Pattern Analysis and Applications (2024) 27:77
https://doi.org/10.1007/s10044-024-01295-8

SHORT PAPER

# Dual model knowledge distillation for industrial anomaly detection

**Simon Thomine**[1] · **Hichem Snoussi**[1]

Received: 18 July 2023 / Accepted: 14 June 2024 / Published online: 2 July 2024


**Abstract**
Unsupervised anomaly detection holds significant importance in large-scale industrial manufacturing. Recent methods have capitalized on the benefits of employing a classifier pretrained on natural images to extract representative features from specific layers, which are subsequently processed using various techniques. Notably, memory bank-based methods, which have demonstrated exceptional accuracy, often incur a trade-off in terms of latency, posing a challenge in real-time industrial applications where prompt anomaly detection and response are crucial. Indeed, alternative approaches such as knowledge distillation and normalized flow have demonstrated promising performance in unsupervised anomaly detection while maintaining low latency. In this paper, we aim to revisit the concept of knowledge distillation in the context of unsupervised anomaly detection, emphasizing the significance of feature selection. By employing distinctive features and leveraging different models, we intend to highlight the importance of carefully selecting and utilizing relevant features specifically tailored for the task of anomaly detection. This article presents a novel approach for anomaly detection, which employs dual model knowledge distillation and incorporates various types of semantic information by leveraging high and low-level semantic information.

**Keywords** Unsupervised · Anomaly · Pattern · Knowledge distillation · Autoencoder · Feature extraction

## 1 Introduction

The field of unsupervised anomaly detection, specifically in industrial applications, has witnessed notable attention, with Convolutional Neural Networks (CNNs) presenting a substantial breakthrough in incorporating effective anomaly detection mechanisms. The effectiveness of CNNs lies in their capability to analyze and process visual data, such as images and textures, by capturing spatial features and patterns [1]. Deep learning has progressively gained momentum in the industry due to its ability to acquire intricate representations from extensive datasets, adapt to different domains, and perform real-time processing. Leveraging the potential of deep learning allows industries to derive enhanced accuracy, automation, and efficiency across various applications, including the detection of anomalies in quality control.

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

In recent times, there has been a proliferation of approaches that leverage extracted features obtained from pre-trained classifiers [2–8]. These classifiers, trained on extensive databases such as ImageNet [9], encapsulate a wealth of informative features at various levels, encompassing both low-level details like contours and color, as well as higher-level features that are more contextual and abstract in nature.

In industry, time efficiency plays a critical role, especially in real-time anomaly detection. Despite their impressive performance, techniques like memory-bank [2, 4] are not practical in real-world industrial environments as they face challenges in effectively handling large volumes of data. Therefore, it is essential to explore alternative approaches that can strike a balance between fast inference speed and high performance to address the specific needs of industrial applications.

✉ Simon Thomine
simon.thomine@utt.fr; simonthomine66@gmail.com

Hichem Snoussi
hichem.snoussi@utt.fr

[1] Department LIST3N, University of Technology of Troyes, 12 Rue Marie Curie, 10300 Troyes, France

One particularly noteworthy method among these approaches is knowledge distillation-based anomaly detection [6, 10], which has demonstrated exceptional performance with real-time inference capabilities. The essence of this method revolves around employing a teacher model, typically a pre-trained classifier, to train a student model to replicate the teacher features outputs on defect-free samples. During the phase of experimentation, in the presence of a defective sample, discrepancies arise between the feature outputs of the student and teacher models, resulting in the computation of an anomaly score and facilitating the localization of the anomaly. This technique has proven to be highly effective in anomaly detection tasks and offers efficient processing times for real-time applications [7, 8, 11].

In a Convolutional Neural Network, low-level features typically represent simple, localized patterns such as edges, corners, and textures. These features contain fundamental visual elements that serve as building blocks for higher-level representations. Mid-level features, on the other hand, capture more complex patterns and structures that emerge from combinations of low-level features. They encode meaningful arrangements of edges, shapes, and object parts, providing a richer depiction of visual content. Finally, high-level features encapsulate semantic concepts and abstract representations of objects or scenes [1].

Conventional methods in anomaly detection often prioritize the utilization of low or mid-level features to mitigate the potential bias introduced by classifiers [2–8]. However, we propose that a meticulous selection of features from multiple teacher networks can offer the opportunity to leverage even more pertinent extracted features, while simultaneously avoiding the interference caused by the classifier bias. By incorporating a diverse range of teacher networks, each with its own set of unique features, we can enhance the overall quality and relevance of the extracted features used in anomaly detection tasks.

Through the utilization of a dual model architecture, the presented approach has the capacity to acquire and distill knowledge from multiple layers, facilitating a comprehensive analysis of the input data. Utilizing deep layers poses challenges due to their inherent bias towards classification tasks and the potential risk of encountering the vanishing gradient issue, particularly in deep networks trained on limited image data. In response, we introduce a novel student architecture design complemented by an autoencoder module. This innovative approach aims to address these challenges by effectively extracting relevant features from the deep features of the teacher model. By harnessing the deep layers for extracting high-level abstract features and the shallow layers for capturing low-level details and fine-grained information, the proposed approach combines these complementary sources to enhance overall anomaly detection and localization performance.

The primary contributions of this paper are outlined as follows:

- A knowledge distillation and autoencoder approach leveraging the deeper layers of an EfficientNet [12], which demonstrates exceptional performance in the realm of detecting defects in textures, achieving state-of-the-art performance.
- A dual model knowledge distillation approach, leveraging high-level features from ResNet [13] and low-level features from EfficientNet to achieve competitive results in both anomaly detection and localization.
- An analysis of feature activation across various pre-trained models coupled with a meticulous formulation of the score calculation function.

