedicated BAL method [50]; only MVTec results are included as the source code is not currently available. To show
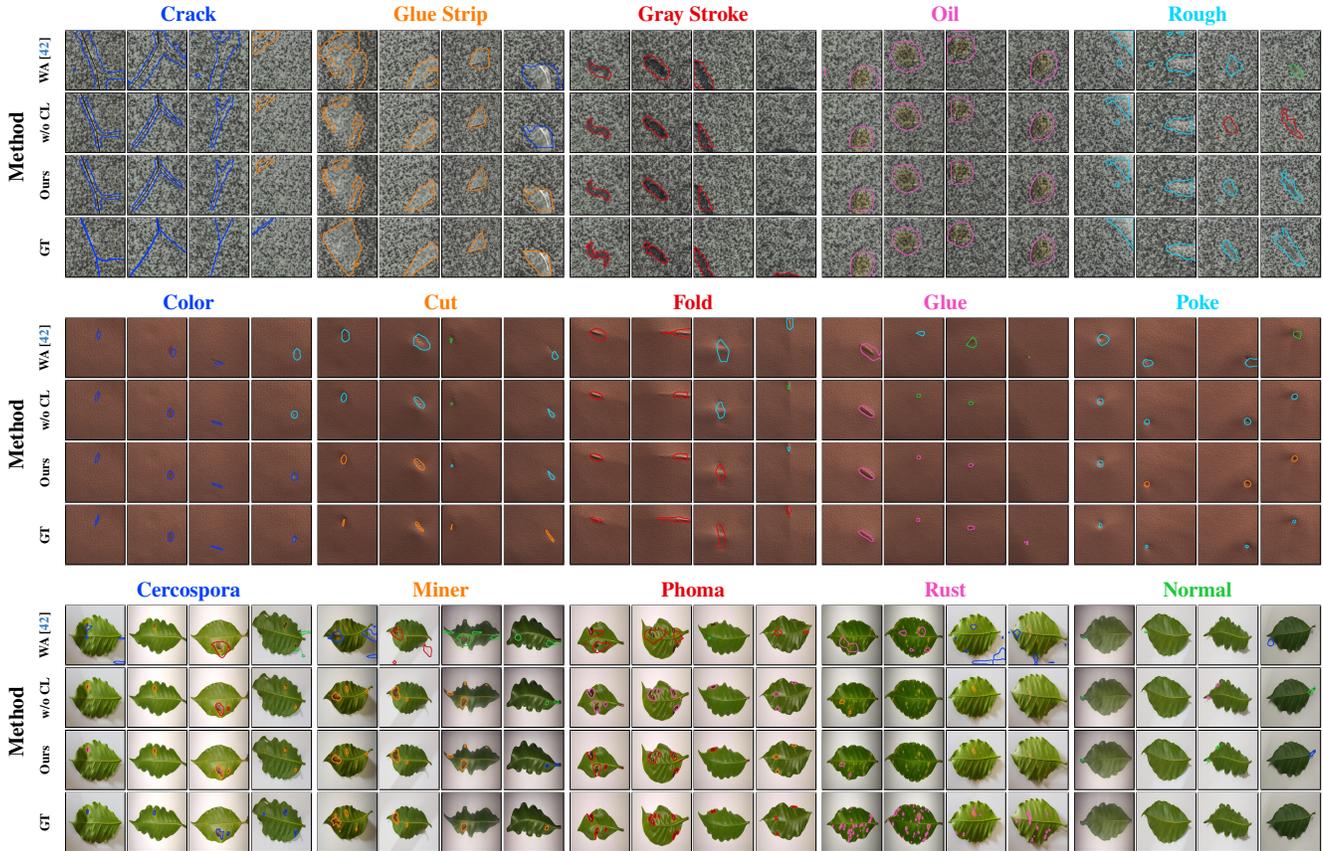
Figure 2. Qualitative evaluation of blind localization and clustering of anomalies. Note that we only show a subset of each cluster; the proportion of misclassifications reflects the true accuracy (rounded down to include more failure cases). The color of a column name indicates the contour color associated with that anomaly type; green is reserved for the normal class (no anomaly). The contours for 'Ours' and '(Ours) w/o CL' differ only through their color because they start from the same binary anomaly localization (our BAL).

the limitations of applying a normality-supervised method on anomaly-contaminated data, we include two prominent methods for anomaly detection: DRAEM [49] and CFA [31]. As this evaluation considers a binary anomaly localization, we use the common metrics: PRO [10] with a false positive rate threshold of 0.3, $AUROC_p$ (pixel level), and $AUROC_i$ (image level). The image-level anomaly score is computed as the maximum over pixel-level scores.

