Adapting Generic Features to A Specific Task: A Large Discrepancy Knowledge Distillation for Image Anomaly Detection

Chenkai Zhang[1]
ckzhang@zju.edu.cn
Tianqi Du[1]
caleb_du@zju.edu.cn
Yueming Wang[1†]
ymingwang@zju.edu.cn

[1] College of Computer Science
Zhejiang University
Hangzhou, China

[†] Corresponding Author

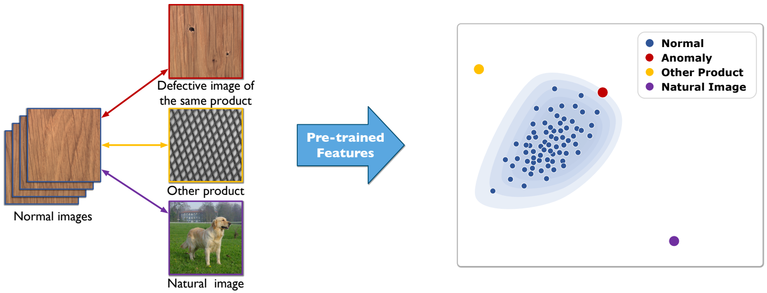
Abstract

Anomaly detection is a challenging task due to the lack of data on unexpected anomalies. Recent approaches using *Knowledge Distillation* (KD) between Teacher-Student (T-S) models have shown great potential for anomaly detection. These techniques use pre-trained models on natural images as the teacher model. However, for industrial images, defects typically occur in a small region, while the global semantics of the anomaly image remain similar to normal images. This situation results in generic features being unable to capture defects well, leading to a loss of discriminability in detecting anomalies. This paper proposes a way to improve this situation by applying learnable feature mappings to adapt the generic features for the data-specific task. Additionally, a novel angular margin loss is introduced to improve the regular training loss of knowledge distillation and ensure larger discrepancies between T-S models on anomalies. Extensive experiments show that the proposed feature mappings and angular loss can effectively improve the feature discriminability for anomaly detection and help state-of-the-art KD-based methods achieve better detection performance.

1 Introduction

Anomaly detection is an important and challenging task in many domains of computer vision. Typically, the population of normal examples and anomalies is heavily imbalanced, and it is often infeasible to enumerate all possible anomalies [8]. To address this issue, unsupervised anomaly detection methods have been proposed, where a normal profile is learned solely from normal examples, and then samples that do not conform to this profile are identified as anomalies. While many approaches rely on image reconstruction, this method has been observed to have poor detection performance as it only learns a pixel-to-pixel mapping [16].

Recently, it has been shown that pre-trained models on large image datasets are highly effective for many downstream tasks. Several recent works [2, 18, 21] follow the idea



(a) Images (b) Features

Figure 1: (a) Comparison of normal images with defective anomaly image, other product image, and natural image. The anomaly exhibits the highest visual similarity to the normal images. (b) Demonstration in the feature space of a pre-trained model (ResNet-18 on ImageNet) reveals that the anomaly sample is the most inseparable from the normal samples.

of *Knowledge Distillation* (KD) [9], which involves training a student model to distill the knowledge of normal data only from a pre-trained teacher model. During testing, the behavior discrepancies between T-S models on unseen anomalous data are used to detect anomalies. This approach leverages the powerful knowledge condensed from large datasets to anomaly detection with less data. Furthermore, using feature-level rather than pixel-level reconstruction avoids the possibility of learning shortcuts for pixel-to-pixel mapping [25].

