2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)

Enhancing Reverse Distillation for Anomaly Detection through Variance Loss Optimization

Ziwei Song
School of Engineering
Unvercity of Yamanashi
Kofu, Japan
g22dtsa2@yamanashi.ac.jp

Prawit Buayai
School of Engineering
Unvercity of Yamanashi
Kofu, Japan
buayai@yamanashi.ac.jp

Xiaoyang Mao*
School of Engineerin
Unvercity of Yamanashi
Kofu, Japan
mao@yamanashi.ac.jp

***Abstract*—Anomaly detection methods using unsupervised learning are widely applied in industrial defect inspection due to their robustness against diverse anomaly types and data imbalances. These methods rely on the inability of feature extractors trained on normal data to replicate anomalous features, enabling anomaly identification through feature deviations. However, real-world challenges, such as high variability within normal data, often hinder the accurate separation of normal and anomalous samples. We propose a novel method specifically designed to improve the performance of the Reverse Distillation (RD) model for anomaly detection. This method addresses the variability of normal data by enabling the model to focus on extracting common features while overlooking intraclass variances. The effectiveness of the proposed method is demonstrated through experiments conducted on MVTec AD dataset. The results demonstrate that the proposed methodology improves performance in both anomaly detection and localization, demonstrating our proposed approach's effectiveness.**

Keywords—Anomaly detection, Unsupervised learning, Reverse distillation

I. Introduction

Anomaly detection (AD) and localization aim to identify anomalous images and locate sub-regions exhibiting abnormalities. These techniques have a wide range of applications [1-4], especially in industrial quality inspection [5-7]. Anomalies often occupy only a small fraction of the entire image, making their detection particularly challenging. Fig. 1 illustrates various types of anomalies found in different industrial products from the MVTec AD dataset [8]. Moreover, the rarity and diversity of anomalies pose significant challenges in the collection and annotation of data required for training discriminative models. Due to their infrequent occurrences and varied manifestations, assembling a sufficiently large dataset becomes exceedingly difficult, resulting in highly imbalanced training sets dominated by normal samples.

Existing methods use various approaches to solve this problem. The augmentation-based anomaly detection methods have achieved state-of-the-art performance in both anomaly detection and localization. For instance, CutPaste [7] creates synthetic anomalies by cutting patches from images and pasting them onto random locations, introducing perturbations to spatial arrangements. Similarly, MSTUnet [8] employs anomaly simulation and masking strategies to generate diverse anomalous training data, enabling end-to-end anomaly detection and localization. SimpleNet [5] augments anomalous data by adding noise at the feature space, and differentiate normal and abnormal by multi-layer perceptron. RealNet [19] generates anomalies using Strength-controllable Diffusion Anomaly Synthesis. However, since the nature of anomalous data is unpredictable, models based on data augmentation may have a limited generalization ability and the risk of failing to detect some anomalies.

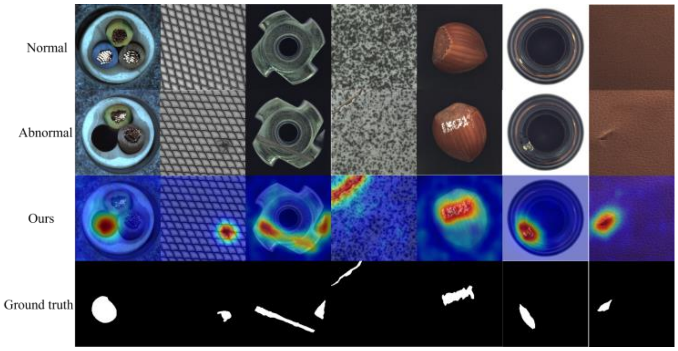

Fig. 1.The visualization of poposed method on MVTec dataset.

Another prominent approach tackles anomaly detection in an unsupervised manner, using only normal samples during training. A common approach in those methods is the use of generative models, such asccr (AE) [9], and Generative Adversarial Network (GAN) [10], to learn the underlying feature distribution of normal data. Embedding-based methods [11,12], extract features from normal images using pre-trained convolutional neural networks. These features are then mapped into a representation space through Gaussian distribution [11] or normalization [12] approaches. Anomalies are identified by measuring the deviation of test sample features from the normal feature distribution within this embedding space. However, these methods struggle with complex data distributions, which can result in poor accuracy in real-word applications.