The results in Table 2 suggest that FCA performs better than the soft weights from Sohn *et al.* on mostly homogeneous textures (MVTec, MTD) but the performance degrades on more complex classes (Leaves). Our approach improves upon FCA in all cases and surpasses the baseline [42] by a large margin, especially on the size-sensitive localization metric PRO [10]. We also demonstrate the benefit of our approach qualitatively in Figure 3. Our results show better localization through a reduced number of false positives.

Improved descriptors. Intuitively, a finer anomaly localization should improve the downstream performance of the clustering, since it depends on the image-level descriptors (Equation (2)), computed using the anomaly maps A_i. We

verify this by ablating both the VAE and our contrastive learning (CL). Ablating CL amounts to using the descriptors D_i directly as instance embeddings, similarly to the weighted aggregation from Sohn *et al.* [42]. VAE ablation is performed by applying contrastive learning using the anomaly scores obtained from zero-shot FCA directly.

Table 3 shows that both components contribute to the performance of our approach. As the textures in MVTec AD are more homogeneous, the benefit of our VAE-based BAL is smaller, and it does not improve the downstream clustering performance on this dataset. Nevertheless, its significance is clearly observed in the results on MTD and Leaves. To show that our improvements are orthogonal to the final feature-clustering method used, we also present this ablation in the supplementary material (Table 6) when using k-means instead of Ward clustering.

We visualize the effect of our anomaly-targeted contrastive learning in Figure 4.

Feature-clustering methods. The last step of our algorithm consists in an off-the-shelf clustering method based on image-level descriptors. It can be observed in Table 4

MVTec AD textures	PRO	AUROC$_p$	AUROC$_i$
ILTM[†] [38]	79.04	85.38	–
DRAEM [49]	31.12	57.08	73.04
CFA [31]	92.51	97.27	88.54
Sohn *et al.* [42]	90.84	96.61	98.69
Zhang *et al.* [50]	91.26	96.80	97.58
FCA [5]	96.92	98.74	99.85
FCA+VAE (ours)	**97.50**	**99.02**	**99.93**

MTD	PRO	AUROC$_p$	AUROC$_i$
DRAEM [49]	22.80	54.22	54.98
CFA [31]	65.27	73.86	62.49
Sohn *et al.* [42]	66.51	72.86	77.00
FCA [5]	72.15	74.48	80.29
FCA+VAE (ours)	**75.53**	**75.32**	**83.87**

Leaves	PRO	AUROC$_p$	AUROC$_i$
DRAEM [49]	33.53	76.91	51.34
CFA [31]	76.92	97.15	81.51
Sohn *et al.* [42]	75.25	97.15	**92.02**
FCA [5]	48.10	88.50	62.05
FCA+VAE (ours)	**77.20**	**97.62**	90.62

Table 2. Binary anomaly detection and localization metrics in a *blind* setting (mixed normal and anomalous data with no annotations). For all metrics a higher value is better.

[†] ILTM [38] results differ slightly in the evaluation protocol.

that agglomerative/hierarchical clustering with Ward linkage performs best, with k-means being a close second. Generally, the results obtained with Ward clustering are better compared to k-means over different methods [42] and parameters. However, as we show in Section 4.4, the variance is also larger, with small differences in features yielding considerable different clusterings. Ideally, the features optimized by our method should make the different anomaly types more easily separable so that the clustering algorithm used would have limited influence on the results. Indeed, in Figure 5 we show that our contrastive learning has the added benefit of making the algorithm more stable.