However, pre-trained models have limited capability for anomaly detection due to two main reasons. First, the teacher models used for KD are pre-trained on natural images for general purposes and are not directly applicable to industrial data due to the domain gap [7]. Second, for the anomaly detection task, anomalous and normal images of a product share the same semantics, as defects occur only in a small region, making it challenging for generic features to obtain sufficient discriminability about them. As shown in Figure 1, normal and anomalous images have very similar content, and the anomalous sample in the feature space is close to the distribution of its normal samples. In contrast, other products or natural images are usually far from this distribution. Therefore, detection methods using pre-trained models face the challenge that the teacher model cannot produce distinguishable features for normal and anomalous data, resulting in only minor discrepancies between the T-S models for unseen anomalous data. Recently, Deng *et al*. [5] attempted to address this issue by designing an asymmetric architecture to increase feature differences and achieve better performance. However, its training process still aims to align the feature space from natural images, so the above problem remains.

This paper presents a method called Feature Mapping with Angular Margin (FMAM) to tackle the above problem. In FMAM, feature mapping (FM) is designed to adapt generic features of pre-trained models to the anomaly detection task, enabling them to obtain more discriminative features for detection. The angular marginal loss (AML) is proposed to improve the learned FM during the training process. Our experimental and analytical results demonstrate that the proposed feature mapping can help different KD-based detection methods achieve better performance on diverse datasets and improve the discriminability of anomalies in the feature space.

Our contributions can be summarized as follows: First, we propose the utilization of learnable feature mappings in pre-trained models, which allows for the acquisition of more discriminative features for the task of anomaly detection. Second, we introduce the application of angular marginal loss to enhance the training process and enable improved learning of feature mappings in pre-trained models. Third, we validate the effectiveness of our proposed method on various datasets and detection methods, demonstrating its efficacy and versatility.

2 Related Work

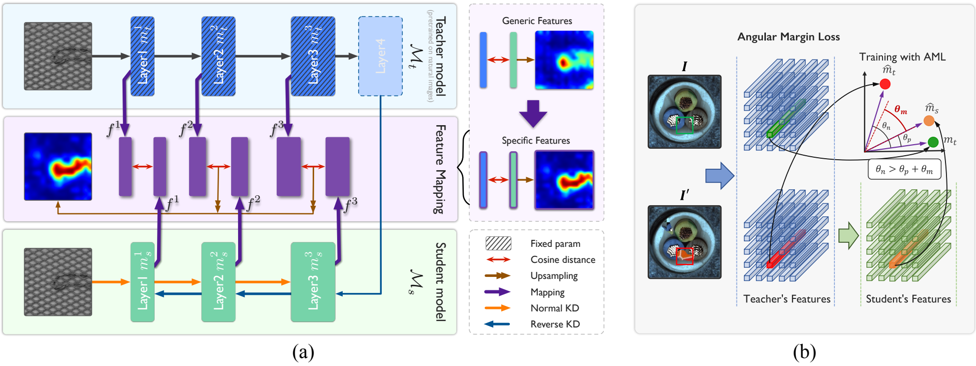
Image Reconstruction-based Methods. Reconstruction-based techniques assume anomalies can be detected through reconstruction errors since models trained on normal data cannot accurately reconstruct them. However, recent research has found that these models can effectively reconstruct anomalies, resulting in missed anomaly detection [16]. Alternative methods, such as feature memory banks [16] and self-supervised learning techniques, have been explored to overcome this issue, including applying strip masks [24], playing jigsaw puzzle games [19], and restoring superpixel segments [12]. However, these methods lack the ability to capture high-level semantic information.

Feature Distillation Methods. In accordance with the principles of *Knowledge Distillation* [9], Park *et al.* [17] aim to distill knowledge by considering the relationships among classifier features in order to enhance feature discriminability. However, their approach primarily focuses on natural images and does not optimize features for anomaly detection tasks. To improve detection performance, some methods employ features extracted by pre-trained models as descriptions of raw images. In these approaches, teacher-student (T-S) models are expected to produce more discrepant features for anomalies during inference. For instance, MRKD [18] proposes multilevel feature alignment, while U-Students [2] ensemble several student models trained on normal data. They suggest that using students with a symmetric architecture can reduce information loss [21]. However, RKD [5] proposes using an asymmetric student with reverse feature flow for distillation to improve the discrepancy between asymmetric T-S models furthe.

