High-Fidelity Zero-Shot Texture Anomaly Localization Using Feature Correspondence Analysis

Andrei-Timotei Ardelean Tim Weyrich
Friedrich-Alexander-Universität Erlangen-Nürnberg
{timotei.ardelean, tim.weyrich}@fau.de

Abstract

We propose a novel method for Zero-Shot Anomaly Localization on textures. The task refers to identifying abnormal regions in an otherwise homogeneous image. To obtain a high-fidelity localization, we leverage a bijective mapping derived from the 1-dimensional Wasserstein Distance. As opposed to using holistic distances between distributions, the proposed approach allows pinpointing the non-conformity of a pixel in a local context with increased precision. By aggregating the contribution of the pixel to the errors of all nearby patches, we obtain a reliable anomaly score estimate. We validate our solution on several datasets and obtain more than a 40% reduction in error over the previous state of the art on the MVTec AD dataset in a zero-shot setting. Also see reality.tf.fau.de/pub/ardelean2024highfidelity.html.

1. Introduction

Anomaly Detection (AD) refers to discerning between elements that abide by a standard of normality and those which do not. Humans are generally able to perform this distinction without the need for an explicit guideline for the standard of normality simply by comparing them to items that agree to the standard [50]. Even further, we can often find anomalous regions from visual imagery without previous knowledge of how a certain object or material should look, by simply pinpointing what stands out in a single, isolated sample [37]. This motivates the search for an automatic system able to perform this task, *i.e.*, zero-shot anomaly localization (ZSAL).

Anomaly detection and localization has a wide range of applications. Automatically finding defects during manufacturing, identifying forgeries, detecting situations that require attention in medical imaging, and discovering inaccuracies in industrial machines are just a few of the domains where an anomaly detection system could bring considerable benefits.

The computer vision community has lately shown increased interest in solving the problem of anomaly detection and localization, encouraged by the success of deep learning methods on various tasks [15]. The primary employed strategy is unsupervised learning, modeling normality from a collection of unblemished items. This removes the need for labeled anomalous data at training time, which can be difficult to acquire; however, most current systems still require numerous curated (normal) samples [22, 33, 36, 42, 51, 56]. To alleviate this requirement, the more challenging task of few-shot [28, 43, 46] and zero-shot [4, 29, 45] AL has recently started to be addressed.

Figure 1. Anomaly localization example. *Left:* input texture; *right:* predicted anomaly map.

We develop a new system designed specifically for anomaly localization that works in a zero-shot setting, identifying the parts that break the homogeneity of a single textured sample (Figure 1). Our main contribution is a novel method for comparing the statistics between different patches in an image or feature map. To quantify the normality of a pixel location one could trivially compute the average of the nearby features and compare them to a global descriptor, however, as we show, the errors obtained by this method are too coarse for a pixel-level localization of anomalies. We analyze different methods for comparing the local statistics of a patch to a (global) reference and show that one can use a bidirectional mapping that implicitly results from the Wasserstein distance to more precisely identify the offending pixels. This insight is the core of our Feature Correspondence Analysis (FCA).

2. Related Work

The problem of anomaly detection can be posed for various types of data such as weather records [52], stock market and financial transactions [1], acoustic monitoring [21], video

surveillance [31], medical imaging [25], manufacturing inspection [27, 35], etc. In this work, we address the detection of anomalies in images, more exactly detecting anomalous regions in otherwise homogeneous or stationary textures. This can be formulated as a multi-class segmentation and classification of anomalous pixels [13, 39], or in a simpler setting, as a binary separation between normal and anomalous regions [20, 22, 34, 43, 53, 56]. We focus on the latter, usually referred to as anomaly localization (AL) despite dealing with pixel-level segmentation (as opposed to localization understood in the context of object detection). AL can be considered a superclass of the anomaly detection task/classification over images, as an image label can be simply computed as the maximum of pixel-wise anomaly predictions [9]. Therefore, AL is more challenging and, for most purposes, more useful compared to image-level classification, making the result explainable and actionable [50]. In the remainder of this section, we briefly address the most relevant methods and refer readers to a survey [35, 50] for a broader insight into anomaly localization literature.

Reconstruction-based methods. Most of the early machine learning methods for anomaly detection are reconstruction-based [50], using a (variational) autoencoder [3, 12, 54], or a generative adversarial network (GAN) [2, 6, 44] to learn to synthesize normal images. At inference, reconstruction errors reveal anomalies. These methods are intuitive; however, they do not incorporate any priors on real images (*e.g.* by pretraining), which makes them dependent on a large set of normal samples. Conversely, our method identifies anomalies with *zero* normal exemplars.

Deep features-based methods. The leading approaches in recent years belong to the class of deep feature-based methods. In essence, these methods leverage features extracted with the help of a larger network, pretrained on vast amounts of data, that serves as a prior. These embeddings have been used in various ways such as taking the k-nearest neighbors at image (DN2 [8]) or sub-image level (SPADE [20]), creating a Teacher-Student feature-reconstruction framework [11], modeling the distribution of features that characterizes each pixel location as a multivariate Gaussian (PaDiM [22]), creating a memory bank of feature patches as a representation of normality (PatchCore [42]), etc. The intuition is that the features from the intermediate layers of a CNN trained on ImageNet [23] capture higher level semantics that can be used to identify anomalies. As we show, our approach also benefits from using deep features.