Following the introductory section, the subsequent segment of this manuscript is dedicated to a comprehensive review of existing literature pertaining to deep learning methodologies employed in unsupervised anomaly detection. Section 3 presents our innovative approach of dual model distillation, accompanied by a precise elucidation of the underlying models. Moving forward, Sect. 4 focuses on conducting a series of experiments to rigorously evaluate the efficacy of our proposed technique. In Sect. 5, an ablation study is conducted to investigate the effects of activating different layers within various models, assess the individual contributions of each component within our model, and meticulously analyze the impact of the score calculation function. The conclusive section offers a summary of the paper's findings and outlines potential avenues for future research.

## 2 Related works

In industrial applications, the comprehensive collection of data for all potential defects in an object or texture is a challenging and time-consuming task, and the failure to account for all defect types can lead to suboptimal performance outcomes [14]. Consequently, this section presents a comprehensive overview of unsupervised anomaly detection methodologies, with a particular emphasis on recent advancements that utilize deep learning techniques.

In early literature, generative models like autoencoders [15–17], generative adversarial networks [18–21], and their variations were employed to reconstruct normal images from anomalous ones. Despite their usefulness, these methods faced challenges in accurately reconstructing intricate textures, occasionally resulting in the reconstruction of faulty samples.

Recently, there has been a growing belief that fine-grained visual features can lead to significant advancements in anomaly detection. In response to this hypothesis,

emerging methods focus on acquiring representations from normal samples, with a prevailing approach in anomaly detection being the utilization of pre-trained models on external image datasets to gain understanding of the normal feature distribution. Utilizing features extracted from pre-trained networks, particularly those trained on extensive datasets like ImageNet [9], has been found to yield superior anomaly detection accuracy compared to directly processing the image itself. These extracted features demonstrate discriminative properties for normal images, enabling the approximation of normal feature distributions and highlighting differences in defect areas. Within this framework, three predominant methodological families have emerged to effectively exploit the extracted features.

PatchCore [2] introduces an algorithm that utilizes a memory bank for anomaly detection by exploiting the correlation between patches within an image. The approach involves storing a subsampled core set of the image and extracting features from a pre-trained backbone network. Subsequently, these features are stored in a memory bank, and the detection of anomalies is achieved by comparing patch-level distances between the core set and a given sample. Similarly, CFA [4] focuses on addressing the issue of biased features from pre-trained networks impacting anomalous localization and proposes an adaptive solution tailored to the target dataset to mitigate such effects. The approach involves obtaining discriminant features through metric learning and demonstrates experimentally that these features enable highly accurate localization of complex anomalies. Notably, CFA utilizes a memory bank that is compressed independently of the target dataset size, achieving promising performance. Nevertheless, it is crucial to acknowledge that these methods have limitations when trained on extensive datasets, as they require significant computational resources for creating memory banks and demand intricate architectural considerations.

Alternative approaches concentrate on estimating the normal pattern distribution using a parametric framework, specifically through the utilization of normalizing flows [22]. These methods have demonstrated remarkable outcomes by incorporating flow-based subnetworks into their pipelines to achieve more accurate approximations of normal feature distributions. During training, flow-based models minimize the negative log-likelihood loss on normal images to align their features with the target distribution, thereby enhancing the performance of the anomaly detection system. Different strategies were used to enhance performance, such as a 2D flow [3] or a cross-scale flow [5].

The concept of knowledge distillation [23] has recently been adapted for unsupervised anomaly detection [6, 24]. This approach involves training a student network on normal samples, using the output features of a pre-trained teacher network that was initially trained for classification tasks. During the testing phase, the student network aims to replicate the output features of the teacher network when provided with defect-free samples. However, its accuracy diminishes when presented with defective samples, enabling the extraction of a meaningful anomaly score. This methodology allows for effective unsupervised anomaly detection by leveraging the knowledge transfer from the teacher network to the student network. Indeed, various methods have applied the principle of knowledge distillation in unsupervised anomaly detection, employing diverse strategies to enhance the performance such as a multi-layer feature selection [6], a reverse distillation approach [7, 8] and a mixed-teacher approach [11].

## 3 Proposed method

This section focuses on our proposed method that leverages various types of semantic information obtained from layers of different pretrained models chosen for their expressive feature capabilities.

### 3.1 Deep features network

The objective of this model is to extract high-level semantic information from an EfficientNet [12], while the ablation study addresses the question regarding the selection of the model and layers.

**Knowledge distillation part**: Given a training dataset of images without anomaly $D = [I_1, I_2, ..., I_n]$, our goal is to extract the information of the $L$ top layers of EfficientNet model. For an image $I_k \in R^{w \times h \times c}$ where $w$ is the width, $h$ the height and $c$ the number of channels, the teacher and student outputs features are defined respectively as $F_t^l(I_k) \in R^{w_l \times h_l \times c_l}$ and $F_t^l(I_k) \in R^{w_l \times h_l \times c_l}$ where $l$ denotes the $l^{th}$ bottom layer. During the training phase, the student model is trained to reproduce the teacher features on normal samples. In the anomaly detection setting, normal samples conform to the identical distribution in both $F_t$ and $F_s$, while out-of-distribution samples are regarded as anomalies.

In the context of anomaly detection, the utilization of techniques such as knowledge distillation and pretrained classifiers often aims to exploit high and mid-level features. This choice is made to mitigate the inherent bias that arises from the classifier's inclination towards its specific task.

Reverse distillation [8] demonstrates that directly duplicating the teacher architecture when training the student can potentially lead to a suboptimal anomaly detector. This occurs because both networks have identical information flow, which may cause the student to inadvertently learn how to reproduce defects as well.