Angular Loss. The proposed AML is related to several angular loss techniques. L-Softmax [13] was the first to employ angular loss to improve the separability issue of softmax in the image classification task, followed by SphereFace [14], CosFace [22], and ArcFace [6], which enhanced L-Softmax for face recognition by adjusting the position of the marginal terms in the cosine function. These angular softmax methods replace the logits output with cosine similarity, and use angular margins to achieve wider separation intervals for the final classifier [13]. A similar term to ours is used by Kim *et al.* [10], where they apply ArcFace to the softmax of classifiers. However, their results indicate a failure to effectively localize defects using this approach. In contrast, our AML is proposed to use cosine-based distances to improve the low feature discriminability of pre-trained models in anomaly detection. Unlike angular softmax, which is used to generate probabilistic predictions and trained with cross-entropy loss, AML directly optimizes cosine distances over the feature maps.

3 Method

In this section, we first review *Knowledge Distillation* (KD) for anomaly detection. Then, we introduce feature mapping to enhance the pre-trained models for better anomaly detection. We also propose angular margin loss to improve the training of feature mappings.



(a) (b)

Figure 2: (a) Schematic diagram of feature mappings combined with both normal and reverse KD-based detection methods to adapt generic features to specific industrial images, resulting in larger differences from the T-S models. (b) Demonstration of training using angular margin loss with synthesized anomalous image.

3.1 Distillation on Feature Map

Let $\mathcal{D}_{train} = \{I_1, I_2, \ldots, I_N\}$ denote a training dataset consisting solely of normal images, where $I \in \mathbb{R}^{h \times w \times c}$. Suppose a model pre-trained on ImageNet is the teacher model $\mathcal{M}_t$. A student model $\mathcal{M}_s$, which has a symmetric or asymmetric architecture with respect to the teacher model, i.e., normal KD or reverse KD in Figure 2, to mimic the outputs of the teacher $\mathcal{M}_t$. The primal KD only transfers the knowledge of logits before softmax for classification. However, for anomaly detection, the knowledge transfer is typically designed to align multilevel intermediate features between the teacher $\mathcal{M}_t$ and student $\mathcal{M}_s$ as follows:

$$\mathcal{L}_{KD} = \frac{1}{L} \sum_{j=1}^{L} D(m_t^j,\, m_s^j), \tag{1}$$

where $m_s^j \in \mathcal{M}_s(I)$ and $m_t^j \in \mathcal{M}_t(I)$ are point-wise features of image I, extracted from the j-th layer from the student and teacher models, respectively. The rationale behind this is that the student model learns the teacher's knowledge solely from the anomaly-free data, and its behavior will be inconsistent with the teacher model on unseen anomalous data. Consequently, such a behavioral discrepancy in test anomalies can be used to detect and locate them. $D(*)$ is a vector-wise distance function to measure discrepancies between their intermediate features. During testing, the unflattened vector-wise distances of features enable us to detect and localize anomalies from different levels.

3.2 Feature Mapping

The models pre-trained on natural images often fail to produce high-discriminative features for industrial anomaly detection due to the following factors: limited pre-training on natural images, similar background semantics in industrial product images (as shown in Figure 1), and the difficulty of detecting anomalies manifested in minor and subtle regions.

To address this challenge, we propose to use feature mappings (FM) on generic features to better adapt pre-trained models for the anomaly detection task. Since knowledge transfer for anomaly detection typically occurs in multiple layers, we will learn separate feature mappings for different levels to achieve the best discrimination between the teacher and student models. Specifically, we denote $f^j(*)$ as the feature mapping at the j-th level, which can be implemented by using convolutional or fully-connected networks. Then, the discrepancy between teacher and student models can be calculated as follows:

