

# CSE: Surface Anomaly Detection with Contrastively Selected Embedding

Simon Thomine<sup>1,2</sup> <sup>a</sup> and Hichem Snoussi<sup>1</sup>

<sup>1</sup>University of Technology Troyes, Troyes, France

<sup>2</sup>AQUILAE, Troyes, France

**Keywords:** Unsupervised, Anomaly, Pattern, Contrastive, Autoencoder, Feature Extraction.

**Abstract:** Detecting surface anomalies of industrial materials poses a significant challenge within a myriad of industrial manufacturing processes. In recent times, various methodologies have emerged, capitalizing on the advantages of employing a network pre-trained on natural images for the extraction of representative features. Subsequently, these features are subjected to processing through a diverse range of techniques including memory banks, normalizing flow, and knowledge distillation, which have exhibited exceptional accuracy. This paper revisits approaches based on pre-trained features by introducing a novel method centered on target-specific embedding. To capture the most representative features of the texture under consideration, we employ a variant of a contrastive training procedure that incorporates both artificially generated defective samples and anomaly-free samples during training. Exploiting the intrinsic properties of surfaces, we derived a meaningful representation from the defect-free samples during training, facilitating a straightforward yet effective calculation of anomaly scores. The experiments conducted on the MVTEC AD and TILDA datasets demonstrate the competitiveness of our approach compared to state-of-the-art methods.


## 1 INTRODUCTION

The unsupervised anomaly detection domain, especially in industrial applications, has attracted considerable attention in the past few years. Convolutional Neural Networks (CNNs) have emerged as a significant breakthrough in this field by introducing effective mechanisms for anomaly detection. The efficacy of CNNs resides in their capacity to analyze and process visual data, including images and surfaces, through the capture of spatial features and patterns. Deep learning has gained increasing momentum in the industry owing to its capacity to derive intricate representations from extensive datasets, adapt to diverse domains, and execute real-time processing. Harnessing the potential of deep learning enables industries to attain heightened accuracy, automation, and efficiency across diverse applications, including the detection of anomalies in quality control.

In the industrial setting, where precision and accuracy are of paramount importance, it is imperative to employ specialized and faultless methods that adhere to stringent standards, minimizing errors and ensuring flawless performance tailored to the specific requirements of the environment.

Recently, there has been a proliferation of approaches capitalizing on extracted features derived from pre-trained classifiers. These classifiers, trained on extensive databases like ImageNet (Krizhevsky et al., 2012), encapsulate a wealth of informative features at various levels, encompassing both low-level details such as contours and color, as well as higher-level features that are more contextual and abstract in nature. These approaches regroup mainly memory banks, normalizing flows and knowledge distillation that all offers impressive results while guaranteeing a decent inference time. The purpose of this paper is to introduce a new method based on pre-trained features that broadens the possibilities in terms of approaches to handle this specific problem while concurrently minimizing inference time.

The primary objective of feature extraction from pre-trained models is to compile the most representative features of the object, emphasizing those that exhibit differences in the presence of an anomaly. Conventional approaches employ various strategies for feature extraction, including sub-sampling of features, normalizing flows, or reconstruction-based approaches. Our conviction lies in the idea that, for effective anomaly detection, guiding the model toward features with optimal "anomaly detection" capabilities

<sup>a</sup>  <https://orcid.org/0009-0001-8989-8720>

ties for our target texture is crucial. To this end, we employ a defect generation method, such as the one introduced in DRAEM (Zavrtnik et al., 2021), to assist the model in extracting features that are responsive to defects. Our model comprises three primary components: a pre-trained feature extractor, an embedder/encoder responsible for aggregating the most representative features, and a decoder designed to avoid a trivial embedded representation. In the process of training the model, two samples are subjected to processing: one being anomaly-free, and the other exhibiting either an absence of anomalies or the presence of an artificially generated defect with a specified probability. Subsequently, the cosine similarity measure is employed as a contrastive loss function, with the objective of minimizing the embedding distance between the two samples if both are anomaly-free, or increasing it otherwise. The anomaly-free embedding of the defect-free sample is then subjected to the decoder to minimize the reconstruction loss, thereby enhancing the diversity of the embedding representation. Following the completion of the training process, a k-means clustering procedure is implemented to extract a predetermined number of clusters, which subsequently functions as a feature bank. In the testing phase, the anomaly score is computed efficiently and accurately by comparing these clusters with the embedding of the test sample. Figure 1 described our proposed score calculation approach compared to other embedding-based approaches. The primary contributions of this paper are outlined as follows:

- An embedder capturing the most representative features of a target surface through the application of a contrastive training approach, showcasing exceptional performance in the domain of texture defect detection and achieving state-of-the-art capabilities.
- A contrastive cosine loss formulated with the intention of amplifying the difference in embedding representation between defective samples and anomaly-free samples, while simultaneously diminishing this difference between two anomaly-free samples.
- A comprehensive training design incorporating a decoder to augment the variability of the embedded features, thereby preventing a trivial representation.
- A k-means clustering approach extracting the most significant clusters for anomaly scoring.