Recognizing the potential bias introduced by the classifier, coupled with the issue stemming from architectural

symmetry, we have chosen to modify the architecture of the student model as a means to mitigate these concerns. Incorporating low-level features presents a challenge in model training due to limited training data and the focus on optimizing deeper layers, necessitating the development of an effective approach to enhance the learning of low-level features. The ablation study section delves into an analysis of the model's behavior, highlighting the necessity of designing a specialized architecture.

To mitigate the aforementioned issue during the training process, we introduced a residual network comprised of residual blocks consisting of $5 \times 5$ convolutions. This architecture mimics the number of filters and the image dimension of the EfficientNet-b0 model. This residual network aims to maintain a consistent information flow during training and prevent the student network from learning the exact information flow of EfficientNet while guaranteeing an effective training process.

**Autoencoder part**:

Despite utilizing the aforementioned architecture, upon visualizing the training results, we observed that there was still a deficiency of information in the defect-free student reconstruction. To address this issue, we introduced a simple yet effective autoencoder module for each selected deep feature of the network. The main goal of integrating the autoencoder module was to address the information gaps present in the student's extracted features, with the purpose of improving the information retrieval capabilities of the student model. The integration of the autoencoder module introduces a novel approach to extracting representations of normality derived from the teacher model. The distinctive design contrast between the student model and the autoencoder module results in a more comprehensive array of information regarding normality. This allowed us to capture missing details in the defect-free reconstructions, thereby improving the overall performance and accuracy of the anomaly detection process.

Both the residual network and the autoencoder module can be used independently based on specific requirements, particularly in scenarios where high inference speed is a crucial factor. The aforementioned architecture is presented in Fig. 1.

### 3.2 Anomaly detection and localization with shallow layers

The preceding model has been tailored for anomaly detection, but it lacks effectiveness in anomaly localization. In conventional knowledge distillation frameworks, anomaly localization is accomplished by upsampling primarily high-level features employed in the model, resulting in a precise anomaly map due to the preservation of the large image dimension.

Given the emphasis of our model on low-level features, the upsampling technique utilized for anomaly localization produces less precise outcomes, resulting in a coarse localization. To mitigate this concern, we introduced a dedicated student network specifically designed for anomaly localization; however, due to the suboptimal performance of the EfficientNet architecture for shallow layers, we opted for an alternative network architecture for the student model.

To ensure a satisfactory balance between inference speed and effective localization capabilities, we employed reduced student [11], a framework based on a ResNet18 [13] teacher, using the outputs from the first two residual blocks. By doing so, we achieved a desirable compromise, allowing for efficient

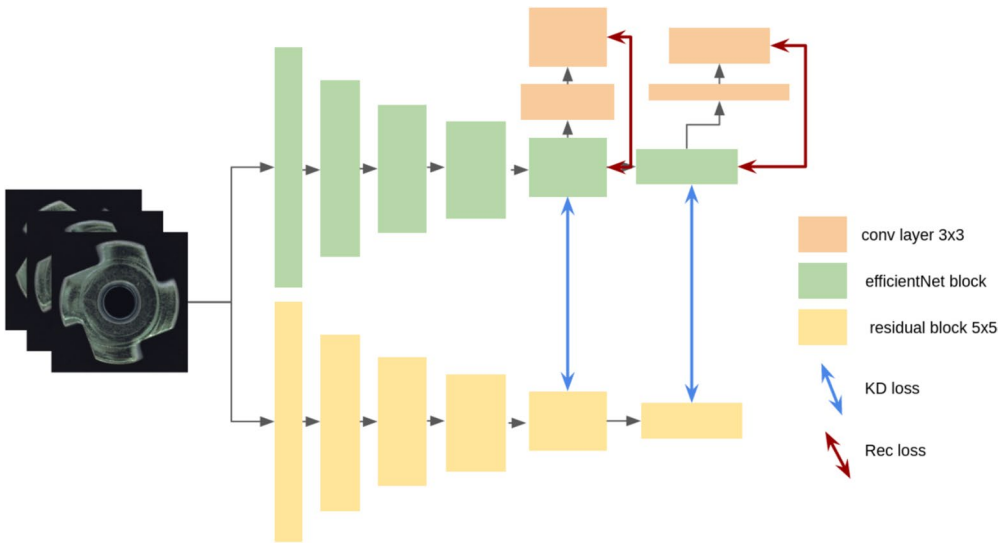


**Fig. 1** Deep Features network architecture

inference times while still retaining reliable localization capabilities. Although this model was primarily introduced for anomaly localization, it also plays a significant role in anomaly detection, particularly for objects. Deep layers within the model tend to struggle with accurately describing objects, making the localization model crucial in enhancing anomaly detection in such cases. An overview of the complete architecture is described in Fig. 2.

### 3.3 Loss and anomaly scoring

The choice of the loss function for our approach involves utilizing the mean squared error between the features of the student model and the teacher model, which is consistent with the prevailing practice observed in existing literature on knowledge distillation methodologies [6, 11]. In the subsequent equations, $F_t$ represents the teacher features, $F_s$ denotes the student features and $I_k$ denotes the input image.

The pixel difference is defined as:

$$M^l(I_k)_{ij} = \frac{1}{2}\|norm(F_t^l(I_k)_{ij}) - norm(F_s^l(I_k)_{ij})\|, \quad (1)$$

with $M^l \in \mathbb{R}^{w_l \times h_l}$,the layer l loss function as:

$$loss^l(I_k) = \frac{1}{w_l h_l}\sum_{i=1}^{w_l}\sum_{j=1}^{h_l} M^l(I_k)_{ij}, \quad (2)$$

and the global loss is written as:

$$loss(I_k) = \sum^{l} \alpha_l loss^l(I_k), \quad (3)$$

with $\alpha_l$ the loss factor for the layer $l$.