$$D^j\left(f^j(m_t^j(I)),\ f^j(m_s^j(I))\right). \tag{2}$$

3.3 Angular Margin Loss

To enhance the feature discriminability for anomaly detection, we propose to train the feature mapping with angular margin loss (AML). Similar to [5], we utilize cosine similarity to measure the difference in feature maps between the T-S models. The cosine-based distance D_{cos} can be derivative from the cosine similarity as follows:

$$D_{cos}(\theta) = 1 - \cos(\theta) = 1 - \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \cdot \|\mathbf{v}_2\|_2}, \tag{3}$$

where θ refers to the angle between two exemplary feature vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, and can be calculated by using *arccos* function. The cosine distance measures the difference between two vectors in the inner product space, thus its value solely depends on the angle rather than the magnitudes of them. By calculating the vector-wise distance along the channel dimension for feature maps, the T-S models are able to generate a 2D anomaly detection maps during inference.

Our approach is based on the concept that the distance between feature vectors can be represented by the angle between them, and aims to improve the ability to differentiate between normal and anomalous features. To achieve this, the proposed AML employs a form of triplet loss [20, 23] to increase θ_n, the angle between normal and anomaly features, while simultaneously minimizing θ_p, the angle between normal features, by a angular margin of θ_m. As shown in Figure 2(b), the angular margin loss requires that the angle of negative pairs exceed the angle of positive pair by a fixed angle margin, i.e., the following requirement needs to be satisfied:

$$\theta_p + \theta_m < \theta_n. \tag{4}$$

Specifically, the learning of AML involves training on two pairs of samples and is optimized by reducing the angles between positive pairs and expanding the angles between negative pairs. According to Equation 3 and Equation 4, the angular margin loss can be defined as follows:

$$D_{AML}(f(m_t),\ f(\hat{m}_s),\ f(\hat{m}_t)) = \max(0,\ D_{cos}(\theta_p + \theta_m) - D_{cos}(\theta_n)), \tag{5}$$

and these angles are

$$\begin{cases} \theta_m \in [0,\ \pi] \\ \theta_p = \arccos[\frac{f(m_t)\cdot f(\hat{m}_s)}{\|f(m_t)\|_2 \cdot \|f(\hat{m}_s)\|_2}] \\ \theta_n = \arccos[\frac{f(m_t)\cdot f(\hat{m}_t)}{\|f(m_t)\|_2 \cdot \|f(\hat{m}_t)\|_2}] \end{cases}, \tag{6}$$

where θ_m is scaler of the angle margin that needs to be specified during the training, and $\hat{m} \in \mathcal{M}(I')$ represents the feature of a negative image I' with synthetic anomalies. These anomalies are not real but are generated using commonly employed strategies in self-supervised anomaly detection [11, 24]. This procedure will be described in the next subsection.

3.4 Anomaly Synthesis

To train AML in Equation 7, we basically follow methods of [26] and [11] to generate synthesized samples. The generation process, as illustrated in Figure 3, involves fusing structure and texture information.

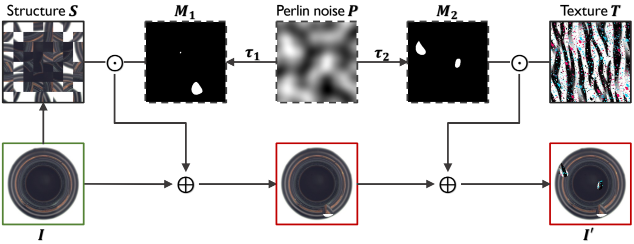


Figure 3: The synthesis process of fusing structure and texture information.

Specifically, the entire image is first divided into 64 patches, which are then randomly rotated and shuffled to obtain the structure information S. Texture information T is drawn from the DTD dataset and resized. To fuse the structure and texture information, masks M_1 and M_2 are generated by binarizing Perlin noise P using symmetry thresholds of τ and $-\tau$. In the figure, the symbol $\odot$ represents the Hadamard product, while $\oplus$ represents element-wise addition. For a more detailed process, interested readers are encouraged to refer to the work by Zavrtanik *et al.* [26].