Subsequent to the introductory section, the following segment of this manuscript is devoted to a comprehensive review of existing literature concerning deep

learning methodologies utilized in unsupervised industrial anomaly detection. Section 3 presents our innovative approach with a precise description of each component. Section 4 is dedicated to a series of experiments to evaluate the efficacy of our proposed model. In section 5, an ablation study is conducted to present the benefits of each component from the contrastive approach relevance to a comparison between training methods for the decoder along with an explanation of the choice of features. A conclusive section offers a summary of the paper’s findings, outlines the limitations and proposes potential avenues for future research.

## 2 RELATED WORK

In the realm of industrial applications, the comprehensive compilation of data pertaining to every potential defect in an object or texture poses a challenging and time-intensive task where neglecting to account for all types of defects can result in sub-optimal performance outcomes (Han et al., 2022). This section provides a thorough overview of methodologies for unsupervised anomaly detection, placing specific emphasis on recent advancements that leverage deep learning techniques.

In early literature, generative models like auto-encoders (Mei et al., 2018; Nguyen et al., 2019; Zavrtnik et al., 2021), generative adversarial networks (Goodfellow et al., ), and their variations (Schlegl et al., 2019; Pourreza et al., 2021; Liang et al., 2022) were employed to reconstruct normal images from anomalous ones. Notwithstanding their utility, these methods encountered difficulties in accurately reconstructing complex objects or surfaces, occasionally leading to the generation of faulty samples.

In recent times, there has been a growing conviction that exploiting fine-grained visual features can contribute significantly to advancements in anomaly detection. Responding to this conjecture, emerging methodologies prioritize the extraction of representations from normal samples, and a prevailing approach in anomaly detection involves utilizing models pre-trained on external images datasets to comprehend the distribution of normal features.

The utilization of features extracted from pre-trained networks, especially those trained on extensive datasets such as ImageNet (Deng et al., ), has been observed to confer superior anomaly detection accuracy when compared to the direct processing of the image itself.

Within this framework, three predominant methods have emerged to exploit the extracted features.

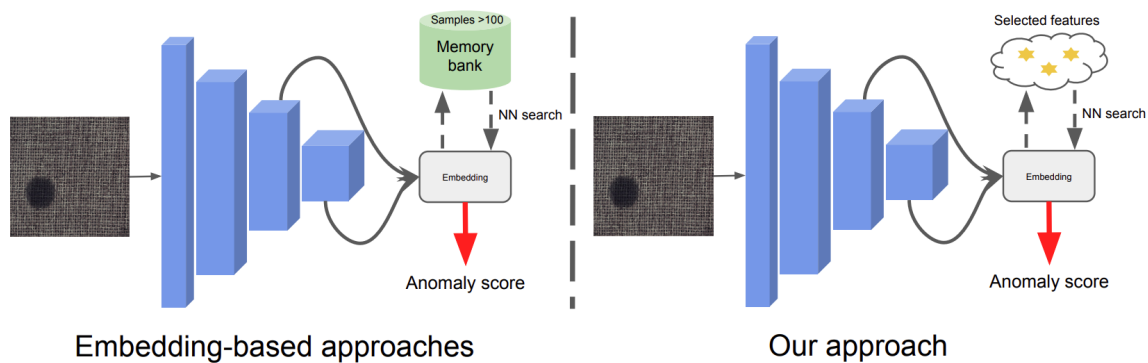


Figure 1: A comprehensive examination of the distinctions between our methodology and alternative embedding-based approaches during the inference phase. Limiting the comparison to a few specifically chosen samples, instead of encompassing the entire set of features, results in a considerable reduction in inference time.

One method focuses on estimating the distribution of the normal pattern within a parametric framework, particularly by employing normalizing flows (Rezende and Mohamed, 2016). In the training phase, flow-based models aim to minimize the negative log-likelihood loss associated with normal images, aligning their features with the target distribution to enhance the performance of the anomaly detection system. Various strategies were employed to improve performance, including the utilization of a 2D flow (Yu et al., 2021) or the adoption of a cross-scale flow (Rudolph et al., 2021).

Alternative approaches employed the concept of knowledge distillation (Hinton et al., 2015) adapted to unsupervised anomaly detection. In this approach, a student network is trained on normal samples, employing the output features of a pre-trained teacher network initially designed for classification tasks. In the testing phase, the objective of the student network is to emulate the output features of the teacher network when given defect-free samples. Nevertheless, its accuracy declines when confronted with defective samples, facilitating the derivation of a meaningful anomaly score. Diverse methods have emerged based on this paradigm such as a multi-layer feature selection (Wang et al., 2021), a reverse distillation approach (Deng and Li, 2022) (Tien et al., 2023) or a mixed-teacher approach (Thomine et al., 2023).