Few-shot methods. Few-shot anomaly detection was recently explicitly addressed with approaches such as normalizing flows [43], hierarchical generative models [46], and feature registration [28]. These methods, however, rely on data augmentation which is problem-specific and may require domain knowledge. Moreover, as observed in [4], they are not significantly better compared to, for example, PatchCore [42] which scales better with the number of samples. The authors of PatchCore even address the concern regarding the performance in a limited normal data setting and shows better results compared to SPADE [20] and PaDiM [22]. Notably, we are interested in the more extreme situation where not a single *normal* image is provided.

Zero-shot methods. Zero-shot anomaly localization considers the case of anomaly detection where the anomalous regions are segmented without a set of unblemished textures to act as guidance. MAEDAY [45] introduces for the first time the task of zero-shot anomaly detection. The method pretrains a transformer-based network which is used to reconstruct a partially masked image at inference. By using this in-painting network, an anomaly score can be computed by identifying the differences between the unmasked image and the reconstructed output. WinCLIP [29] introduced a new paradigm for ZSAL using a vision-language foundation model (CLIP [41]) which quickly gained traction [5, 14, 16]. These methods use text prompts to discriminate between normal and anomalous patches relying on the capacity of the multimodal foundation model to learn this distinction through large-scale training. Aota *et al*. [4] developed a method for zero-shot anomaly detection and localization specifically for textures. For each pixel, the local features are averaged and compared to the k-nearest neighbors in the same image. Our method is most similar to the latter as it compares local features with globally aggregated information, and it is designed to work on textures and not generic objects as [45] and [29]. Academic works that explicitly solve ZSAL only recently emerged; however, the task bears similarities to texture perception [30], image saliency [17], texture stationarity analysis [38], and weathering estimation [7].

3. Algorithm Design

This section describes the design decisions that went into building our method. We analyze how different components of a zero-shot patch-based anomaly localization system affect its performance, and we also introduce a novel procedure for estimating the anomaly degree at each spatial location.

We consider the following attributes of an AL method, identified as desirable: *high sensitivity at high specificity*, *ability to scale to higher resolutions*, and *fast running time*. Importantly, we focus on a zero-shot scenario, and we are mainly interested in textures, which are largely homogeneous, save for the anomalous regions themselves.

As a generic framework for zero-shot anomaly localization, we propose the following self-similarity formulation to obtain the anomaly map A from an image I:

$$A(x, y; F, S, R) = \sum_{F_r \in R(F(I))} S(x, y, F(I), F_r) \ . \tag{1}$$

Such an AL system is defined by three different components: feature extraction (F), patch statistics comparison (S), and

PRO ↑ / AUROC ↑	Colors	RandProj	Steerable	TEM	VGG
Moments	46.51 / 75.62	40.66 / 73.33	64.21 / 80.78	53.97 / 75.64	61.96 / 83.82
Histogram	50.43 / 77.80	53.74 / 80.48	70.64 / 84.43	68.47 / 84.71	73.17 / 88.44
SWW	58.62 / 83.62	62.21 / 85.89	**73.08 / 87.77**	74.48 / 89.36	77.40 / 91.44
FCA (ours)	**63.30 / 85.76**	**66.28 / 87.62**	71.75 / 86.99	**75.33 / 90.18**	**81.08 / 92.58**

Table 1. Preliminary experiment, comparing our patch statistics method to different baselines. Compared in terms of two metrics: PRO(0.3) and AUROC. The best results are highlighted in bold.

reference selection (R). Simply put, the anomaly score A at location (x, y) is computed as the sum of the costs when comparing features within one or more patches containing (x, y) with a set of references F_r. We note that the proposed definition is a superset of the discrete form of the stationarity measure introduced in [38]. While not explicitly designed for anomaly localization, by isolating the influence of each spatial location in the stationarity measure from [38], one can use it as an anomaly localization score. The main difference is that Moritz *et al*. assume the reference set R consists of all patches in $F(I)$, which, as we show, is suboptimal.

3.1. Feature Extraction

We evaluate the effect of different feature extractors $F(I) \rightarrow \mathbb{R}^{H \times W \times C}$ and confirm the findings of previous work that pretrained neural networks provide useful features for AL. Table 1 compares five feature extraction functions F. The metrics used for evaluation are detailed in Section 4. We consider using the colors directly ($F(I) = I$), convolving the image I with a set of random kernels, Steerable Filters [26], Laws' texture energy measure (TEM [32]), and neural features from a simple pretrained VGG19 network [48]. In this preliminary experiment, all feature extractors operate on a single resolution and have a small receptive field, *i.e*., the images are scaled to 256×256, and the feature maps have the same resolution, with $C \approx 128$ channels (except for colors, where $C = 3$). The random projections are inspired by [24], where they are used in the context of the Sliced Wasserstein Distance, and consist of normalized random 5×5 kernels; we use only one level of steerable filters (full spatial resolution); finally, we use the concatenated output of the first two convolutional layers of the VGG network, having an effective receptive field of 5×5. We use a patch size of 25×25 for all stationarity measures, which is large enough to capture the difference in appearance between normal and anomalous regions.