Finally, feature mappings with AML are trained as follows:

$$\mathcal{L}_{AML} = \frac{1}{L}\sum_{j=1}^{L} D_{AML}(f^j(m_t^j),\ f^j(\hat{m}_s^j),\ f^j(\hat{m}_t^j)). \tag{7}$$

4 Experiment

In this section, we present extensive experiments conducted on different datasets to evaluate the effectiveness of our proposed methods. We apply these methods to both symmetric and asymmetric KD detection methods and compare them with recent methods. Additionally, we conduct analytical experiments to further validate the proposed method.

4.1 Experiment Setup

4.1.1 Datasets and Metrics

Datasets: MVTec Dataset [1] is an industrial image anomaly detection dataset including 5,354 high-resolution images in 15 categories of objects and textures. ZJU-Leaper [27] is a fabric dataset that contains 98,777 images of 15 patterns with different texture complexity. For data of each category, the training set only includes defect-free images, and the test set comprises both defect-free images and defective images with various types of detects. For all experiments, images are resized to a size of 256×256.

Evaluation Metrics: We adopt the Area Under the Curve of Receiver Operating Characteristics curve (AUC of ROC) to evaluate the performance on image-level and pixel-level detection results.

Table 1: The performance comparison on MVTec dataset and ZJU-Leaper Dataset.

Category	Method	MVTec		ZJU-Leaper	
		Image AUC	Pixel AUC	Image AUC	Pixel AUC
Feature Space	PuzzleAE [19]	71.1	80.7	69.1	68.2
	FCDD [15]	86.6	92.5	58.0	61.6
	SPADE [3]	85.5	96.0	83.3	88.8
	PaDiM [4]	90.3	96.1	84.8	86.3
Symmetric KD	MRKD [18]	87.7	90.7	86.9	82.3
	NKD	94.7	96.6	84.9	92.7
	NKD+FMAM	**96.7**	**96.9**	**88.6**	**93.6**
Asymmetric KD	RKD [5]	96.1	97.1	89.8	93.8
	RKD+FMAM	**98.2**	**97.3**	**91.9**	**94.7**

4.1.2 Anomaly Scores

According to Equation 3, we obtain a set of D^j from the various layers of the T-S models, where j represents the vector-wise discrepancy map in the j-th layer. To merge these results, bilinear interpolation Φ is used to upsample each D^j to the image size, and then add them together. In order to reduce noise and improve the interpretability of final detections, we apply a Gaussian filter with sigma=4, as recommended in many previous studies. For the classification, averaging the pixel score map S_L directly is not reasonable for images with only small anomalies. Thus, we average only the top k values (empirically set as 100) of S_L to obtain the sample-level score S_C. The pixel-level score map S_L an image-level score S_C can be formulated as follows:

$$S_L = \sum_{j=1}^{L} \Phi(D^j(m_t^j,\ m_s^j)), \quad S_C = \frac{1}{k}\sum_{i=1}^{k} \mathrm{top}_k(S_L). \tag{8}$$

4.1.3 Implementation Details

In all experiments, the ResNet-18 pre-trained on ImageNet is used as the backbone for both teacher and student. To measure the discrepancy, we use features from the first three layers of the four-layer architecture (i.e. $j = \{1,2,3\}$). The symmetric student model is implemented following contemporary works [2, 18, 21], and the asymmetric student is implemented according to Reverse Distillation [5]. For the implementation of feature mappings, we empirically find that two-layer MLPs are good enough. In order to ensure that the feature mapping does not rely on randomly initialized features of the student model, our training process consists of two stages. In the first stage, we train the ordinary T-S models for 50 epochs using the KD loss. We employ the Adam optimizer with a learning rate of $1e^{-3}$ and a weight decay of $1e^{-4}$. In the second stage, we optimize the T-S models with FMAM are optimized. This stage involves training for an additional 50 epochs using the Adam optimizer with a learning rate of $1e^{-4}$ and an L1 weight decay of $1e^{-5}$. Furthermore, it is worth noting that there are two implementations of the feature loss [5]. We will discuss both implementations and their respective results in the supplementary materials.