An alternative solution involves knowledge distillation (KD) frameworks [13,14]. Bergmann et al. [15] introduced a teacher-student framework for unsupervised anomaly detection, demonstrating strong performance on various datasets. Salehi et al. [16] implemented a teacher-student network architecture where knowledge is transferred from the teacher to the student. The student is trained exclusively on normal samples, allowing it to capture their distribution. Consequently, when anomalous queries are encountered during inference, the student is expected to generate out-of-distribution representations. Deng and Li [17] identified limitations in teacher-student models due to similar architectures and identical data flow. They proposed Reverse Distillation (RD), utilizing a One-Class Bottleneck Embedding

979-8-3315-0976-7/25/$31.00 ©2025 IEEE

(OCBE) to transfer teacher outputs to the student, achieving high performance with low latency.

Despite notable advancements in anomaly detection, there remain challenges in achieving high accuracy, reliable and fast inference for real-world applications. Reverse Distillation (RD) [17], with its multi-layer feature comparison approach, demonstrates superior effectiveness compared to GAN-based methods. By leveraging hierarchical feature representations, RD achieves more precise anomaly detection and localization. However, it still encounters limitations in real-world scenarios. As noted in [20], the distillation task and OCBE module in RD fail to provide compact representations.

To address these issues, we propose a targeted improvement to the RD framework by eliminating variance among normal data features. Unlike other solutions that alter model architecture and increase inference time, our approach introduces a novel variance loss, seamlessly integrated into the RD framework. This variance loss specifically minimizes intra-class variance within normal data by calculating cosine distances among batch samples during feature extraction. The resulting tightly clustered feature space for normal samples enhances compactness, suppresses the propagation of normal data variability through convolutional layers, and improves the model's ability to distinguish normal from abnormal instances.

The contributions of the proposed method are summarized as follows:

1) We proposed a simple yet effective loss, which enable the RD model to focus on common features while eliminating the variance of normal data.

2) The integration of the proposed loss into typical models is outlined, including the selection of feature layers, the use of cosine distance as the similarity metric, and the determination of optimal weight parameters to ensure effective application and enhanced performance.

3) The proposed method was validated MVTec AD dataset, demonstrating its ability to enhance the performance on both anomaly detection and anomaly localization tasks.

II. Related Work

A briefly reviews previous efforts anomaly detection in this section.

Augmentation-based anomaly detection makes use of both normal and anomalous data. Most of the anomalous data is obtained by adding noise, shadow, mask, etc. to the anomaly free data. CutPaste [18] proposed a simple strategy for generating synthetic anomalies for anomaly detection, which involves cutting image patches and pasting them to random locations in a large image. CNNs are trained to discriminate between images from normal and enhanced data distributions. AnoSeg [28] goes with anomalous samples generated by utilizing hard augmentation, adversarial learning and coordinate channel concatenation to train a GAN model. SimpleNet [5] enhances anomaly detection by augmenting anomalous data through the addition of noise directly in the feature space. RealNet [19] proposed Strength-controllable Diffusion Anomaly Synthesis to generate nature anomalies and refines features, balancing detection performance and efficiency.

Auto-Encoder and GANs are commonly used methods for unsupervised anomaly detection. Since models are trained solely on normal data, they struggle to reconstruct anomalous regions, resulting in large reconstruction errors. Some methods [21-23], consider anomaly detection as a repair problem, where patches in the image are masked randomly and then, the model is trained to predict the information of the mask. DRÆM [6] proposed a network that is trained for discrimination in an end-to-end manner. GAN [6] networks consist of a generator, which tries to generate samples similar to real data, and a discriminator, which tries to distinguish between real data and generated samples. The Structural Similarity Index (SSIM) [24] and cosine distance are widely used in the training of unsupervised methods [25], to measure the difference between input and output. Anomaly localization is generated as the pixel-level difference between the input image and its reconstructed image.