As shown in Table 1, the embeddings obtained from the VGG network consistently outperform other types of features, including traditional texture analysis methods [26, 32].

3.2. Patch Statistics Comparison

The function $S(x, y, F(I), F_r)$ evaluates the degree of anomaly at position (x, y) given its local context in the feature maps $F(I)$, by comparing with the reference F_r. The function should analyze how do the local statistics around (x, y) differ from the statistics in F_r. In this subsection we describe different options for S, together with their limitations, and introduce our Feature Correspondence Analysis (FCA) method for comparing patch statistics.

Moments. In general, only a small region around a certain location is needed to identify an anomaly. This leads to a trivial patch statistics comparison method, computed by averaging the features around (x, y), *i.e*.,

$$S(x, y) = \left\| \frac{1}{T^2} \sum_{(x', y') \in P_{xy}} F(I)(x', y') - \text{avg}(F_r) \right\|_2^2 , \quad (2)$$

where $F(I)$ and F_r have been omitted from S for brevity, and P_{xy} denotes a patch of size $T \times T$ centered in (x, y). The definition can be easily extended to include spatial weighting (*e.g*., Gaussian) and moments of higher order, becoming equivalent to the method of moments from [38] when using RGB colors directly as features.

Histogram. Moritz *et al*. [38] propose another two options for computing the stationarity measure, which can be described in our conceptual framework as using a histogram-based patch statistics comparison over RGB colors, and steerable filters, respectively. The histogram-based algorithm can be described as:

$$S(x, y) = \text{hist}\Big(\bigcup_{(x', y') \in P_{xy}} F(I)(x', y') \Big) \ominus \text{hist}(F_r) , \quad (3)$$

where $\ominus$ gives the earth mover's (Wasserstein) distance between the two histograms. As in the case of moments, when computing the histogram one can employ spatial weighting to increase the importance of the pixels closer to (x, y).

Sample-weighted Wasserstein (SWW). The previous methods have limited expressive powers, specifically because they consider the distribution inside a patch as a whole, unable to pinpoint "outlieredness" of individual samples.

That ability conveniently occurs in an efficient implementation of the 1-D Wasserstein distance when operating on individual samples drawn from distributions. If two sets of samples have the same size, the Wasserstein distance can be obtained by sorting the samples and then summing over the absolute differences between the elements corresponding to the same rank [24]. That comparison of samples of the same rank within a sorting can be seen as a bijective mapping

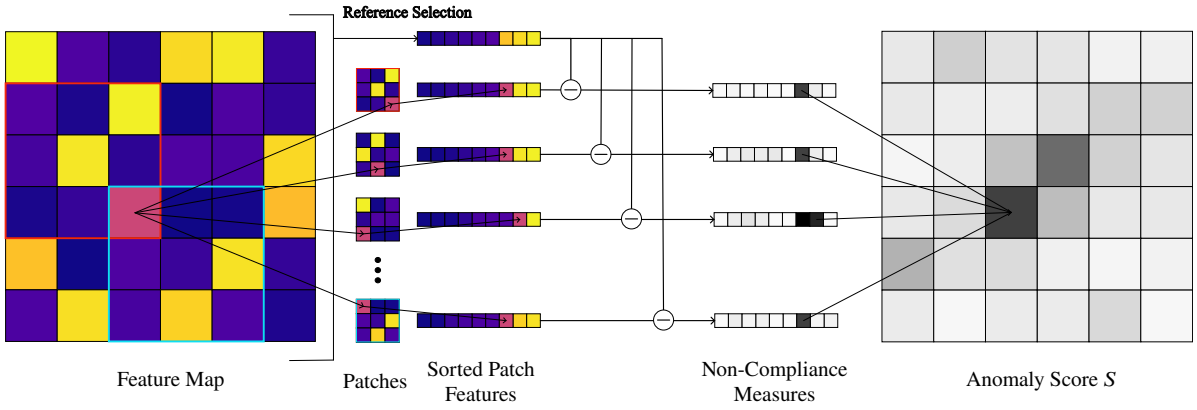

Figure 2. Depiction of our Feature Correspondence Analysis (FCA). All patches surrounding a pixel are compared against the *reference*. The correspondence errors of the pixel in all contexts are aggregated to obtain the final anomaly score. The $\ominus$ denotes the absolute difference.

between two sample sets, and the difference between corresponding samples is an immediate measure for those samples' non-compliance with the respective other distribution. That resulting non-compliance calculation translates directly to a coarse anomaly measure of the corresponding feature channel for each pixel; summing them across channels results in an error score $M(x, y; P)$ for each pixel (x, y) in a patch P.

Subsequently, we aggregate the per-patch error map M into an anomaly measure for the center of the patch P. At this point, averaging $M(\cdot, \cdot; P)$ would yield the exact Wasserstein distance, and it would be equivalent to the previously defined histogram method (with bins $\to \infty$). Instead, we use a Gaussian-weighted average to increase the spatial sensitivity of the resulting anomaly score $S(x, y)$.