4.2 Results

Table 1 presents the performance of different methods on the MVTec and ZJU-Leaper datasets, reported with sample-level and pixel-level detection evaluation metrics. The results of feature-

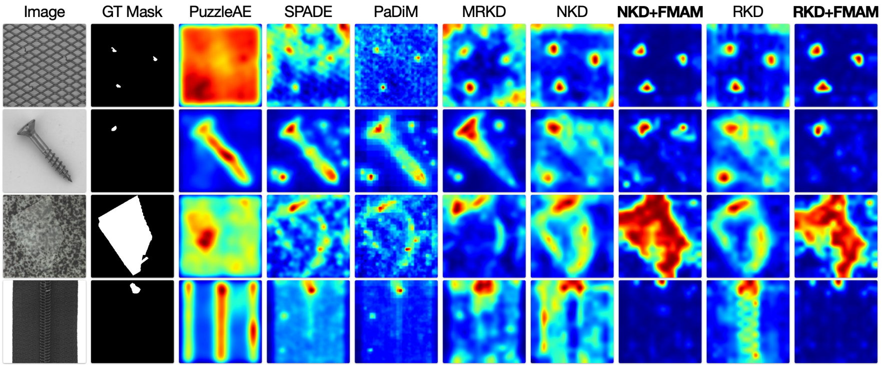
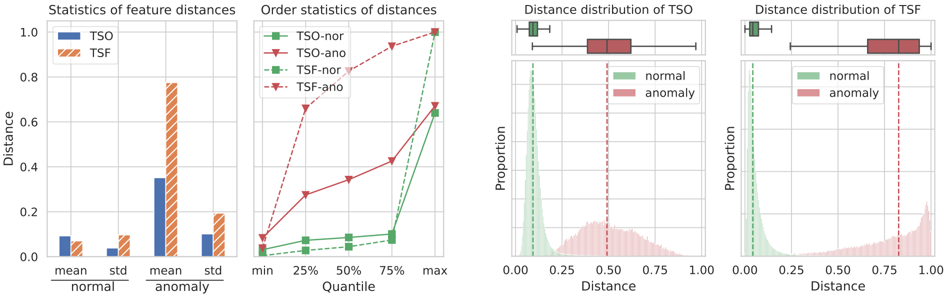


Figure 4: Visual comparison with other SOTA detection methods on the MVTec dataset.

based methods, except for PuzzleAE [19] and FCDD [15], are obtained from ResNet-18 pre-trained on the ImageNet. NKD (Normal Knowledge Distillation) [2, 5, 18, 21] is a representative baseline of symmetric KD methods, with an improved implementation based on these conventional KD detection methods. RKD (Reverse Knowledge Distillation) [5] is a baseline of asymmetric KD methods where the student uses a reverse feature flow to the teacher model, resulting in better performance. Although RKD uses an asymmetric structure to extend feature discrepancy between T-S models, their features are still derived from a model pre-trained on natural images, so there is still room for the proposed FMAM to make improvements. The table shows that both NKD and RKD can obtain improvements from the proposed feature mapping. Notably, while the improvements in pixel-level metrics are less noticeable, the visualizations in Figure 4 demonstrate that the proposed feature mapping can help models produce less noisy heatmaps. It also suggests that the pixel-level AUC may be less informative due to the massive amount of normal pixels, as reported by Zhang *et al.* [27].

4.2.1 Analysis Experiment

To examine the effectiveness of FMAM in enhancing KD detection methods, we conduct a comprehensive analysis of the feature distances between the teacher and student models. For this purpose, we consider the original T-S models (TSO) as well as the T-S models trained



(a) distances

(b) deviations

Figure 5: Statistics of feature distances of distance deviations

with FMAM (TSF).