Memory banks approaches rely on diverse defect-free samples to accumulate pertinent features, thereby establishing a bank of features dedicated to the comparison with new samples. PatchCore (Roth et al., 2021) uses a pre-trained classifier to extract specific layers and then gathers features based on their awareness and sub-samples these features. Subsequently, these features are deposited in a memory bank, and the detection of anomalies is accomplished by comparing

patch-level distances between the core set and a given sample. Nonetheless, it is crucial to acknowledge that these methods face limitations when trained on extensive datasets, as they demand significant computational resources for the establishment of memory banks and necessitate intricate architectural considerations.

Other approaches rely on the generation of custom defects. Significantly, the DRAEM method (Zavrtanik et al., 2021), introduces a discriminatively trained auto-encoder to generate textural defects using the DTD (Describable Textures Dataset) dataset (Cimpoi et al., 2014) and Perlin noise. The CutPaste (Li et al., 2021) and MemSeg (Yang et al., 2022) approaches have also suggested the generation of structural defects to introduce diversity into the defect pool. The employed methodologies demonstrate exceptional outcomes and hold promise for textural anomaly detection, given the inherent properties of surfaces that render the generation of defects comparatively more straightforward.

### 3 PROPOSED METHOD

This section is devoted to delineating our proposed methodology, which capitalizes on distinct subcomponents to achieve efficient training and precise outcomes. Our approach relies on a contrastive training process that exploits synthesized anomalies and utilizes deep features extracted from a pre-trained model to derive a precise embedding. The complete architecture is shown in Figure 2.

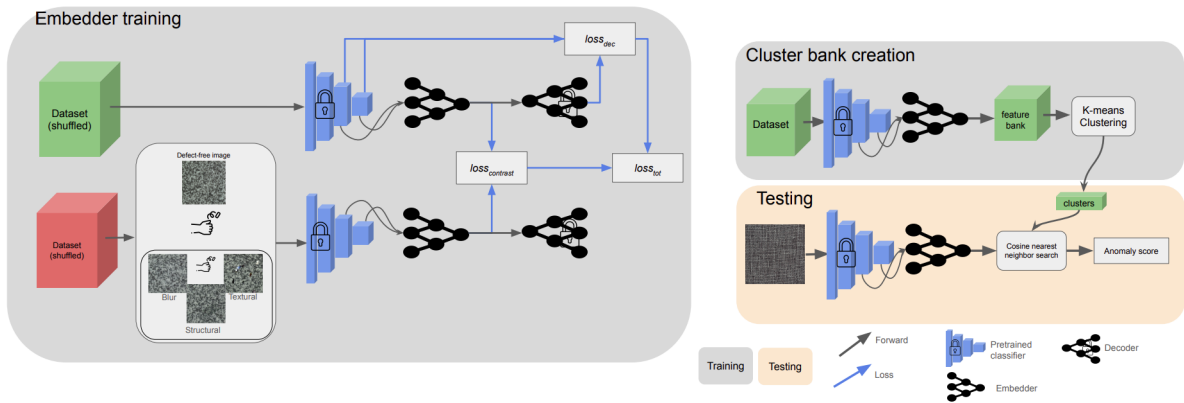


Figure 2: The complete training process. The training of the embedder constitutes the initial step, followed by the computation of clusters derived from the embedding representations.

### 3.1 Image Corruption with Synthesized Anomalies

To conduct contrastive training, it is imperative to generate anomalies. In alignment with contemporary literature, our anomaly detection process is based on Perlin Noise generation and encompasses various types of anomalies, including structural anomalies (Yang et al., 2022), textural anomalies utilizing the DTD dataset (Zavrtanik et al., 2021) (Cimpoi et al., 2014), and a novel blurry noise introduced through a straightforward application of Gaussian noise with a randomly generated kernel applied to the original image. The complete process of defect generation is detailed in Figure 3. Every category of defect manifests with equal probability during the training process to ensure a balanced training regimen and prevent bias towards any particular anomaly type. It is crucial to note that defects are randomly generated during the training process rather than pre-existing before training. This approach aims to mitigate overfitting and enhance the model’s capacity to effectively address a diverse range of defects.

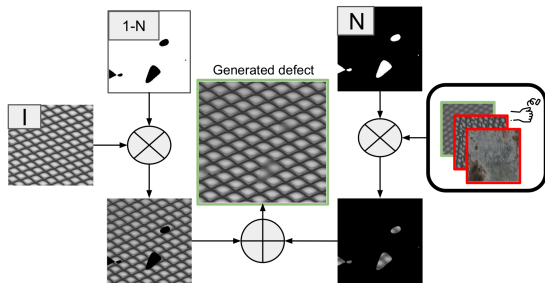


Figure 3: The defect generation process.  $N$  is the mask generated by thresholding a Perlin noise and  $(1-N)$  denote its negation.  $I$  is the original image.