In previous knowledge distillation methods [6, 8, 11], the common practice involves employing mean squared error, followed by upsampling and selecting the maximum value from the resulting anomaly map as the anomaly score. We hold the belief that cosine similarity is better suited for this specific problem, considering its focus on global similarity, particularly in relation to textures. Moreover, given the presence of different features with varying sizes, it is crucial to consider the method of combining these features in accordance with the desired impact of each feature. Two potential options were considered for our approach: the first involved adding each layer with a predetermined factor, chosen to maximize anomaly detection based on prior knowledge, while the alternative approach involved using feature multiplication to ensure equal influence for each selected feature.

After thoughtful consideration, we have opted for a model-specific approach, incorporating both methods based on the significance of features and our knowledge, enabling us to capitalize on the respective strengths of each method and customize their implementation accordingly. In the context of the EfficientNet component, we

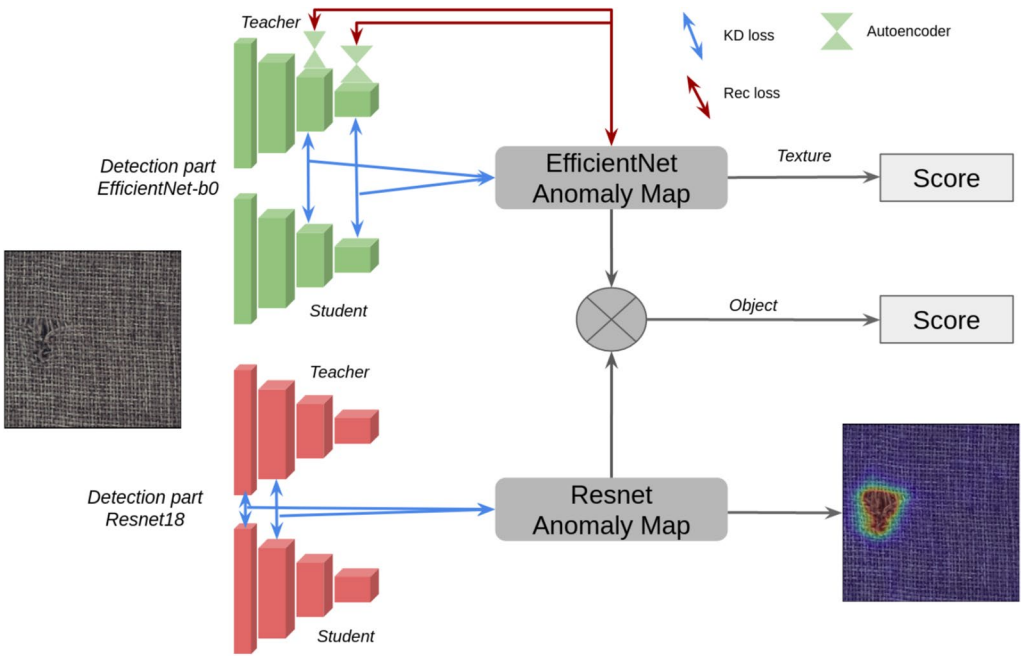


**Fig. 2** This figure illustrates the complete model architecture, combining the EfficientNet architecture for deep layers with the ResNet architecture for shallow layers and localization tasks. The upper section harnesses the deep features of a pre-trained EfficientNet model through an asymmetric student architecture and an autoencoder module, demonstrating strong performance, particularly in texture anomaly detection. In contrast, the lower section focuses on anomaly localization by utilizing features with bigger spatial dimensions

calculate the score function using cosine similarity and feature multiplication for the selected features, as we consider them to be highly significant for addressing the anomaly detection problem. The following equations employs $l_e$ to denote the layers of the efficient student-teacher part and $l_r$ to represent the layers of the ResNet student-teacher part.

The pixel-score is defined as:

$$pscore^{l_e}(I_k)_{ij} = 1 - \frac{(F_t^{l_e}(I_k)_{ij})^T . F_s^{l_e}(I_k)_{ij}}{\|F_t^{l_e}(I_k)_{ij})\| \|F_s^{l_e}(I_k)_{ij}\|}. \quad (4)$$

The total EfficientNet score is then defined as:

$$\text{score}_{\text{effNet}}(I_k) = \max\left(\prod^{l_e} \text{UpSample}\left(\text{pscore}^{l_e}(I_k)\right)\right). \quad (5)$$

For the ResNet component and the localization score, we used the pixel difference introduced in Eq. 1, and we apply an upsampling:

$$L_{map}(I_k) = \sum^{l_r} UpSample(M^{l_r}(I_k)). \quad (6)$$

where $L_{map}(I_k) \in \mathbb{R}^{W \times H}$ and $l_r$ are the layers of the ResNet student. A Gaussian filter is then used on the normalized $L_{map}$ to smooth the defect localization [6]. The localization score is then calculated with the pixel-wise similarity between the ground truth mask and the normalized and blurred $L_{map}(I_k)$. To infer the ResNet detection score, we take the maximum value of the generated anomaly map.

**Discussion.** It is crucial to underscore the distinctions between our proposed approach and MixedTeacher [11] for several reasons. Firstly, MixedTeacher employs an EfficientNet backbone for the student, which we have demonstrated to be suboptimal for preserving the essential information content of the pretrained student. To address this limitation, we have developed a novel backbone specifically tailored for the deep student model, ensuring the optimal retention of crucial information. Secondly, our approach integrates an autoencoder module specifically designed to extract a richer representation of normality from the teacher model. Lastly, our method introduces an innovative score calculation technique to harness the strengths of each student.