Some methods focus on modeling the distribution of normal patterns through a parametric approach, embedding normal data features into a specialized space while mapping anomalous features away from the normal clusters within the embedding space. SPADE [26] embeds normal images into a memory bank that stores the normal feature, leveraging the K-nearest neighbor algorithm to compute the anomaly score. PaDiM [11] utilizes multivariate Gaussian distribution to embed the extracted anomaly patch features. The normal flow is utilized in DifferNet [27] to map the features of the normal image to a specified Gaussian distribution and the anomaly score is estimated by calculating the likelihood value.

Knowledge distillation method utilizes a teacher-student framework to transfer knowledge from a large, pre-trained model (teacher) to a smaller, more efficient model (student), enabling the student model to mimic the teacher's performance while maintaining lower computational costs. Bergmann et al. [15] introduced the teacher-student framework into anomaly detection, demonstrating strong performance on various datasets. Salehi et al. [16] employed multiresolution knowledge distillation to effectively identify unusual features across multiple levels of feature representation. STFPM [29] employs a feature pyramid matching mechanism between a pre-trained teacher network and a student network, generating feature discrepancies to detect anomalies during inference. Deng and Li [17] introduced Reverse Distillation, which uses an encoder-decoder structure to transfer knowledge from the teacher to the student, employing distinct data flows for each model. Tien et al. [20] improved feature compactness and anomaly signal suppression in RD models through multi-task learning. achieving higher accuracy while maintaining fast inference speed.

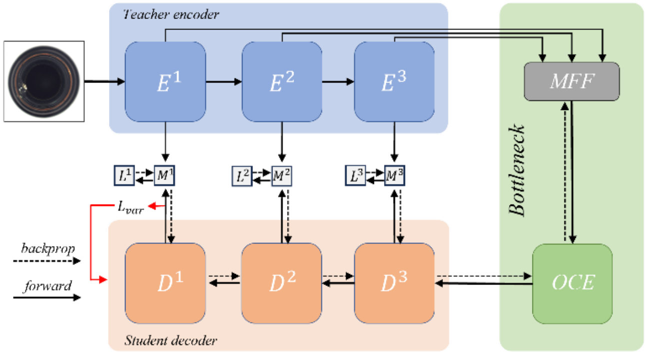

Fig. 2.The RD model structure with proposed loss.

III. Proposed Method

Anomalies are typically identified by detecting deviations from the normal data distribution. Most existing methods, including Reverse Distillation method, rely on distinguishing anomalies from normal data through their feature discrepancies. However, a critical limitation of RD is the inherent variability within normal data, which often propagates through its multi-layer feature comparison framework. This variability can blur the boundary between normal and anomalous features, reducing the model's accuracy in anomaly detection.

To address this, we propose an enhancement tailored to the RD framework by introducing a variance-suppressing mechanism that minimizes intra-class variability within normal data. Our approach introduces a novel variance loss, which operates during feature extraction by calculating cosine distances among batch samples of normal data. This loss encourages features of normal samples to cluster more tightly in the representation space, reducing their variability while preserving the structural integrity of RD’s multi-scale comparison. By focusing exclusively on suppressing variance within normal data in the RD framework, our method amplifies its ability to capture shared characteristics of normal data, enabling sharper differentiation of anomalies. Importantly, this improvement retains RD's inherent strengths, such as efficient inference and robust multi-layer feature analysis, while significantly enhancing its anomaly detection performance.

A. Reverse Distillation Model with Overlooking Variance in Trainig

Unlike traditional data distillation methods, Reverse Distillation creatively applies the data distillation approach within an encoder-decoder structure. This method disrupts the typical consistency of data flow in conventional distillation and challenges the structural similarity between student and teacher models. The RD model comprises three components: the teacher encoder, the bottleneck module, and the student decoder. The teacher encoder E, based on a pre-trained WideResNet on ImageNet [30], is responsible for extracting comprehensive representations. The bottleneck module consists of two sub-modules: Multi-Feature Fusion (MFF) and One-Class Embedding (OCE). The OCBE module compresses multi-scale features into a compact bottleneck representation. The student decoder D features a reversed but symmetrical architecture relative to E. During training, D is designed to replicate the behavior of the teacher encoder.