### 3.2 Anomaly Detection Specific Embedding

To achieve efficient defect detection, the embedding is trained through a contrastive process, wherein the embedder is presented with pairs of images. These pairs consist of either two defect-free samples or one anomalous sample paired with one defect-free sample. Each scenario occurs with equal probability. Subsequently, the embedder is trained to augment the dissimilarity between features for antagonistic samples, while reducing it for correct samples.

In the context of surfaces, conducting contrastive training poses challenges, as a texture with a minor defect remains highly similar to a defect-free texture. To alleviate this issue, we opted to train our feature embedder using deep features extracted from a pre-trained model. Deep features offer the advantage of possessing a substantial receptive field and a relatively low resolution. Consequently, the features of a defective sample are highly likely to encompass a substantial portion of the image.

To retain spatial information and simplify the embedder architecture, we opted to exclusively employ convolutions with a kernel size of one. For enhanced capabilities, the embedder possesses the capacity to utilize features from various deep layers and efficiently fuse them without incurring any additional inference time cost.

Given a training dataset of images without anomaly  $D = [I_1, I_2, \dots, I_n]$ , our goal is to extract the relevant feature from the  $L$  top layers of a pre-trained model. For an image  $I_k \in R^{w \times h \times c}$  where  $w$  is the width,  $h$  the height,  $c$  the number of channels and  $l$  the  $l^{th}$  bottom layer, the output features are noted as  $F^l(I_k) \in R^{w_l \times h_l \times c_l}$ . The embedded feature is denoted as  $E(I_k)$ , signifying the embedding of the features extracted

from the image  $I_k$  by the pre-trained model. When presented with another image  $I_m$ , our aim is to enhance the disparity between  $E(I_k)$  and  $E(I_m)$  in the case of a defective  $I_m$ , while reducing this difference if  $I_m$  is non-defective.

The design of the embedder is straightforward, featuring a sequence of pointwise convolution layers, complemented by a ReLU layer, a batch normalization layer, and culminating in an average pooling layer that acts as a smoothing component. In the event of input features from multiple layers, the features are initially upsampled to match the size of the largest features and subsequently concatenated before being fed into the embedder.

### 3.3 Contrastive Cosine Loss

Our contrastive loss relies on cosine similarity, as opposed to the conventional mean square error. This choice is driven by the superior results observed and the absence of a margin parameter, which can be challenging to optimize. The cosine similarity is defined as:

$$\text{CosSim}(E(I_k), E(I_m)) = \frac{E(I_k) \cdot E(I_m)}{\|E(I_k)\| \|E(I_m)\|} \quad (1)$$

The cosine contrastive loss function is defined as:

$$\text{loss}_{\text{contr}} = \begin{cases} 1 + \text{CosSim}(E(I_k), E(I_m)) & \text{if } I_m \text{ is defective} \\ 1 - \text{CosSim}(E(I_k), E(I_m)) & \text{otherwise} \end{cases} \quad (2)$$

where  $\text{CosineSim}(E(I_k), E(I_m)) \in [-1; 1]$ . The objective of this loss function is to enhance the similarity of features from defect-free samples and amplify the discrepancy between features otherwise.

### 3.4 Decoder Loss

During the training of our model using only the contrastive loss, we encountered an issue of trivial representation in our embedding. This manifested as all embedded features being identical to each other. This phenomenon is attributed to the absence of diversity requirements in the training objective. To mitigate this phenomenon, we introduced a decoder designed to reconstruct features from the embedder dimension to the original dimension. The objective was to ensure diversity, as the decoder would be unable to reconstruct the original dimension from a trivial representation. Significant to note is that the decoder remains untrained throughout the training process and is initialized with random weights. Further details on this aspect are elaborated in the

ablation study. This decoder process is done only on the defect-free image  $I_k$  and the reconstruction of the layer  $l$  is noted as  $R^l(I_k)$ .

The pixel-loss function is defined as :

$$\text{ploss}^l(I_k)_{ij} = \frac{1}{2} \|F^l(I_k)_{ij} - R^l(I_k)_{ij}\| \quad (3)$$

with  $\text{ploss}^l \in \mathbb{R}^{H_l \times W_l}$ , the layer  $l$  loss function as :

$$\text{loss}^l(I_k) = \frac{1}{w_l h_l} \sum_{i=1}^{w_l} \sum_{j=1}^{h_l} \text{ploss}^l(I_k)_{ij} \quad (4)$$

and the decoder loss is written as:

$$\text{loss}_{\text{dec}}(I_k) = \sum_{l=1}^L \text{loss}^l(I_k) \quad (5)$$

The decoder process is described in Figure 4.