## 4 Experiments

### 4.1 Implementation details

We used EfficientNet-b0 [12] and ResNet18 [13] pretrained on ImageNet as backbones respectively for the detection and localization part. The training and inference processes were conducted on an RTX 3080ti. In order to maintain consistency with other unsupervised approaches during the evaluation process, the images were initially resized to $256 \times 256$ pixels and then further processed through center-cropping to a final size of $224 \times 224$ pixels. The dataset was split into a training set, comprising 70% of the data, and a validation set, containing the remaining 30%. During training, we monitored the validation loss and retained the checkpoint corresponding to the lowest loss value. To optimize the model's parameters, we utilized the ADAM optimizer [25] with a learning rate of 0.005. To dynamically adjust the learning rate during training, we employed a scheduler that effectively reduced the learning rate when the model's performance reached a plateau. The training process spanned 100 epochs with a batch size of 8.

### 4.2 Experiments on MVTEC AD dataset

We used the area under the receiver operating characteristic curve (AUROC) to assess the image-level and pixel-level anomaly detection performance, utilizing the generated anomaly map. Our evaluation was conducted on the MVTEC AD dataset [26] (dataset link), a widely recognized and demanding benchmark comprising 5 texture classes and 10 object categories. The proposed methodology was tailored specifically for detecting defects in textures, and we report the corresponding results in Tables 1 and 2. The outcomes for the object categories are presented separately in Table 3.

**Table 1** Anomaly detection results with %-AUROC on MVTEC AD textures

| Category | CFA [4] | PatchCore [2] | FastFlow [3] | RD++ [7] | Mixed-Teacher [11] | Ours |
|---|---|---|---|---|---|---|
| carpet | 97.3 | 98.7 | 99.4 | **100** | 99.8 | **100** |
| tile | 99.4 | 98.7 | **100** | 99.7 | **100** | **100** |
| wood | **99.7** | 99.2 | 99.2 | 99.3 | 99.6 | **99.7** |
| leather | **100** | **100** | 99.9 | **100** | **100** | **100** |
| grid | 99.2 | 98.2 | **100** | **100** | 99.7 | **100** |
| Mean | 99.1 | 99.0 | 99.7 | 99.8 | 99.8 | **99.9** |

**Table 2** Anomaly localization results with %-AUROC on MVTEC AD textures

| Category | CFA [4] | PatchCore [2] | FastFlow [3] | RD++ [7] | Mixed-Teacher [11] | Ours |
|---|---|---|---|---|---|---|
| carpet | 98.9 | 99.1 | 98.9 | **99.2** | 98.8 | **99.2** |
| tile | 95.6 | **96.6** | 95.6 | **96.6** | 93.7 | **96.6** |
| wood | 95 | 94.1 | 95.3 | **95.8** | 92.4 | **95.8** |
| leather | 99.3 | **99.6** | 99.4 | 99.4 | 98.7 | 99.4 |
| grid | 98.1 | 98.7 | 99.2 | **99.3** | 97.5 | **99.3** |
| Mean | 97.4 | 97.6 | 97.7 | **98.1** | 96.2 | **98.1** |

**Table 3** Anomaly detection and localization results with %-AUROC on MVTEC AD objects

| Category | CFA [4] | PatchCore [2] | FastFlow [3] | RD++ [7] | Ours |
|---|---|---|---|---|---|
| bottle | **100**/**98.8** | **100**/98.6 | **100**/98.6 | **100**/**98.8** | **100**/98.7 |
| cable | 98.8/**99.0** | **99.5**/98.4 | 96.2/98.6 | 99.2/98.4 | 99.1/93.4 |
| capsule | 97.3/**99.1** | 98.1/98.8 | 96.3/99 | **99**/98.8 | 86.7/97.4 |
| hazelnut | **100**/98.9 | **100**/98.7 | 99.4/98 | **100**/**99.2** | **100**/98 |
| metal nut | **100**/**99.2** | **100**/98.4 | 99.5/98.8 | **100**/98.1 | 99.9/96.4 |
| pill | 97.9/**98.9** | 96.6/97.4 | 94.2/97.6 | **98.4**/98.3 | 96.9/95.9 |
| screw | 97.3/98.9 | 98.1/99.4 | 83.9/96.6 | **98.9**/**99.7** | 85.1/96.8 |
| tooth-brush | **100**/99.0 | **100**/98.7 | 83.6/98 | **100**/**99.1** | 83.9/98.6 |
| transistor | **100**/**98.1** | **100**/96.3 | 97.9/97.1 | 98.5/94.3 | 98/82.8 |
| zipper | **99.6**/**99.0** | 99.4/98.8 | 95.1/95.5 | 98.6/98.8 | 94.5/98.7 |
| Mean | 99.1/**98.9** | **99.3**/98.4 | 94.6/97.8 | **99.3**/98.4 | 94.4/95.7 |

Table 1 demonstrates the superiority of our approach over previous state-of-the-art methods in terms of anomaly detection on textures, as evidenced by a remarkable mean %-AUROC score of 99.94. Moreover, as seen in Table 2, our method showcases competitive anomaly localization results on textures, with a mean performance that rivals the current state-of-the-art methodology.

On objects, as exposed in Table 3 the results are somewhat nuanced, with some objects achieving similar or close-to-state-of-the-art performances, while others yield unsatisfactory results. We have observed that on objects with low scores, the detected defects often occur in the background of the image when the background is not entirely homogeneous. This is likely due to the highly abstract information derived from the deep layers of EfficientNet.

### 4.3 Experiments on TILDA dataset

To assess our detection capabilities on textures, we also performed experiments on the TILDA dataset [27] (datas et link), which regroups 8 different fabric textures. The Table 4 showcases our results on this dataset.