Upon cursory observation, this may resemble the weighted, sliding-window histogram calculations of Moritz *et al*. [38]; however, Moritz *et al*. compute weighted distributions before calculating their metric, whereas we preserve the original patch distribution but weight the influence of each sample's non-compliance score on the final anomaly score. The equation for this sample-weighted Wasserstein method is:

$$S(x, y) = \sum_{(x', y') \in P_{xy}} M(x', y'; P_{xy}) G_{\sigma_w}(x' - x, y' - y) \ , \qquad (4)$$

where $G_{\sigma_w}(\Delta_x, \Delta_y)$ is a spatial weighting function, for which we use a Gaussian with variance σ_w^2. Note that we introduce this method as a conceptual bridge between comparing histograms and our FCA.

FCA. The previous definition of SWW allows us to separate the context size and the amount of smoothing in the aggregation through the parameter of the Gaussian; however, the final anomaly score for any location (x, y) uses as context only the patch P_{xy}. We further leverage the bijective mapping from SWW by computing the anomaly score at location (x, y) as the sum of the matching errors for position (x, y) in the context of all surrounding patches, which gives:

$$S(x, y) = \sum_{(x', y') \in P_{xy}} M(x, y; P_{x'y'}) G_{\sigma_p}(x' - x, y' - y) \ . \qquad (5)$$

Please note the change of parameters in M compared to the SWW equation (4). The main difference is that instead of considering one context patch P_{xy} when computing $S(x, y)$, we consider all patches that contain (x, y), and aggregate the contribution of the location (x, y) in all of these contexts. Anomalies are generally considered smooth and all available datasets present anomalies as binary blobs that mark anomalous regions, rather than continuous scores depicting the contribution of each pixel to the anomaly. To attend to this, we introduce Gaussian smoothing $\mathcal{G}_{\sigma_s}$ after matching errors, yielding the final formula:

$$S(x, y) = \sum_{(x', y') \in P_{xy}} \mathcal{G}_{\sigma_s}\big(M(\cdot, \cdot; P_{x'y'})\big)(x, y)\, G_{\sigma_p}(x' - x, y' - y) \ . \qquad (6)$$

The workflow of the algorithm is illustrated in Figure 2. We name this novel method Feature Correspondence Analysis (FCA), as it computes the anomaly score based on the correspondence of features from patches to a reference.

In Figure 3, we showcase the effect of the proposed method on an artificial problem. We run FCA without smoothing (Equation 5) to show how our formulation allows significantly better localization of the source of the error when comparing the patch features statistics to the reference. While running the histogram method with a small patch size would improve the first result, it would fail in the second example because it contains a contextual (also called conditional [49]) anomaly.

We also compare the Histogram method and our FCA on a real case example from MVTec AD [9], in Figure 4.

3.3. Reference Selection

We analyze several options for the set of references $R(F(I))$. An intuitive solution is to use all the patches

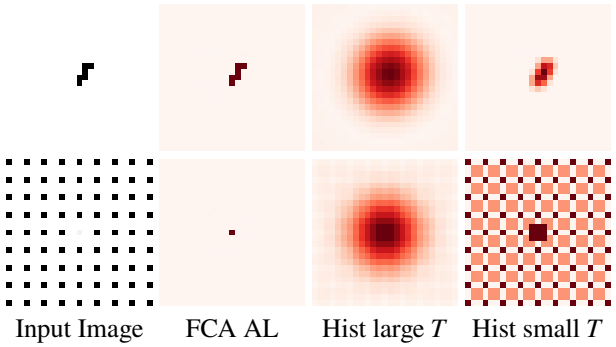

Figure 3. Anomaly localization for 2 synthetic examples when comparing patch statistics using FCA versus histograms.

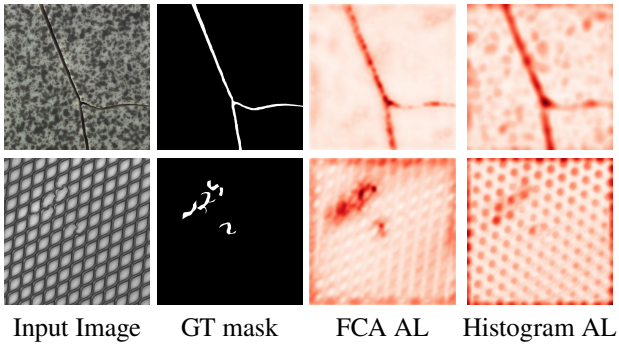

Figure 4. Anomaly localization maps for 2 textures from MVTec AD [10] when comparing patch statistics using FCA versus histograms.

in the image as references, however, this amounts to computing the pairwise distances between all patches in an image which can be very time-consuming, scaling poorly as the image resolution increases. Choosing a single patch at random is fast but is a poor approximation of the global statistics.

One can alternatively use one reference that aggregates the global information (*e.g.*, global average for moments, and the histogram over the whole feature map, for the histogram-based patch statistics comparison). In the case of SWW and FCA, we choose F_r as:

$$\underset{F_r}{\operatorname{argmin}} \sum_{(x,y)} A(x, y; \cdot, \cdot, R = \{F_r\}) \ . \tag{7}$$

The feature set that minimizes the Wasserstein distance across all patches has a closed-form solution, obtained by taking the median over the features at each sorted position individually, *i.e.*, compute the median for each order statistic for each feature channel. We analyze the performance of the global statistic aggregation method and the trade-off between the number of random patches used and performance in Table 2.