To improve the RD model's ability to distinguish between normal and anomalous patterns, we propose incorporating a novel variance loss (L_{var}) into the training framework. The architecture of RD method with L_{var} is shown in Fig. 2. The loss function for training the improved RD model is as follows:

$$L = L_{model} + \lambda L_{var} \tag{1}$$

Where L_{model} represents the original reconstruction loss in the RD framework, which measures the similarity between the teacher and student features. L_{var} is the proposed variance loss, which minimizes the variance among features extracted from anomaly-free images. λ is a user controllable parameter to adjust the weight of the variance loss to balances the contributions of L_{model} and L_{var}.

B. Implementation Details

To achieve optimal performance, we systematically analyzed the design and integration of L_{var}.

1) Integration Points of L_{var}

We evaluated the effect of adding L_{var} at different stages of the student decoder, ranging from D^1 to D^3. This analysis allowed us to identify the most effective stage where variance minimization contributes to feature compactness without compromising reconstruction accuracy.

2) Distance Metric

The cosine distance is employed to represent the Cosine similarity, which was got the best performance in the ablation study. Assume the training batch size of RD model is N. a batch of feature maps $\{f_1, f_2 \dots f_N\}$ was got, where f_i is the extracted feature of the i_th image of batch size N. We first reshape each feature map into a vector by follows function:

$$F_i = \{\text{flatten}(f_i)\} \tag{2}$$

Where F_i is the flattened feature vector derived from f_i , reducing its dimensions from $w \times h \times c$ to a one-dimensional vector. And calculate the mean feature vector from the reshaped feature maps by

$$\bar{F} = \frac{1}{N}\sum_{i=1}^{N} F_i \tag{3}$$

Then, compute the cosine similarity between each feature vector and the mean feature vector get by

$$cos_similarity(F_i, \bar{F}) = \frac{F_i \cdot \bar{F}}{\|F_i\|\|\bar{F}\|} \tag{4}$$

Finally, the L_{var} is obtained by calculating the mean of the cosine distances between all flattened feature vectors F_i and $\bar{F}$. Mathematically, L_{var} get by:

$$L_{var} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - cos_similarity(F_i, \bar{F})\right) \tag{5}$$

The proposed method promotes a tighter clustering of normal samples in the feature space, facilitating a clearer separation from anomalies.

TABLE I. COMPARISON OF RD MODEL WITH OUR METHOD ON MVTEC AD DATASET ON IMAGE-BASED ANOMALY DETECTION AND PIXEL-BASED ANOMALY DETECTION.

Method	***RD***	***Ours RD***
Carpet	98.9 / 98.9	**99.9** / **99.3**
Grid	**100** / **99.3**	**100** / 99.2
Leather	**100** / 99.4	**100** / **99.5**
Tile	99.3 / 95.6	**99.5** / **95.8**
Wood	99.2 / 95.3	**99.5** / **95.6**
Bottle	100 / 98.7	**100** / **98.9**
Cable	95.0 / 97.4	**98.6** / **97.9**
Capsule	96.3 / **98.7**	**97.5** / **98.7**
Hazelnut	99.9 / 98.9	**100** / **99.2**
Metal Nut	**100** / **97.3**	**100** / **97.3**
Pill	96.6 / **98.2**	**97.8** / 98.1
Screw	97.0 / **99.6**	**99.2** / **99.6**
Toothbrush	99.5 / **99.1**	**100** / 98.8
Transistor	96.7 / 92.5	**97.5** / **93.1**
Zipper	**98.5** / 98.2	98.4 / **98.7**
Average	98.5 / 97.8	**99.2** / **98.0**

3) *Analyzed the Batch Size Setting*

Batch size plays a crucial role in our method, as the variance loss is calculated by comparing feature vectors within a single batch. An appropriately chosen batch size ensures that the method captures intra-batch consistency effectively, promoting tighter clustering of normal samples in the feature space and facilitating a clearer separation from anomalies. The impact of batch size settings is further analyzed in the ablation study.

4) *Investigated the λ Setting*

The hyperparameter λ balances the influence of L_{model} and L_{var} . A higher λ prioritizes minimizing feature variance, making the model robust to intra-class variability, while a lower λ emphasizes reconstruction fidelity. Choosing an appropriate λ is essential for achieving an optimal trade-off between these objectives.