**Table 4** Anomaly detection results with %-AUROC on TILDA textures

| Category | CFA [4] | RD++ [7] | DBFAD [28] | Ours |
|---|---|---|---|---|
| tilda1 | 88.4 | 93.6 | **96.9** | **96.9** |
| tilda2 | 86.5 | 96.3 | 95.8 | **97.8** |
| tilda3 | 89.7 | 86.3 | 92.5 | **94.4** |
| tilda4 | 83.6 | 75.4 | 75.0 | **88.6** |
| tilda5 | **91.2** | 75.1 | 87.2 | **91.2** |
| tilda6 | 85.7 | **90.0** | 88.6 | 89.2 |
| tilda7 | 82.4 | **86.5** | 70.7 | 71.9 |
| tilda8 | 48.1 | 47.6 | 61.9 | **74.5** |
| Mean | 80.9 | 81.4 | 83.6 | **88.1** |

### 4.4 Inference speed

By leveraging small pretrained network backbones, our methodology has successfully attained state-of-the-art outcomes in both inference speed and AUROC. This aspect was a key consideration during the design of our approach, given that industrial demands frequently necessitate rapid inference speeds, particularly for textures that involve extensive surface control. The inference speed results are shown in Table 5.

### 4.5 Discussion

In this section, we showcase the superior performance of our method over state-of-the-art approaches in both texture unsupervised anomaly detection and localization. Despite these achievements, our approach encounters challenges in detecting anomalies within certain types of objects. However, it excels in terms of inference speed, offering millisecond-level latency and providing flexibility in choosing the desired subtask to accomplish. This is particularly significant as the localization of anomalies may not always be necessary in an industrial process, emphasizing the practical applicability and efficiency of our approach.

## 5 Ablation study

### 5.1 EfficientNet-b0

In this section, we present our observations on the visual activation of features in relation to defects within textures. These observations aim to highlight the underlying factors that motivated our architectural choices for the proposed method.

#### 5.1.1 Layer activation visualization

Fig. 3 illustrates the remarkable activation of deep EfficientNet-b0 features in response to defective textures, surpassing the activation observed in layer 2 of ResNet, which is commonly utilized in the literature for defect detection based on pretrained models. This finding supports our architectural choice of leveraging deep EfficientNet-b0 features in our proposed method for enhanced defect detection performance.

#### 5.1.2 Classic training student-teacher efficientnet

EfficientNet-b0 has been previously employed with the intention of combining two models to enhance both defect detection and localization [11]. They employed an identical model for training the student, thereby failing to fully leverage the model's capabilities. In contrast, as elucidated in Sect. 3, we opted for a distinct architecture featuring residual connections and a 5x5 convolutional layer, which proved crucial in achieving satisfactory training outcomes.

Despite encountering challenges during the training process using a mirror model, it is worth noting that the powerful nature of EfficientNet-b0 features enabled us to achieve good results. This fact becomes evident when considering the visualized outcomes, as depicted in Fig. 4. The training procedure demonstrates indications of underfitting, vanishing gradients, and convergence towards a uniform distribution of activations. This observation further underscores the inherent strength of EfficientNet-b0 features that, even with a uniform distribution, still demonstrates the capability to effectively discriminate the majority of defects by utilizing cosine distance measurements as demonstrated in Table 6.

### 5.2 Model part separated

Now let us examine each component of the texture detection model individually, with the first section dedicated exclusively to the analysis of the EfficientNet model and the second part focusing on the autoencoder module.

**Table 5** Comparison of pretrained based approach in terms of inference time and frame per second

| Category | PatchCore [2] | FastFlow [3] | RD(Resnet18) [8] | RD(WR50 [29]) [8] | Ours Det | Ours Loc |
|---|---|---|---|---|---|---|
| FPS | 5.8 | 21.8 | 62 | 33 | 83 | 167 |
| Latency (ms) | 172 | 45.9 | 16 | 30 | 12 | 6 |

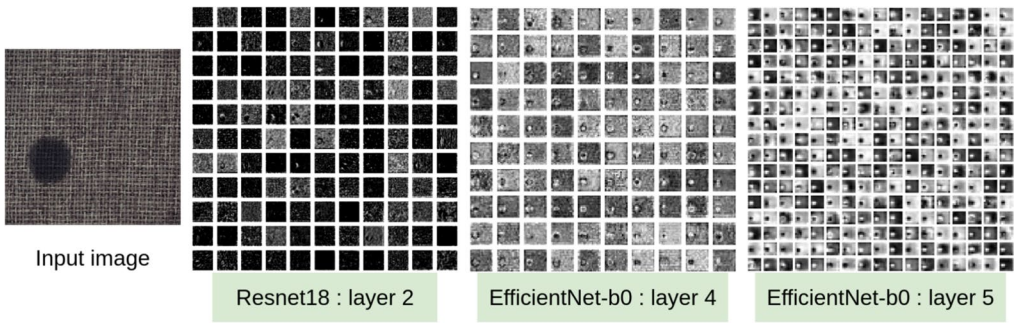


**Fig. 3** Feature activation to a defective texture from different layers of different models

**Table 6** Anomaly detection results with %-AUROC on MVTEC AD textures with EfficientNet-b0 architecture for both the student and the teacher

| category | carpet | wood | tile | leather | grid | mean |
|---|---|---|---|---|---|---|
| miror student | 100 | 99.4 | 100 | 99.8 | 95.6 | 98.9 |