Using the median works well when the texture is homogeneous but struggles to capture the global statistics for multimodal textures (*e.g.*, structured textures with the period

PRO / Time [s]	Hist	SWW	FCA
Random (1)	60.95 / 1.1	59.56 / 5.7	62.01 / 9.2
Random (3)	67.04 / 1.1	66.79 / 8.2	69.57 / 18.7
Random (10)	72.99 / 1.2	73.23 / 17.3	75.91 / 52.3
Random (100)	74.55 / 1.7	75.69 / 134	78.45 / 482
All	74.01 / 380	– / 84984	– / 314577
Global	73.17 / 1.1	77.40 / 5.7	81.08 / 9.2

Table 2. Analysis of the effect of the Reference Selection method. We report the PRO(0.3) metric as well as the running time per image. Variants that would be unreasonably slow to be used in practice were marked with "–", and only the time was reported.

larger than the patch size). To avoid this issue, one can use the pairwise distances and discard the outliers by considering only the closest k distances. In this case, R selects the k-nearest neighbors (k-NN) over all patches in the feature maps, with respect to the cost $S(x, y; F(I), F_r)$. In Section 4.3, we only report results using k-NN references when running on low-resolution feature maps, due to the high running time of this method. We note that employing a WideResnet-50 [55] as feature extractor, using the first moment for patch statistics comparison, and taking the k-NN for reference selection yields a system equivalent to Aota *et al.* [4].

3.4. Final Method and Implementation Details

In accordance with the observations made in this section, we design our final anomaly localization system to use neural features from a pretrained neural network, evaluate the local statistics with the newly introduced FCA, and use the median for reference selection as a balance between fidelity and speed. Following recent work on anomaly detection [20, 22, 42], we use a WideResnet-50 network [55] and extract the features computed by the second convolutional block, yielding feature maps with 512 channels. Because the output of this block has a resolution 8 times smaller than the input, and FCA can handle relatively large context sizes, we choose to run the method at full resolution and not resize it as a preprocessing step as done in previous work [4, 22, 42]. All patch statistics comparison variants, including our FCA, have been implemented in PyTorch [40], utilizing CUDA acceleration, and ran on an NVIDIA RTX A5000 GPU. We use the same hyperparameters for all experiments, setting $\sigma_p = 3.0, \sigma_s = 1.0$. The patch size T should be set depending on the size of the feature maps. We use $T = 9$ when running at full dataset resolution and $T = 3$ for consistency with Aota *et al.* [4] when running at 320×320.

4. Experiments

We compare our approach with state-of-the-art methods in zero-shot anomaly detection as well as a few other adapted baselines. Several datasets are considered in order to assess the robustness of the proposed approach.

4.1. Datasets

MVTec AD. Currently, the dataset most used in the context of anomaly detection is the MVTec AD dataset [9, 10]. We use the 5 texture classes, accumulating over 500 test images and their (manually annotated) segmentation masks. The resolution of these images ranges from 840×840 to 1024×1024 pixels. Past works [4, 9, 42] propose various preprocessing and postprocessing setups, consisting of resizing and cropping to various resolutions. For a fair evaluation, we compute the metrics at full resolution, following the original evaluation script from the dataset provider [9]. The only adaption performed is cropping to the center before evaluation to avoid computing metrics on the edges of the images where most methods do not provide reliable scores [4].

Woven Fabric Textures. Bergmann *et al*. [12] introduced a small dataset for the task of defect segmentation containing two woven fabric textures (denoted WFT from here on). For each of them, 50 test images and segmentation masks are provided. The resolution of the images is 512×512.

DTD-Synthetic. Aota *et al*. [4] constructed an artificial dataset to evaluate anomaly detection methods on more diverse data, including anisotropic textures. The dataset is based on the Describable Texture Dataset [19] on which various types of defects were artificially added. The textures are also randomly rotated and cropped, eventually yielding 1304 images of small resolution (180×180 to 384×384). Following [4], all images are resized to a fixed 320×320.

Aitex. The Aitex dataset [47] contains uniform fabric textures. The defects have been manually annotated in the original images of size 256×4096. Following standard practice, we split the images into square pieces. Additionally, we discard all frames that are not completely covered by the texture and images that do not contain any anomaly. For consistency with [4], we resize the images to 320×320.

4.2. Metrics

The main metric for anomaly localization is the threshold-independent AUROC (area under the receiver operating characteristic curve). This metric is not very sensitive to spatially small anomalies. To account for this, [10] introduced the PRO(0.3) metric which weighs the size of each anomalous region, and only computes the integral up to a False Positive Rate of 0.3. Since the purpose of the proposed method is to obtain a more detailed anomaly segmentation, the PRO metric is our most important indicator. We additionally report the pixel-level maximum F_1 score [18, 57], corresponding to the (F_1-)optimal threshold. Our contribution deals with anomaly localization and does not focus on image-level labels (computed as the maximum across the anomaly scores map). Therefore, the anomaly classification metric, AUROC_c, is only reported on the MVTec AD dataset and omitted for other experiments.