Figure 5(a) presents statistics of the feature distances. The mean and std analyses on the left side of the figure demonstrate that TSO and TSF have similar feature differences in normal areas. In comparison, TSF can yield more significant feature differences in anomalous areas compared to TSO. The order statistics for the different quartiles of these distances on the right side reveal that TSF can obtain not only greater feature differences in anomalous regions (dashed red line above the solid red line) but also smaller feature differences in normal regions (dashed green line below the solid green line). These indicate that the T-S models using generic features can effectively increase the differences between anomalous features and reduce the differences between normal features through FMAM, thus improving the discriminability of generic features for anomalies.

On the other hand, Figure 5(b) presents specific distributions of feature distances. The results show that the distribution of feature distances for TSF has a larger distribution interval about normal and abnormal compared to TSO, which enables the feature differences to better reflect the abnormalities in the test image. The medians of distributions (indicated by the dashed lines) and the box-and-whisker plot at the top further illustrate the superiority of TSF in the distribution of feature distances.

In summary, our experimental results demonstrate that FMAM can effectively improve the feature discriminability for detecting anomalies with generic features. As a result, T-S models pre-trained on natural images can achieve better performance on the anomaly detection task.

4.2.2 Ablation Studies

In order to analyze the contribution of the proposed Feature Mapping and Angular Margin (FMAM) method, several ablation studies are conducted on the MVTec dataset.

Table 2 presents the results of the RKD model training using different distance functions of Equation 1. Our findings confirm the suggestion by [5] that using cosine similarity to measure the feature differences of KD-based models can yield better detection performance. Therefore, all our experiments employ cosine distance for training Teacher-Student (T-S) models.

Table 3 displays the ablation results of the two components of the proposed FMAM,

Table 2: Ablation of distance functions of RKD (Equation 1)

Distance	Image AUC	Pixel AUC
L1	83.3	96.1
L2	82.1	95.4
cos	96.1	97.1

Table 3: Ablation of FMAM

Model	Image AUC	Pixel AUC
RKD	96.1	97.1
RKD+FM	96.8	97.2
RKD+AML	96.9	97.2
RKD+FMAM	98.2	97.3
RKD+FM(L1ML)	93.9	96.6
RKD+FM(L2ML)	93.0	90.5

Table 4: Ablation of other backbones

Model	Image AUC	Pixel AUC
RKD(res34)	98.3	97.2
RKD(res34)+FMAM	98.3	97.4
RKD(res50)	98.5	97.6
RKD(res50)+FMAM	98.8	97.9
RKD(wres50)	98.5	97.7
RKD(wres50)+FMAM	99.1	98.1

Table 5: Margins of AML

Margin	Image AUC	Pixel AUC
5°	97.9	97.1
15°	98.2	97.3
30°	98.1	97.3
60°	97.4	97.2
90°	97.0	97.2

namely Feature Mapping (FM) and Angular Margin Loss (AML). The first row represents the baseline performance of the RKD method. Subsequent rows demonstrate the individual utilization of FM or AML in training the T-S models, resulting in only marginal performance improvements. However, when FM and AML are combined, a significant enhancement in performance is observed. To explore alternative margin loss formulations, such as employing L1-norm and L2-norm based margin losses (referred to as L1ML and L2ML), for training T-S models, further investigation was conducted. The results demonstrate that the traditional Euclidean triplet loss fails to yield any improvement, and even negatively impacts the performance of T-S models. It is worth noting that all T-S models were trained in an end-to-end manner.

Table 4 presents an investigation of the proposed FMAM method using different backbone models. The results reveal that larger models generally display superior detection capabilities. Remarkably, our method significantly enhances the detection performance across different backbone models. Notably, our method achieves outstanding results when employing the largest backbone, WideResNet-50, with an image AUC of 99.1 and a pixel AUC of 98.1.