Ultimately, the total loss can be expressed as:

$$\text{loss}_{\text{tot}}(I_k) = \text{loss}_{\text{dec}}(I_k) + \alpha \cdot \text{loss}_{\text{contr}}(I_k) \quad (6)$$

with  $\alpha$  the weighting factor. In our experimental setup,  $\alpha$  is configured to 10.

A description of the decoder architecture for multiple layers can be seen in Figure 4.

### 3.5 Anomaly Scoring and Memory Bank

Cutting-edge memory bank methodologies necessitate the utilization of a memory bank whose scale aligns with that of the training dataset, thereby maximizing accuracy. By depending on shallow and mid-level features, these methodologies necessitate a larger number of defect-free samples to enhance the likelihood of aligning with the features of a defect-free sample during the inference process. In contrast, leveraging deep features and concentrating on surfaces obviates the requirement for a comprehensive memory bank, as features characterized by a high level of abstraction lack fine-grained details such as edges and contours. To obtain computable features for deriving an anomaly score, we employed the k-means algorithm on the embeddings of all elements within the defect-free training dataset, utilizing a variable number of clusters based on the texture's diversity. In pursuit of a domain-generalized approach, a greater number of clusters may be employed compared to a texture characterized by regular samples. In our experiments with public datasets, we configured the number of clusters to one, thereby rendering our cluster equivalent to the computation of the mean of defect-free training samples. The anomaly score is subsequently determined by calculating the cosine similarity with all clusters and selecting the minimum distance. The process is described in Figure 2.

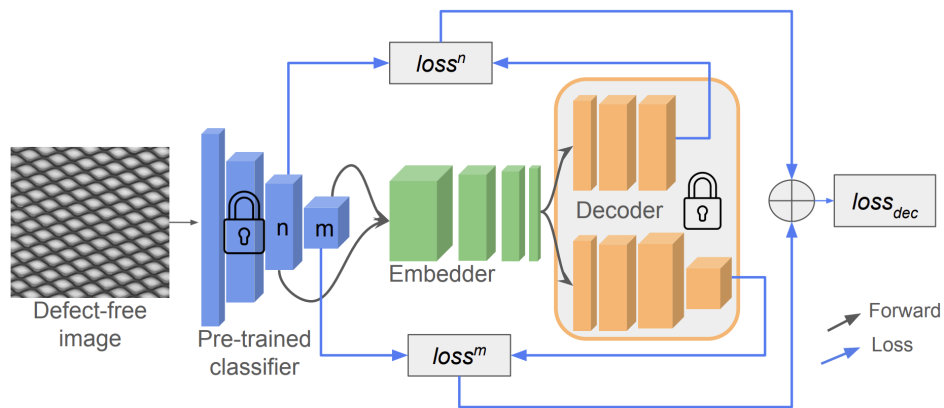


Figure 4: The decoder process for multi-layer embedder. Throughout the training process, both the pre-trained classifier and the decoder remain in a frozen state.

## 4 EXPERIMENTS

### 4.1 Implementation Details

We used the deep layers of an EfficientNet-b3 (Tan and Le, 2020) pre-trained on ImageNet as pre-trained extractor. The training and inference processes were conducted on an RTX 3090ti. In order to maintain consistency with other unsupervised approaches during the evaluation process, either the images were resized to 256x256 pixels and then further processed through center-cropping to a final size of 224x224 pixels for the dataset MVTEC AD, or conducted the evaluation under identical conditions using the anomalib library (Akcaý et al., 2022) for the TILDA (DFG, 1996) dataset. During training, the dataset was split into a training set, comprising 70% of the data, and a validation set, containing the remaining 30%. Throughout the training phase, we systematically tracked the validation loss, preserving the checkpoint corresponding to the minimum recorded loss value. To optimize the model’s parameters, we utilized the ADAM optimizer (Kingma and Ba, 2017) with a learning rate of 0.0004. To expedite convergence, we implemented a one-cycle learning rate scheduler (Smith, 2018) and conducted training over 100 epochs, utilizing a batch size of 8. All experiments presented were conducted utilizing the deep layers of EfficientNet-B3, employing input sizes of 136x14x14 and 384x7x7, along with an embedding dimension set at 64x7x7.

### 4.2 Experiments on Surface Datasets

We used the area under the receiver operating characteristic curve (AUROC) to assess the image-level

anomaly detection performance.

Our evaluation was conducted in different surfaces datasets namely the MVTEC AD dataset (Bergmann et al., 2019) and the TILDA dataset (Xie et al., 2021). These datasets compile a substantial amount of textural samples representing various conceivable scenarios.

#### 4.2.1 MVTEC AD Surfaces

The widely recognized and demanding benchmark MVTEC dataset gathers 5 surfaces and 10 objects in the realm of industrial inspection. Since our method is designed for unsupervised surface defect detection, we evaluate only on the 5 surfaces. An overview of the dataset is shown in 5. The results of our evaluation are depicted in Table 1.