4.3. Results

We compare our final method against several existing methods for zero-shot and few-show anomaly localization. The results on the MVTec AD dataset are reported in Table 3 below. We compare our system against MAEDAY, an image-reconstruction-based zero-shot method [45], WinCLIP [29], SAA [14], and April-GAN [16] based on visual-language models, and Aota *et al*. [4] which employs a WideResnet as feature extractor, and uses a simple average for patch statistics comparison, combined with a k-NN search. Despite being multi-shot methods, we include PatchCore [42] and RD++ [51] for reference. We additionally adapt methods that were not explicitly designed for ZSAL but are related in scope, for a more complete comparison. Bellini *et al*. [7] propose a method for weathering arbitrary textures from a single image, and uses an age-estimation procedure as the first step in their pipeline. The age-estimation procedure targets the same goal, to highlight regions in an image that stray away from the pristine appearance. Saliency-RC [17] as a saliency detection method highlights parts of the image that stand out, which is related to anomaly localization. However, as mentioned by the authors, the method's performance on textures is limited.

	PRO (0.3)	AUROC	AUROC_c
PatchCore all†	93.64	97.52	98.96
RD++ all† [51]	96.06	98.06	99.80
Saliency [17]	22.92	58.41	46.92
Bellini *et al*. [7]	50.75	76.06	36.90
MAEDAY† [45]	–	75.20	88.90
WinCLIP† [29]	71.5	89.06	99.64
SAA+ [14]	64.79	77.82	93.86
April-GAN [16]	92.57	96.51	97.61
Aota *et al*. [4]	93.82	97.47	**99.67**
Ours_{320}	95.46	97.74	99.21
Ours_{320} + k-NN	95.58	97.77	99.17
Ours	**97.18**	**98.73**	99.58

Table 3. Quantitative comparison on MVTec AD. We note with † results taken from different papers (may be evaluated slightly differently, as discussed in Section 4.1). The subscript $_{320}$ marks running our method at the lower resolution. PatchCore and RD++ are included for reference despite being multi-shot methods.

The proposed method improves the localization of anomalous textures significantly compared to the previous state-of-the-art zero-shot method of Aota *et al*. by using a more precise method for comparing patch statistics. Moreover, our system even outperforms prominent multi-shot anomaly detection methods PatchCore and RD++. April-GAN localizes the anomalies significantly better than methods following the same paradigm (WinCLIP, SAA+), which is consistent with April-GAN having won the zero-shot visual anomaly and novelty detection challenge at CVPR 2023. Nonetheless, on