Speed assessment Despite being composed of several components and requiring two stages of neural network training, our method is relatively fast. Concretely, for the MTD dataset which contains 1344 images, the entire process takes a total of 14 minutes (7.5% for VAE training, 66.5% for contrastive learning, and 26% for FCA, image-level descriptors calculations and agglomerative clustering). For reference, the method of Sohn *et al.* [42] takes 30 minutes due to a time complexity quadratic in the number of images.

Application domain We design our approach specifically for textures, a significant category which occurs often in civil-
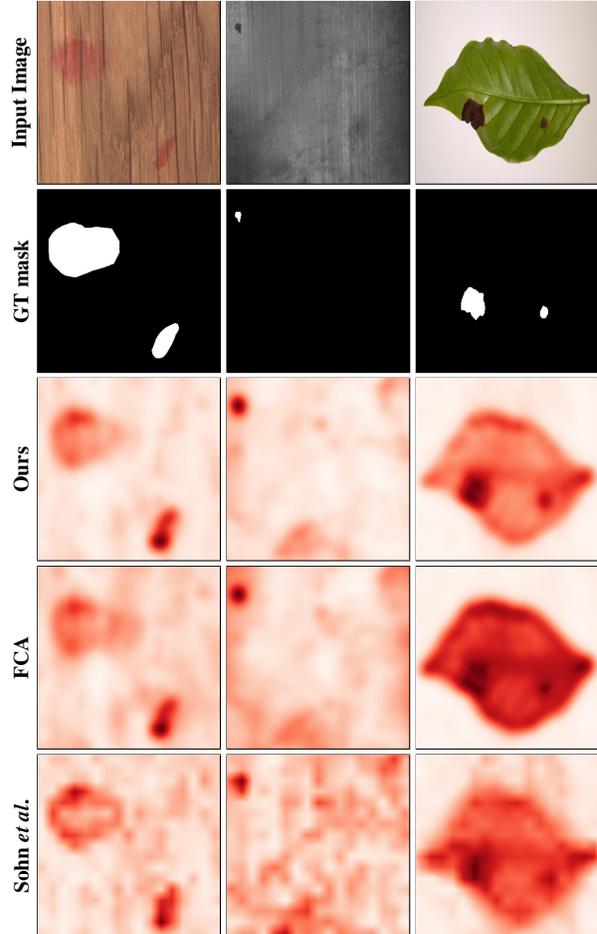

Figure 3. Qualitative comparison of anomaly maps produced by different approaches to BAL. The images' borders are cropped out.

structures or industrial inspection. For an extended discussion on the scope of our method see S7 in the supplementary.
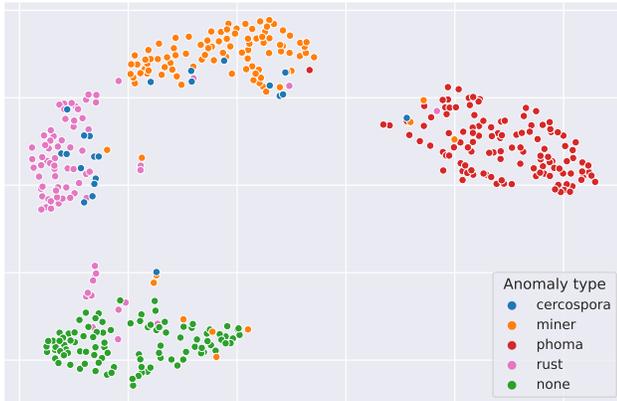
4.4. Sensitivity Analysis

We analyze the performance of our algorithm under different values of τ (see Equation (2)), the single most important parameter of the method. A small value of τ focuses the weights on a single element in the feature maps, the one with the highest anomaly score. On the other hand, a large τ will push the weights toward a uniform distribution, minimizing the effect of using anomaly scores to focus on the regions of interest. Figure 5 shows that different datasets have different optimal values for τ. Nevertheless, the range $1.5 \cdot 10^{-3}$ to $3.0 \cdot 10^{-3}$ seems broadly acceptable for all datasets.

We note here that the improvement added by contrastive learning is visible for various values of τ. Furthermore, the variance of the NMI metric is greatly reduced, as well as the gap between agglomerative (Ward) clustering and k-means, which speaks to the stability and robustness of our algorithm.