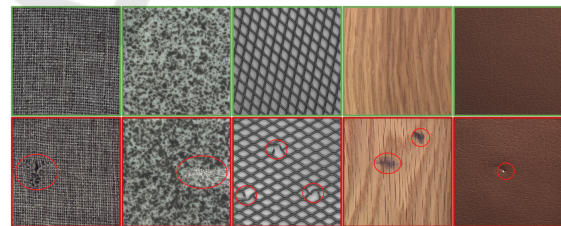


Figure 5: An overview of MVTEC AD surfaces. The figure’s upper section contains defect-free samples, whereas defective samples are situated in the lower part. Red encirclement highlights the defects.

Table 1 illustrates the competitive efficacy of our methodology relative to contemporary approaches, exhibiting a mean Area Under the Receiver Operating Characteristic (AUROC) comparable to leading models and demonstrating state-of-the-art performance on wood surface.

Table 1: Anomaly detection results with AUROC on MVTEC surfaces.

Category	CFA (Lee et al.,)	PatchCore (Roth et al., 2021)	FastFlow (Yu et al., 2021)	RD++ (Tien et al., 2023)	MixedTeacher (Thomine et al., 2023)	Ours
carpet	97.3	98.7	99.4	<b>100</b>	99.8	<b>100</b>
tile	99.4	98.7	<b>100</b>	99.7	<b>100</b>	99.3
wood	<b>99.7</b>	99.2	99.2	99.3	99.6	<b>100</b>
leather	<b>100</b>	<b>100</b>	99.9	<b>100</b>	<b>100</b>	<b>100</b>
grid	99.2	98.2	<b>100</b>	<b>100</b>	99.7	99.6
Mean	99.1	99.0	99.7	<b>99.8</b>	<b>99.8</b>	<b>99.8</b>

#### 4.2.2 TILDA Dataset

Our methodology was additionally evaluated on the TILDA (Xie et al., 2021) textile datasets encompassing a diverse collection of 8 distinct textile types from plain fabric to patterned fabric. Various examples from defective samples are illustrated in Figure 6. Results are depicted in Table 2.

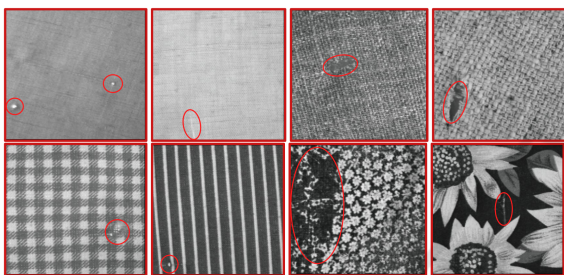


Figure 6: An overview of defective samples from the TILDA dataset. Red encirclement highlights the defects.

The outcomes presented in Table 2 exemplify the competitiveness of our approach in comparison to other state-of-the-art methods. Our methodology showcases a mean Area Under the Receiver Operating Characteristic (AUROC) superior to alternative tested methods, and notably, it achieves a superior AUROC for 4 out of the 8 fabric types considered in the evaluation.

#### 4.3 Inference Speed

An essential advantage of our approach lies in its inference speed, which is primarily constrained by the selection of the pre-trained model employed for feature extraction. The architecture of the embedder, coupled with the straightforward comparison with one or a few clusters during inference, does not substantially increase the inference time. This critical advantage establishes our method as the fastest among counterparts employing the same pre-trained model. Furthermore, it stands out as a comparably swift solution even when compared to methods utilizing a smaller pre-trained model for feature extraction. This distinction is particularly noteworthy as such methods often incorporate a secondary model to extract pertinent anomaly detection information, thereby potentially introducing additional computational overhead.

An inference speed comparison is shown in Table 3.

## 5 ABLATION STUDY

### 5.1 Comparison with a Simple Classifier

To evaluate the effectiveness of our contrastive training approach, we conducted a comparative analysis with a traditional binary classifier. This classifier was trained on defect-free samples and artificially generated anomalous samples. We maintained consistency by extracting the same deep features from EfficientNet-B3. In contrast to our contrastive training methodology, the binary classifier was trained using standard binary classification techniques rather than adopting a contrastive learning framework.

The results obtained not only showcase the descriptive capability of the deep layers of EfficientNet but also affirm the superiority of our approach when compared to a straightforward classifier. It is noteworthy to highlight that the results achieved by the classifier remain highly impressive and are comparable to state-of-the-art methods from two years ago in the context of surface defect detection. Results are shown in Table 4 for the surfaces of the MVTEC AD dataset.