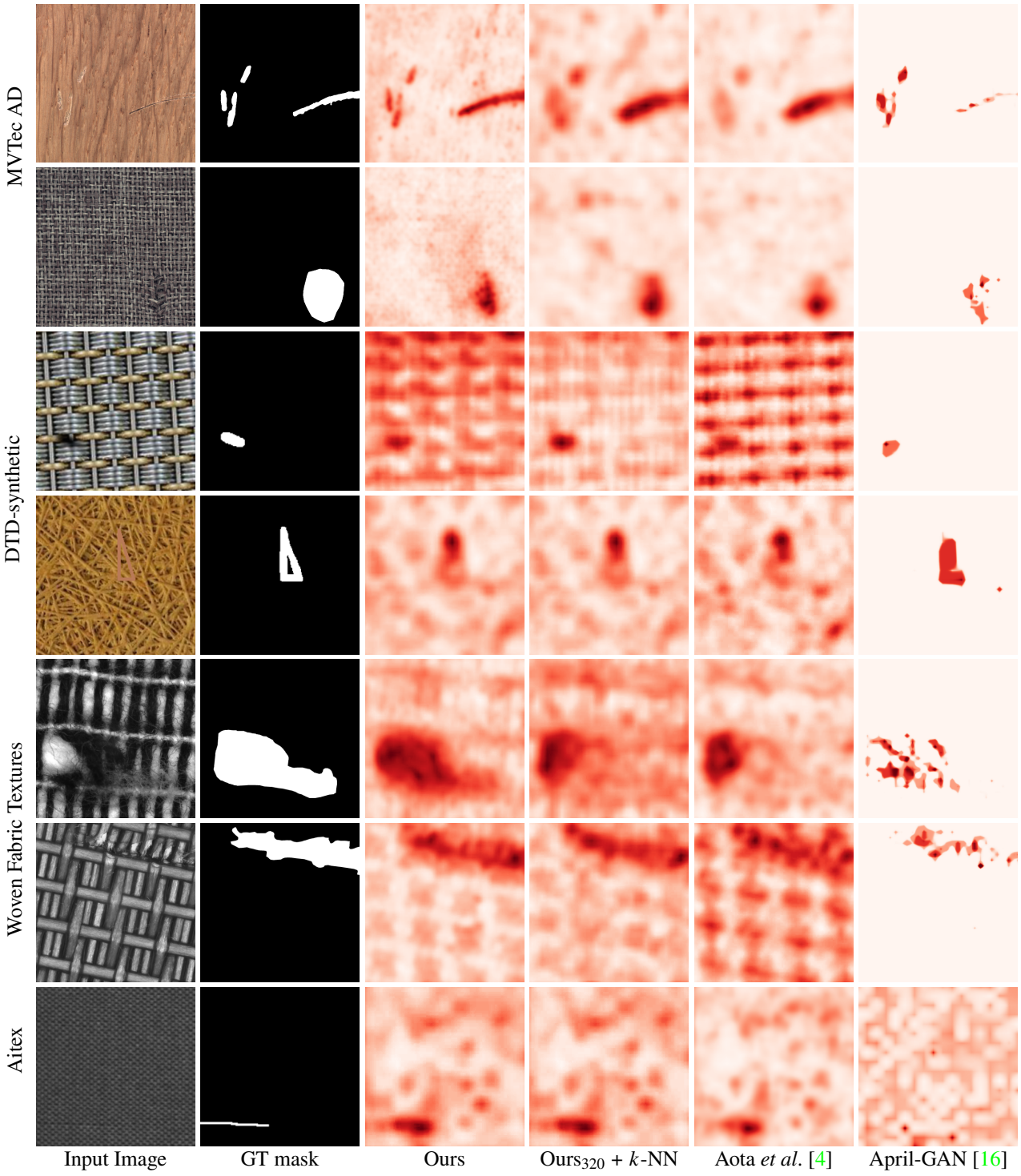

Figure 5. Qualitative comparison on challenging examples. The images are shown after cropping to center.

textures the method is clearly outperformed by our approach.

We additionally run experiments on the DTD-Synthetic dataset [4], Woven Fabric Textures [12], and Aitex [47]. In these experiments we only compare with the identified leading methods in ZSAL, *i.e.*, [4] and [16]. The results are presented in Table 4 and show that our method consistently improves upon prior art in all metrics. We find that using k-NN for reference selection can improve our results at the

	PRO (0.3)	AUROC	F_1
DTD-Synthetic			
April-GAN [16]	88.50	95.32	52.40
Aota *et al.* [4]	94.32	98.00	65.96
Ours	94.71	98.03	69.87
Ours + k-NN	**95.93**	**98.51**	**71.79**
WFT			
April-GAN [16]	84.97	94.90	71.51
Aota *et al.* [4]	84.59	96.11	72.07
Ours_{320}	73.54	93.09	65.86
Ours_{320} + k-NN	86.24	96.19	72.76
Ours	**89.57**	**98.26**	**79.13**
Aitex			
April-GAN [16]	72.62	85.90	35.23
Aota *et al.* [4]	87.11	96.70	61.09
Ours	91.07	97.51	62.39
Ours + k-NN	**91.24**	**97.52**	**62.62**

Table 4. Quantitative comparison on DTD-Synthetic, Woven Fabric Textures (WFT), and Aitex.

cost of a higher running time. Importantly, as the resolution increases, the value added by our FCA also grows. This can be seen in the results on the WFT (512×512) and MVTec AD (1024×1024) datasets where running our method at full resolution outperforms the lower resolution + k-NN variant.

In Figure 5 we present a qualitative comparison to the leading zero-shots methods [4, 16]. We show the anomaly predictions on challenging samples from each dataset. Compared to [4], the anomaly maps produced by our method have higher fidelity, with more precise localization (rows 1, 2, 4), fewer false positives (rows 3, 6), and more complete coverage of the anomalous regions (rows 5, 7). April-GAN generally fails to detect the entire anomaly. For more visualizations please see the supplementary material.

In addition to the study of various design choices in our system with respect to feature extraction and patch statistics comparison (Table 1), reference selection (Table 2), image size and k-NN usage (Tables 3 and 4), we present a direct ablation and a sensitivity analysis for our parameters T, σ_p and σ_s in the supplementary material. We also include there a brief analysis of possible failure cases for our method.

5. Discussion

The results suggest the proposed method predicts anomaly scores with high fidelity. While FCA generally performs better compared to other methods, it has a relatively high complexity. Computing local moments or histograms can be done efficiently thanks to the separability of the Gaussian kernel. This does not apply to FCA which requires sorting the values inside each sliding window. Table 5 reports the computational complexity and running time of patch-comparison-based methods for anomaly localization. The summary shows that despite the added complexity, our method scales sensibly with image resolution.

Method	Complexity	Time [s] 320×320 / 1024×1024
Moments	$O(NTD)$	0.04 / 0.05
Histogram	$O(NTBD)$	0.06 / 0.08
Aota *et al.*	$O(NTD + N^2D)$	1.10 / 199‡
Ours	$O(NT^2 \log(T)D)$	0.71 / 1.07
Ours + k-NN	$O(N^2T^2 \log(T)D)$	3.97 / 392‡

Table 5. Complexity analysis. D: number of features; T^2: patch area; N: image pixel count; B: bins per histogram. Inference time is computed at 320×320 and 1024×1024 resolutions.
‡Aota *et al.* and Ours + k-NN scale poorly with large image sizes.

Due to the large and varied datasets, our experiments support the advantage of our method robustly; however, we observe that the manually defined ground-truth masks of MVTec AD, WFT, and Aitex inevitably introduce a level of subjectivity to those ground-truth references. In some cases, as for instance shown in Figure 6, significantly different ground truth interpretations would have been possible, relativizing the accuracy score in such cases.

A limitation of our system is that, by design, it only works on textures. Generic objects can have very different feature statistics in different regions which would not be handled correctly by our method.

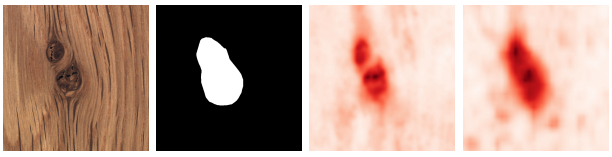

Figure 6. Manual ground-truth annotations remain subjective where multiple plausible interpretations exist.