IV. EXPERIMENT

In this section, we evaluate the proposed method on MVTec AD dataset [8]. a dataset for industrial inspection with high-resolution images of 15 object categories and pixel-level annotations. Specifically, the MVTec AD dataset contains a total of 3466 unlabeled images and 1888 labeled images, with an average resolution of 700×700. The training set consists of 3629 images, all of which are anomaly-free, while the test set contains 1725 images, including both normal and anomalous samples. The dataset is divided into 5 texture classes and 10 object classes, covering 73 types of defects such as scratches, dents, contamination, and deformations.

A. *Implementation Details*

The RD model is trained with the following settings: an input image size of 225 × 225, a learning rate of 0.005, and a batch size of 16. For the evaluation criteria, we show the results by using the area under the receiver operating characteristic curve (AUROC), which plots the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. Each point on the curve represents a specific TPR-FPR value corresponding to a different threshold. All programs run on a server with an AMD EPYC 7543P 32-core Processor and four A100 GPUs.

B. *Experimental Results of RD Model on MVTec*

We applied our approach to the RD model on $D^1$of student encoder. and evaluated it on the MVTec AD dataset. The results, as shown in Table I, demonstrate that our method achieved superior performance in both anomaly detection and anomaly localization tasks. Specifically, in anomaly detection, our approach outperformed the original RD model in 14 out of 15 categories, with significant improvements in categories like Carpet, from 98.9% to 99.9%, and Cable, from 95.0% to 98.6%. For anomaly localization, our method delivered better results in 12 categories, with particularly significant gains in Toothbrush, from 99.5% to 100%, and Transistor, from 96.7% to 97.5%. This indicates that our method is not only capable of detecting anomalies but also excels at pinpointing their exact locations. The average performance of our approach improved from 98.5% to 99.2% on anomaly detection task and 97.8% to 98.0% on anomaly localization task, underscoring its robustness and effectiveness. This demonstrates that the introduction of our variance loss and its integration into the RD framework effectively improve the model's capacity to differentiate anomalies while maintaining high precision.

The visualization of the anomaly localization task is presented in Fig. 3. Compared to the original RD method, our approach generates more precise and concentrated heatmaps for anomaly regions. For instance, On the grid category, RD(Ours) achieves precise localization, avoiding RD's false detections. the RD(Ours) accurately captures the anomaly on the pill surface with a stronger and more concentrated heatmap. On the carpet category, RD(Ours) eliminates noise interference, providing clean anomaly localization, while RD highlights noisy regions despite detecting anomalies. Additionally, RD(Ours) accurately captures the anomaly on the pill surface with a stronger and more focused heatmap. This improvement highlights the robustness of our variance loss and its ability to enhance feature compactness.

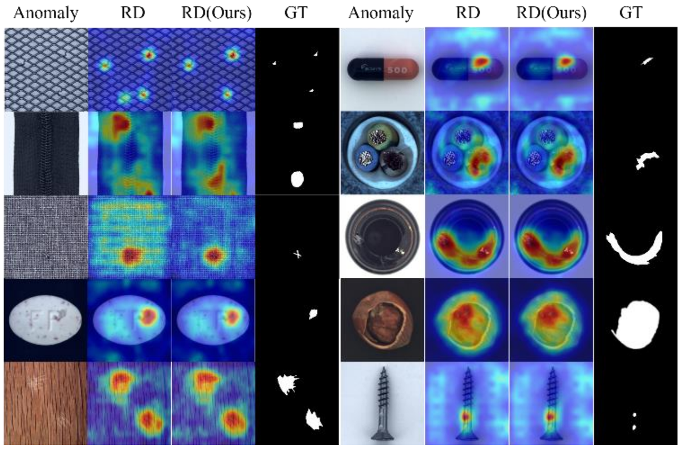

Fig. 3.Visualization examples of proposed method and comparison with the RD method on the MVTec AD dataset.

TABLE II. EFFECTIVENESS VALIDATION OF ANOMALY DETECTION (TIME IN SECONDS).

Metric	With L_{var}	Without L_{var}	Overhead
RD	3308.5	3358.4	1.5%

TABLE III. ABLATION STUDY ON PERFORMANCE WITH VARIOUS DISTANCE METRICS ON METVC DATASET.

Metric	RD	L1	L2	Cos distance
Image level	98.5	56.44	52.93	**99.19**
Pixel level	97.8	63.63	61.62	**97.98**

TABLE IV. THE BATACH SIZE IMPACT TO RD MODEL ON MVTEC DATASET.