Method	MVTec AD textures			MTD			Leaves		
	NMI	ARI	F_1	NMI	ARI	F_1	NMI	ARI	F_1
Ours w/o VAE, CL	0.756	0.682	0.781	0.176	0.166	0.507	0.333	0.235	0.443
Ours w/o VAE	**0.807**	**0.733**	**0.811**	0.167	0.214	0.539	0.528	0.507	0.706
Ours w/o CL	0.769	0.693	0.775	0.142	0.161	0.513	0.512	0.461	0.588
Ours	0.790	0.716	0.806	**0.221**	**0.392**	**0.670**	**0.712**	**0.736**	**0.829**

Table 3. Ablation study of the improved anomaly localization (VAE) and contrastive learning (CL). Details in the main text.

NMI	MVTec	MTD	Leaves
k-means	0.778	0.211	0.689
Agglomerative-Ward	**0.790**	**0.221**	0.712
Gaussian Mixture	0.750	0.157	**0.724**
Spectral	0.757	0.189	0.629

Table 4. Comparison in terms of NMI of different feature-clustering methods applied to the descriptors produced by our method.

(a) Descriptors with contrastive learning

(b) Descriptors without contrastive learning

Figure 4. Visualization of the initial image-level descriptors compared to the descriptors computed after contrastive learning. The vectors are projected to two dimensions using t-SNE.
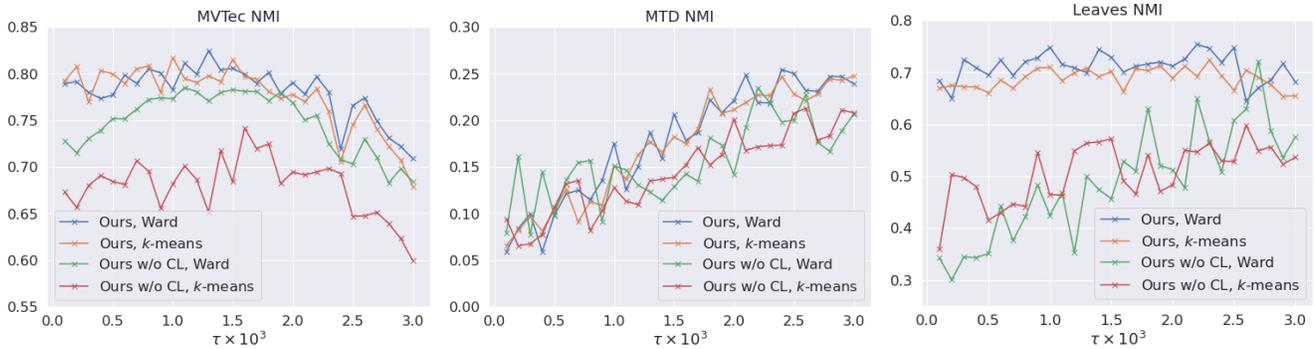

Figure 5. Analysis of clustering performance in terms of NMI for varying values of τ. Contrastive learning improves stability.

5. Conclusion

In this work, we introduce a novel pipeline for clustering anomalies in textured surfaces. Firstly, we extend the state-of-the-art zero-shot method of Ardelean and Weyrich [5] for anomaly localization, and enable incorporating additional information from multiple images. This vastly improves the performance on more complex, non-stationary images, and establishes a new state of the art for the blind anomaly localization task. Secondly, we leverage the predicted anomaly maps to mine positive and negative pairs and employ contrastive learning. This improves the descriptiveness of the neural features, increasing the separability of the anomaly classes. We validate the contributions extensively by comparing with prior art on three different datasets, and we include a comprehensive ablation study that clarifies the role of each component of our method.