To examine the impact of different margins, a line search for θ is conducted and the results are reported in Table 5. For non-negative features after ReLU activation, a reasonable range of margin values is considered as $[0^\circ, 90^\circ]$. Our findings reveal that a margin of 15° yields the most favorable results for the FMAM method.

5 Conclusion

This paper argues that the models pre-trained on natural images produce suboptimal discriminative features for industrial images. To address this issue, we propose the utilization of feature mappings to adapt pre-trained models and obtain superior feature discriminability for the anomaly detection task. Additionally, training feature mappings with the proposed angular margin loss can further increase feature discriminability. Through extensive experiments, we demonstrate the effectiveness of the proposed approach.

6 Acknowledgments

This work was partly supported by grants from the National Natural Science Foundation of China (62336007), the Key R&D Program of Zhejiang (NO.2022C03011), the Chuanqi Research and Development Center of Zhejiang University, the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-002), and the Fundamental Research Funds for the Central Universities.

References

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proc. CVPR*, pages 9584–9592, June 2019.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed

Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings. In *Proc. CVPR*, pages 4183–4192, 2020.

[3] Niv Cohen and Yedid Hoshen. Sub-Image Anomaly Detection with Deep Pyramid Correspondences. *arXiv:2005.02357 [cs]*, February 2021.

[4] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Proc. ICPR*, November 2020.

[5] Hanqiu Deng and Xingyu Li. Anomaly Detection via Reverse Distillation from One-Class Embedding. In *Proc. CVPR*, March 2022.

[6] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 1939-3539.

[7] Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, March 2022. ISSN 1558-2531.

[8] Jiawei Han, Jian Pei, and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, June 2011. ISBN 978-0-12-381480-7.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *Proc. NeurIPS*, March 2015.

[10] Taehyeon Kim, Eungi Hong, and Yoonsik Choe. Deep Morphological Anomaly Detection Based on Angular Margin Loss. *Applied Sciences*, 11(14):6545, January 2021. ISSN 2076-3417.

[11] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In *Proc. CVPR*, 2021.

[12] Zhenyu Li, Ningyang Li, Kaitao Jiang, Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Superpixel Masking and Inpainting for Self-Supervised Anomaly Detection. In *Proc. BMVC*, 2020.

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-Margin Softmax Loss for Convolutional Neural Networks. In *Proc. ICML*, November 2017.

[14] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *Proc. CVPR*, January 2018.

[15] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable Deep One-Class Classification. In *Proc. ICLR*, March 2021.

[16] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-guided Normality for Anomaly Detection. In *Proc. CVPR*, pages 14360–14369, March 2020.

[17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *Proc. CVPR*, pages 3967–3976, 2019.

[18] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution Knowledge Distillation for Anomaly Detection. In *Proc. CVPR*, 2021.

[19] Mohammadreza Salehi, Ainaz Eftekhar, Niousha Sadjadi, Mohammad Hossein Rohban, and Hamid R. Rabiee. Puzzle-AE: Novelty Detection in Images through Solving Puzzles. *Computer Vision and Image Understanding*, 2022.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proc. CVPR*, pages 815–823, June 2015.

[21] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-Teacher Feature Pyramid Matching for Unsupervised Anomaly Detection. In *Proc. BMVC*, March 2021.

[22] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *Proc. CVPR*, April 2018.

[23] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *Proc. NeurIPS*, volume 18. MIT Press, 2005.

[24] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning Semantic Context from Normal Samples for Unsupervised Anomaly Detection. In *Proc. AAAI*, volume 35, pages 3110–3118, May 2021.

[25] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A Unified Model for Multi-class Anomaly Detection. In *Proc. NeurIPS*, October 2022.

[26] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proc. ICCV*, September 2021.

[27] Chenkai Zhang, Shaozhe Feng, Xulongqi Wang, and Yueming Wang. ZJU-Leaper: A Benchmark Dataset for Fabric Defect Detection and a Comparative Study. *IEEE Transactions on Artificial Intelligence*, 1(3):219–232, December 2020. ISSN 2691-4581.