6. Conclusion

In this work, we put forward a generic framework for performing zero-shot anomaly localization. We identify the importance of the different components and suggest a new approach that significantly improves upon prior art. The most important novelty is the proposed FCA for patch statistics comparison which enables high-fidelity anomaly localization that scales well with large textures. The performance of the method is validated on several datasets offering a comprehensive overview of the advantages of the method and the trade-off between running time and prediction quality.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956585.

References

[1] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016. 1

[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 (IJCNN)*, 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. 1, 2, 5, 6, 7, 8

[5] Matthew Baugh, James Batten, Johanna P Müller, and Bernhard Kainz. Zero-shot anomaly detection with pre-trained segmentation models. *arXiv preprint arXiv:2306.09269*, 2023. 2

[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

[7] Rachele Bellini, Yanir Kleiman, and Daniel Cohen-Or. Time-varying weathering in texture space. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2, 6

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

[9] 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. 2, 4, 6

[10] 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, 6

[11] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020. 2

[12] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *VISIGRAPP*, 2018. 2, 6, 7

[13] Rozi Bibi, Yousaf Saeed, Asim Zeb, Taher M Ghazal, Taj Rahman, Raed A Said, Sagheer Abbas, Munir Ahmad, and Muhammad Adnan Khan. Edge ai-based automated detection and classification of road anomalies in vanet using deep learning. *Computational intelligence and neuroscience*, pages 1–16, 2021. 2

[14] 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. 2, 6

[15] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021. 1

[16] 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. 2, 6, 7, 8

[17] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 2, 6

[18] Nancy Chinchor. Muc-4 evaluation metrics. In *Fourth Message Understanding Conference 22–29*. Morgan Kaufmann, 1992. 6

[19] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 6

[20] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2, 5

[21] Clayton Cooper, Jianjing Zhang, Robert X Gao, Peng Wang, and Ihab Ragai. Anomaly detection in milling tools using acoustic signals and generative adversarial networks. *Procedia Manufacturing*, 48:372–378, 2020. 1

[22] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 1, 2, 5

[23] 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. 2

[24] Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. In *European Conference on Computer Vision*, pages 544–560. Springer, 2022. 3

[25] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021. 2

[26] William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991. 3

[27] Ruei-Jie Hsieh, Jerry Chou, and Chih-Hsiang Ho. Unsupervised online anomaly detection on multivariate sensing time series data for smart manufacturing. In *IEEE conference on service-oriented computing and applications (SOCA)*, pages 90–97, 2019. 2

[28] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 1, 2

[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, June 2023. 1, 2, 6

[30] B. Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2):84–92, 1962. 2

[31] Phulpreet Kaur, M Gangadharappa, and Shalu Gautam. An overview of anomaly detection in video surveillance. In *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 607–614. IEEE, 2018. 2

[32] Kenneth I. Laws. Rapid texture identification. In *Optics & Photonics*, 1980. 3

[33] 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. 1

[34] 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

[35] Jiaqi Liu, Guoyang Xie, Jingbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *arXiv preprint arXiv:2301.11514*, 2023. 2

[36] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, June 2023. 1

[37] David Lowe. *Perceptual organization and visual recognition*, volume 5. Springer Science & Business Media, 2012. 1

[38] Joep Moritz, Stuart James, Tom S.F. Haines, Tobias Ritschel, and Tim Weyrich. Texture stationarization: Turning photos into tileable textures. *Computer Graphics Forum (Proc. Eurographics)*, 36(2):177–188, 2017. 2, 3, 4

[39] Hugo S Oliveira, João F Teixeira, and Hélder P Oliveira. Lightweight deep learning pipeline for detection, segmentation and classification of breast cancer anomalies. In *International Conference on Image Analysis and Processing*, pages 707–715. Springer, 2019. 2

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

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

[42] 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. 1, 2, 5, 6

[43] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1907–1916, 2021. 1, 2

[44] 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

[45] Eli Schwartz, Assaf Arbelle, Leonid Karlinsky, Sivan Harary, Florian Scheidegger, Sivan Doveh, and Raja Giryes. MAEDAY: MAE for few and zero shot AnomalY-detection. *arXiv preprint arXiv:2211.14307*, 2022. 1, 2, 6

[46] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8495–8504, 2021. 1, 2

[47] Javier Silvestre-Blanes, Teresa Albero-Albero, Ignacio Miralles, Rubén Pérez-Llorens, and Jorge Moreno. A public fabric database for defect detection methods and results. *Autex Research Journal*, 19(4):363–374, 2019. 6, 7

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[49] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Conditional anomaly detection. *IEEE Transactions on knowledge and Data Engineering*, 19(5):631–645, 2007. 4

[50] 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. 1, 2

[51] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24511–24520, June 2023. 1, 6

[52] S Wibisono, M Anwar, Aji Supriyanto, and I Amin. Multivariate weather anomaly detection using dbscan clustering algorithm. *Journal of Physics: Conference Series*, 1869:012077, 04 2021. 1

[53] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 2

[54] 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 (IC-ICTES)*, pages 1–5, 2019. 2

[55] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vi-*

sion Conference (BMVC), pages 87.1–87.12. BMVA Press, September 2016. 5

[56] 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. 1, 2

[57] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 6