References

[1] Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A deep learning library for anomaly detection. In *IEEE International Conference on Image Processing*, pages 1706–1710. IEEE, 2022. 3

[2] Samet Akçay, Amir Atapour-Abarghouei, and T. Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. *International Joint Conference on Neural Networks*, pages 1–8, 2019. 2

[3] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015. 2

[4] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023. 2, 3, 4

[5] Andrei-Timotei Ardelean and Tim Weyrich. High-fidelity zero-shot texture anomaly localization using feature correspondence analysis. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1, 2, 3, 4, 5, 7, 8

[6] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *BrainLes@MICCAI*, 2018. 2, 4, 1

[7] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020. 2

[8] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2018. 2, 4, 1

[9] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019. 5

[10] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 5, 6, 1, 2, 3

[11] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023. 3

[12] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. 2

[13] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 2

[14] Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018. 4, 1

[15] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 3

[16] N Chinchor. Muc-4 evaluation metrics in proc. of the fourth message understanding conference 22–29, 1992. 5

[17] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2021. 3

[18] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2

[19] Antoine Cordier, Benjamin Missaoui, and Pierre Gutierrez. Data refinement for fully unsupervised visual inspection using pre-trained networks. *arXiv preprint arXiv:2202.12759*, 2022. 2

[20] David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. 5

[21] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event Proceedings, Part IV*, pages 475–489. Springer, 2021. 2

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[23] José G.M. Esgario, Renato A. Krohling, and José A. Ventura. Deep learning for classification and severity estimation of coffee leaf biotic stress. *Computers and Electronics in Agriculture*, 169:105162, 2020. 5

[24] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 2

[25] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006. 4

[26] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022. 3, 5, 2

[27] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36:85–96, 2020. 5

[28] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985. 5

[29] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 3

[30] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *International Conference on Computer Vision*, pages 9865–9874, 2019. 3

[31] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454, 2022. 2, 6, 7, 3

[32] Yujin Lee, Harin Lim, and Hyunsoo Yoon. Selformaly: Towards task-agnostic unified anomaly detection. *arXiv preprint arXiv:2307.12540*, 2023. 3, 5

[33] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 2

[34] Hongbo Liu, Kai Li, Xiu Li, and Yulun Zhang. Unsupervised anomaly detection with self-training and knowledge distillation. In *IEEE International Conference on Image Processing*, pages 2102–2106. IEEE, 2022. 2

[35] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024. 2

[36] Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *European Conference on Computer Vision*, pages 735–751. Springer, 2020. 3

[37] Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022. 3, 5, 2

[38] Ashay Patel, Petru-Daniel Tudosiu, Walter HL Pinaya, Mark S Graham, Olusola Adeleke, Gary Cook, Vicky Goh, Sebastien Ourselin, and M Jorge Cardoso. Self-supervised anomaly detection from anomalous training data via iterative latent token masking. In *International Conference on Computer Vision Workshops*, pages 2402–2410, 2023. 3, 4, 7

[39] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 3

[40] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Margarethe Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging*, 2017. 2

[41] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19540–19549, 2023. 3

[42] Kihyuk Sohn, Jinsung Yoon, Chun-Liang Li, Chen-Yu Lee, and Tomas Pfister. Anomaly clustering: Grouping images into coherent clusters of anomaly types. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5479–5490, 2023. 3, 4, 5, 6, 7, 2

[43] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 2022. 2

[44] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285. Springer, 2020. 3, 5

[45] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 5

[46] Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O Arik, Chen-Yu Lee, and Tomas Pfister. Self-supervise, refine, repeat: Improving unsupervised anomaly detection. *Transactions on Machine Learning Research*, 2022. 2

[47] Sanyapong Youkachen, Miti Ruchanurucks, Teera Phatrapomnant, and Hirohiko Kaneko. Defect segmentation of hot-rolled steel strip surface by using convolutional auto-encoder and conventional image processing. *International Conference of Information and Communication Technology for Embedded Systems*, pages 1–5, 2019. 2

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pages 87.1–87.12. BMVA Press, 2016. 4, 3

[49] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *International Conference on Computer Vision*, pages 8330–8339, 2021. 2, 6, 7, 3

[50] Jie Zhang, Masanori Suganuma, and Takayuki Okatani. That's bad: blind anomaly detection by implicit local feature clustering. *Machine Vision and Applications*, 35(2):31, 2024. 2, 3, 5, 7

[51] Leixin Zhou, Wenxiang Deng, and Xiaodong Wu. Unsupervised anomaly localization using vae and beta-vae. *arXiv preprint arXiv:2005.10686*, 2020. 2, 4, 1

[52] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *Medical Image Computing and Computer Assisted Intervention*, pages 289–297. Springer, 2019. 2, 4, 1