### 5.2 Decoder Initialization and Training

As outlined in Section 3, we employ a decoder with frozen weights initialized randomly during the training process. While unconventional, we present our results with varying decoder initialization approaches: a decoder trained prior to embedder training, a decoder trained concurrently with the embedder, and a frozen decoder with random weights. Additionally, we offer an explanation for this unconventional methodology. The results of the first aforementioned approach are presented in Table 5.

Our conjecture posits that confining the decoder’s training exclusively to defect-free samples could induce a bias towards features crucial for reconstruction, potentially overlooking those essential for defect detection. This phenomenon results in a form of “concurrent” training between the embedder and the decoder. On the other hand, the random weight

Table 2: Anomaly detection results with AUROC on TILDA surfaces.

Category	PaDiM (Defard et al., )	CFA (Lee et al., )	Reverse distillation (Deng and Li, 2022)	Ours
tilda1	89.1	88.4	<b>94.8</b>	90.2
tilda2	88.4	86.5	88.2	<b>92.0</b>
tilda3	80.1	89.7	<b>91.4</b>	84.8
tilda4	45.9	<b>83.6</b>	59.6	80.0
tilda5	61.2	83.2	67.4	<b>88.2</b>
tilda6	79.1	85.7	78.7	<b>93.0</b>
tilda7	81.1	<b>82.4</b>	78.6	79.7
tilda8	45.8	48.1	<b>84.5</b>	68.2
Mean	71.3	80.9	80.4	<b>84.5</b>

Table 3: Comparison of pre-trained based approach in terms of inference time and frame per second.

Category	PatchCore (Roth et al., 2021)	FastFlow (Yu et al., 2021)	RD(Deng and Li, 2022)	RD (Deng and Li, 2022)	Ours
Extractor	WideResnet50 (Zagoruyko and Komodakis, )	WideResnet50	WideResnet50	Resnet18 (He et al., )	EfficientNet-b3
FPS	5.8	21.8	33	62	56
Latency (ms)	172	45.9	30	16	18

Table 4: AUROC obtained a simple classifier trained on efficientNet-b3 deep features on MVTEC surfaces.

category	carpet	wood	tile	leather	grid	mean
classifier	99.2	99.1	98.0	100	94.5	98.2

Table 5: Anomaly detection results with AUROC on MVTEC surfaces.

Category	No decoder	Trained before	Trained together	Random
carpet	99.5	99.7	99.6	100
tile	98.4	98.4	98.7	99.3
wood	99.9	100	99.9	100
leather	100	100	99.9	100
grid	99.3	99.6	98.4	99.6
Mean	99.4	99.5	99.3	99.8

initialization provides a reconstruction with a statistically balanced mix of both representative features and those pertinent to defect detection. This randomness in reconstruction aligns optimally with our training objective. An alternative option could have involved training the decoder on a combination of generated defective samples and defect-free samples. However, this approach yielded unsatisfactory results due to the limited training capacity of the decoder and the imperative for a compact architecture to ensure expeditious inference.

### 5.3 Relevance of Deep Features

In contrast to prevailing methodologies that utilize shallow and mid-level features from pre-trained models to mitigate bias towards specific classification tasks, our approach relies on deep features. These deep features, characterized by a lower resolution and a considerable number of filters, exhibit a pronounced bias toward classification making them unusable for object defect detection. This unconventional choice is elucidated by various considerations, encompassing the utilization of the contrastive loss function and

the inherent characteristics of surface defect detection. In the context of a surface, a defect typically affects only a small portion while leaving the remainder unaffected. To optimize the effectiveness of the contrastive loss, it is advantageous to extract deep features where the defect, if discernible, occupies a more substantial portion of the feature space. This is achieved by employing deep features with a larger receptive field and lower resolution. Given that the defect constitutes a significant portion of the image, the contrastive loss methodology becomes particularly beneficial. In contrast to objects, surfaces exhibit regularity, and the bias towards classification does not introduce misleading information. Indeed, as illustrated in Figure 7, the features extracted from surfaces primarily capture regular patterns. However, when a defect emerges, it becomes readily discernible. These two considerations have been instrumental in guiding our decision regarding the selection of features.

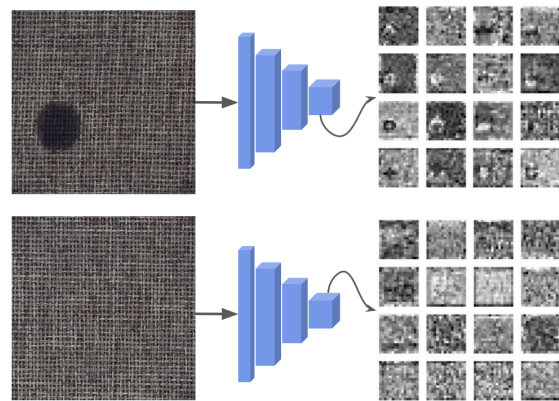


Figure 7: A sample of features extracted from the layer of size 136x14x14 from EfficientNet-b3 .