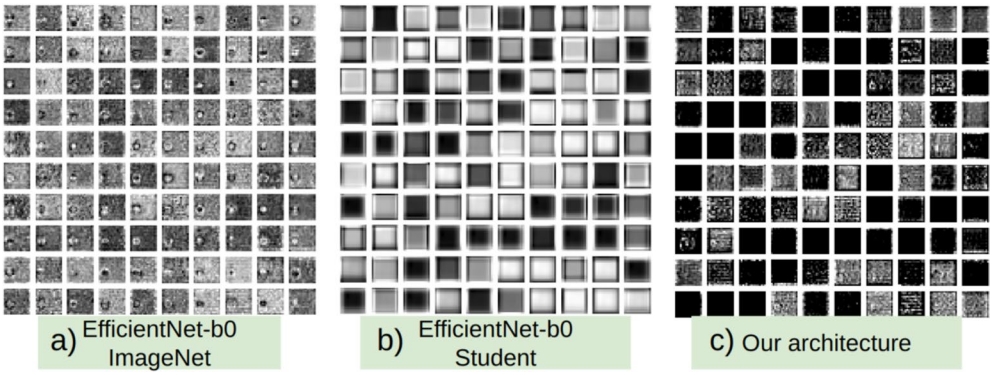


**Fig. 4** Comparison of feature activations between pretrained EfficientNet-b0, student EfficientNet-b0, and our model. (a) The features extracted from the EfficientNet-b0 teacher model clearly depict the presence of a defect. (b) Features from an efficient-b0 student model indicate that the model has learned to produce uniform reconstructions, with identical values regardless of the input data. (c) Features obtained from our architecture obscure the defect, enabling a more accurate score calculation

**Table 7** Anomaly detection on MVTEC AD textures for knowledge distillation (KD), autoencoder (AE) and the mixed results

| category | carpet | wood | tile | leather | grid | mean |
|---|---|---|---|---|---|---|
| KD | 100 | 99.7 | 99.9 | 100 | 99.9 | 99.90 |
| AE | 100 | 99.5 | 99.5 | 99.9 | 96.7 | 99.12 |
| AE+KD | 100 | 99.7 | 100 | 100 | 100 | **99.94** |

#### 5.2.1 Knowledge distillation model part

The student-teacher component, which relies on the EfficientNet model, serves as the primary element of our proposed architecture. As depicted in Fig. 4, we observe that our architecture successfully reconstructs certain informative EfficientNet-b0 features even in the absence of defects. This finding suggests that our architecture is capable of overcoming the training challenges encountered with the mirrored EfficientNet model, thereby achieving improved performance and better feature representation. The results of the knowledge distillation approach in terms of anomaly detection are shown in Table 7.

#### 5.2.2 Autoencoder part

The incorporation of the autoencoder module in our architecture was prompted by the observation, as illustrated in Fig. 4, that despite the utilization of a specifically designed architecture, we were unable to achieve complete reconstruction of all the features. Consequently, we proposed the idea of incorporating an autoencoding block specifically tailored to reconstruct the deeper EfficientNet layers, with the hope that the resulting reconstructed layers would provide us with additional information for an improved anomaly scoring. Table 7 presents the results obtained by utilizing the autoencoder alone as well as in conjunction with the knowledge distillation model. Notably, even with solely the autoencoder part, remarkable results can still be attained, suggesting the possibility of utilizing the encoder independently to optimize inference speed while maintaining excellent performance.

**Table 8** Comparison between the different score calculation methods. The use of anomaly map multiplication technique enhances defect detection performance for both Mean Squared Error (MSE) and Cosine similarity metrics

| category | MSE + add | MSE + mul | Cosine + add | Cosine + mul |
|---|---|---|---|---|
| mean AUROC | 99.84 | 99.92 | 99.90 | 99.94 |

### 5.3 Anomaly scoring impact

This section aims to demonstrate the influence of our multiplicative cosine similarity score function in comparison to conventional approaches such as mean squared error (MSE) and layer addition. The comparison results are presented in Table 8 demonstrating the superiority of the multiplication approach and suggesting that all feature maps have an equal influence on the overall result. This observation implies that the selected layers exhibit complementary characteristics, with each compensating for the information gaps of the others. The multiplication in the score calculation allows for the full expression of

the respective qualities of both features in terms of defect detection, further emphasizing their collective contribution to the overall performance.

**Discussion.** The observed enhancement in performance resulting from the integration of the autoencoder module and the adoption of the cosine distance metric over the mean squared error might not exhibit a pronounced effect and may not achieve statistical significance. Nonetheless, the autoencoder in isolation demonstrates competitive efficacy and remains a viable methodology independently. Furthermore, our investigation into score calculation functions underscores the superiority of feature multiplication over addition.

## 6 Conclusion

In this article, we proposed a novel dual model knowledge distillation approach that effectively tackles the challenges of detection and localization as distinct tasks. Specifically, we focused on the detection part of our model, emphasizing textures, and conducted a comprehensive analysis of features extracted from various layers of different pretrained models, spanning from shallow to deep layers. Through visual observations and extensive testing, we successfully designed a trainable architecture that harnesses the most expressive features for effective defect detection on textures. The separation of anomaly detection and localization into two distinct models offers several advantages. Firstly, it provides flexibility for users who may be solely interested in either anomaly detection or localization, allowing them to focus on the specific aspect they require. Secondly, this separation enables the utilization of deeper features with larger receptive fields for detection purposes, while using shallower features with smaller receptive fields for localization. This approach offers greater flexibility in selecting appropriate layers for each task, potentially enhancing the performance and adaptability of the overall system. From a perspective standpoint, we assumed that EfficientNet features derived from a pretrained classifier exhibit superior performance compared to feature extractors based on other pretrained models. Exploring other pre-trained architectures should be considered in relation to diverse problem domains and varying technical constraints. Furthermore, we hypothesize that leveraging EfficientNet features in a more efficient manner could potentially enhance both performance and inference time compared to our proposed method.

**Data Availibility Statement** The datasets analyzed during the current study are available in the MVTEC website (https://www.mvtec.com/company/research/datasets/mvtec-ad) and the TILDA website (https://lmb.informatik.uni-freiburg.de/resources/datasets/tilda.en.html).

## Declarations

**Conflict of interest** The authors declared no potential Conflict of interest with respect to the research, authorship and/or publication of this article.

## References


1. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–44. https://doi.org/10.1038/nature14539
2. Roth K, Pemula L, Zepeda J, Scholkopf B, Brox T, Gehler P (2022) Towards Total Recall in Industrial Anomaly Detection. In:2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 14298–14308. IEEE, New Orleans, LA, USA. https://doi.org/10.1109/CVPR52688.2022.01392
3. Yu1 J, Zheng Y, Wang X, Li W, Wu Y, Zhao R, Wu L (2021) FastFlow: unsupervised anomaly detection and localization via 2D normalizing flows. ArXiv **abs/2111.07677**
4. Lee S, Lee S, Song BC (2022) CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. IEEE Access 10:78446–78454. https://doi.org/10.1109/ACCESS.2022.3193699
5. Rudolph M, Wehrbein T, Rosenhahn B, Wandt B (2021) Fully convolutional cross-scale-flows for image-based defect detection. 2022 IEEE/CVF winter conference on applications of computer vision (WACV), 1829–1838
6. Wang G, Han S, Ding E, Huang D (2021) Student-teacher feature pyramid matching for anomaly detection. In: British machine vision conference. https://api.semanticscholar.org/CorpusID:240070818
7. Tien TD, Nguyen AT, Tran NH, Huy TD, Duong STM, Nguyen CDT, Truong SQH (2023) Revisiting reverse distillation for anomaly detection. In:2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 24511–24520. IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/CVPR52729.2023.02348
8. Deng H, Li X (2022) Anomaly Detection via Reverse Distillation from One-Class Embedding. In:2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 9727–9736. IEEE, New Orleans, LA, USA. https://doi.org/10.1109/CVPR52688.2022.00951
9. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, pp 248–255
10. Bergmann P, Fauser M, Sattlegger D, Steger C (2020) Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In:2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4182–4191. https://doi.org/10.1109/CVPR42600.2020.00424 . https://ieeexplore.ieee.org/document/9157778/
11. Thomine S, Snoussi H, Soua M (2023) MixedTeacher : Knowledge Distillation for fast inference textural anomaly detection. In:Proceedings of the 36th international conference on machine learning, Lisbonne. https://doi.org/10.48550/arXiv.2306.09859
12. Tan M, Le Q (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In:Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research, vol. 97, pp. 6105–6114. https://proceedings.mlr.press/v97/tan19a.html
13. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR), 770–778

14. Han S, Hu X, Huang H, Jiang M, Zhao Y (2022) Adbench: Anomaly detection benchmark. Adv Neural Inform Process Syst 35:32142–32159
15. Mei S, Yudan W, Wen G (2018) Automatic Fabric Defect Detection with a Multi-Scale Convolutional Denoising Autoencoder Network Model. Sensors 18, 1064. https://doi.org/10.3390/s18041064
16. Nguyen QP, Lim KW, Divakaran DM, Low KH, Chan MC (2019) GEE: a gradient-based explainable variational autoencoder for network anomaly detection. 2019 IEEE conference on communications and network security (CNS), 91–99
17. Zavrtanik V, Kristan M, Skocaj D (2021) DRÆM - A discriminatively trained reconstruction embedding for surface anomaly detection. In:2021 IEEE/CVF international conference on computer vision (ICCV), pp. 8310–8319. IEEE, Montreal, QC, Canada. https://doi.org/10.1109/ICCV48922.2021.00822 . https://ieeexplore.ieee.org/document/9710329/
18. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: advances in neural information processing systems, pp. 2672–2680
19. Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth UM (2019) f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med Image Anal 54:30–44
20. PourReza M, Mohammadi B, Khaki M, Bouindour S, Snoussi H, Sabokrou M (220) G2D: generate to detect anomaly. 2021 IEEE winter conference on applications of computer vision (WACV), 2002–2011
21. Liang Y, Zhang J, Zhao S, Wu R-C, Liu Y, Pan S (2022) Omni-frequency channel-selection representations for unsupervised anomaly detection. IEEE Trans Image Process 32:4327–4340
22. Rezende D, Mohamed S (2015) Variational inference with normalizing flows. In: proceedings of the 32nd international conference on machine learning. Proceedings of machine learning research, vol. 37, pp. 1530–1538. Lille, France. https://proceedings.mlr.press/v37/rezende15.html
23. Hinton G, Vinyals O, Dean J (2015) Distilling the Knowledge in a Neural Network. In: NIPS deep learning and representation learning workshop. http://arxiv.org/abs/1503.02531
24. Salehi M, Sadjadi N, Baselizadeh S, Rohban MH, Rabiee HR (2020) Multiresolution knowledge distillation for anomaly detection. 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 14897–14907
25. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. In: CoRR, San Diego. https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8
26. Bergmann P, Fauser M, Sattlegger D, 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), pp. 9584–9592. https://doi.org/10.1109/CVPR.2019.00982
27. Computer Vision Group F Tilda Dataset. https://lmb.informatik.uni-freiburg.de/resources/datasets/tilda.en.html
28. Thomine S, Snoussi H (2023) Distillation-based fabric anomaly detection. Textile Res J 405:175. https://doi.org/10.1177/00405175231206820
29. Zagoruyko S, Komodakis N (2016) Wide residual networks. In: proceedings of the British machine vision conference 2016, BMVC 2016, York, UK, September 19-22, 2016. http://www.bmva.org/bmvc/2016/papers/paper087/index.html