## 6 CONCLUSION

In this article, we introduced a novel method for unsupervised surface anomaly detection, centered around a contrastively selected embedding designed to aggregate the most pertinent features for the task of defect detection. Leveraging the representational capabilities of deep features extracted from a pre-trained model, our approach achieves state-of-the-art performance in surface defect detection on both the MVTEC AD dataset and the TILDA dataset. Through the employment of a compact network comprised of pointwise convolutions and a judicious selection of samples for inference comparison, our method ensures that inference speed is solely contingent on the chosen pre-trained classifier for deep feature extraction. This design leads to state-of-the-art performance in terms of model latency. However, it is crucial to acknowledge the potential limitations of our method. The primary constraint is associated with the choice of the feature extractor and our substantial reliance on its representational power. As we focus on deep features, it becomes challenging to unbiased the extracted features if the anomaly is not discernible within them. Another constraint lies in the process of defect generation during training, which significantly slows down model training, resulting in a relatively extended training duration compared to other state-of-the-art approaches. In conclusion, we posit that this methodology holds considerable promise in the field of surface defect detection, and we earnestly encourage researchers to explore and further investigate such approaches.

## REFERENCES

- Akçay, S., Ameln, D., Vaidya, A., Lakshmanan, B., Ahuja, N., and Genc, U. (2022). Anomalib: A deep learning library for anomaly detection.
- Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). MVTEC AD — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592. IEEE.
- Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. (2014). Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613. IEEE.
- Defard, T., Setkov, A., Loesch, A., and Audigier, R. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *2021 ICPR International Workshops and Challenges*.
- Deng, H. and Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN: 1063-6919.
- DFG (1996). TILDA textile texture-database.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. In *Advances in neural information processing systems*. 2014.
- Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. (2022). ADBench: Anomaly detection benchmark.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*.
- Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. In *2015 International Conference on Learning Representations (ICLR)*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, volume 60, pages 84–90.
- Lee, S., Lee, S., and Song, B. C. CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. In *IEEE Access Volume 10 Pages 78446-78454 2022*.
- Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021). Cut-Paste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE Conference on Computer Vision and Pattern Recognition*.
- Liang, Y., Zhang, J., Zhao, S., Wu, R., Liu, Y., and Pan, S. (2022). Omni-frequency channel-selection representations for unsupervised anomaly detection. In *IEEE Transactions on Image Processing 2022*.
- Mei, S., Wang, Y., and Wen, G. (2018). Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model. In *Sensors 2018*, volume 18, page 1064.
- Nguyen, Q. P., Lim, K. W., Divakaran, D. M., Low, K. H., and Chan, M. C. (2019). GEE: A gradient-based explainable variational autoencoder for network anomaly detection. In *IEEE Conference on Communications and Network Security (CNS) 2019*.
- Pourreza, M., Mohammadi, B., Khaki, M., Bouindour, S., Snoussi, H., and Sabokrou, M. (2021). G2d: Generate to detect anomaly. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2002–2011. IEEE. event-place: Waikoloa, HI, USA.
- Rezende, D. J. and Mohamed, S. (2016). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning 2016*.
- Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. (2021). Towards total recall in in-

- dustrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rudolph, M., Wehrbein, T., Rosenhahn, B., and Wandt, B. (2021). Fully convolutional cross-scale-flows for image-based defect detection. In *Winter Conference on Applications of Computer Vision (WACV) 2022*.
- Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. In *Medical Image Analysis 54*, volume 54, pages 30–44.
- Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay.
- Tan, M. and Le, Q. V. (2020). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning 2019*.
- Thomine, S., Snoussi, H., and Soua, M. (2023). MixedTeacher: Knowledge distillation for fast inference textural anomaly detection. In *2023 International Conference on Computer Vision Theory and Applications (VISAPP 2023)*, pages 487–494.
- Tien, T. D., Nguyen, A. T., Tran, N. H., Huy, T. D., Duong, S. T. M., Nguyen, C. D. T., and Truong, S. Q. H. (2023). Revisiting reverse distillation for anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, G., Han, S., Ding, E., and Huang, D. (2021). Student-teacher feature pyramid matching for anomaly detection. In *The British Machine Vision Conference (BMVC)2021*.
- Xie, H., Zhang, Y., and Wu, Z. (2021). An improved fabric defect detection method based on SSD. In *AATCC Journal of Research Volume 8. 2021*, volume 8, pages 181–190.
- Yang, M., Wu, P., Liu, J., and Feng, H. (2022). MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities. In *Engineering Applications of Artificial Intelligence Volume 119*, page 15.
- Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. (2021). FastFlow: Unsupervised anomaly detection and localization via 2d normalizing flows.
- Zagoruyko, S. and Komodakis, N. Wide residual networks.
- Zavrtnik, V., Kristan, M., and Skočaj, D. (2021). DRAEM – a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),2021*.