Metric	4	8	16	32
Image level	97.87	99.14	**99.19**	99.03
Pixel level	97.21	**97.98**	**97.98**	97.91

TABLE V. THE IMPACT OF FEATURE BLOCK OF RD MODEL ON MVTEC DATASET.

Metric	D^1	D^2	D^3	$D^1 + D^2 + D^3$
Image level	**99.19**	99.0	99.08	98.63
Pixel level	**97.98**	97.78	97.98	97.71

C. Effectiveness Validation

Our approach adds additional losses without changing the model structure, only training time is affected. We compare the time consumption of the anomaly detection methods with and without our proposed approach to training process. The results are shown in Table II. Compared to the original method, using our method only incurred an additional time overhead of 1% to 2%, with the maximum overhead being only 4%. This shows that while the additional loss function introduces some computational overhead, the cost is within an acceptable range.

D. Ablation Analysis

To thoroughly understand the impact of each component of our method, we conducted a series of ablation studies by varying one factor.

Table III presents the performance of the RD model using various distance metrics. We evaluated several common distance measures, and the results demonstrate that cosine distance performs best, achieving the highest scores in both detection and localization tasks. Those results indicate that directly calculating the distance, such as L1 and L2 distances, between features within a batch has minimal impact on reducing variance among feature vectors and may even have negative effects

Table IV illustrates the impact of batch size on detection and localization accuracy. As shown, both detection and localization accuracy initially increase with larger batch sizes, reaching peak performance at a batch size of 16. And when the batch size is further increased to 32, there is a slight decrease in accuracy for both detection and localization metrics, but still better than the original results. It suggests that larger batch sizes help smooth the feature distribution of normal data but may reduce the model's sensitivity to subtle abnormal features.

We also investigated the influence of features extracted from different network layers, as presented in Table V. The full feature set comprises block1 (D^1), block2 (D^2) and block3 (D^3). The results reveal that lower-level features, such as those in D^1, yield higher detection and localization accuracy compared to higher-level features. Specifically, D^1 achieves the best performance for both tasks, suggesting that features from lower layers provide finer-grained details, which are important for accurate anomaly detection and localization.

Additionally, to investigate the effect of weight values λ of the proposed method, we performed experiments with a range of λ settings. The results are shown in Fig. 4. When λ is set too low, the variance loss does not effectively suppress intra-class variability, limiting its ability to guide the model. However, a large λ overemphasizes variance loss, disrupting the balance of the model and leading to degraded performance. Therefore, selecting an appropriate λ value is crucial to achieving an optimal balance between overlooking internal variability and maintaining model stability.

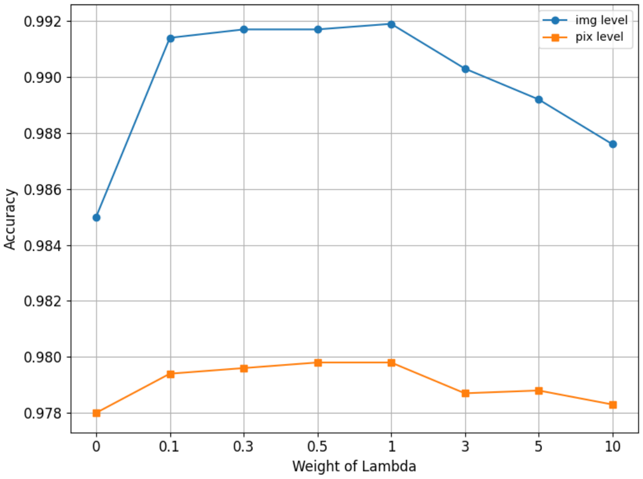

Fig. 4. The λ weight impact on accuracy of RD model on MVTec dataset

V. CONCLUSION

In this study, we propose a new method to improve anomaly detection by minimizing intra-class differences in normal data features. Building upon the RD model, we introduced a specialized loss function to reduce variance within normal data, enhancing feature compactness and improving the model's ability to distinguish anomalies. Through ablation analysis, we identified the optimal placement for integrating this loss within the RD framework and determined the corresponding weighting for best effectiveness. Our approach is simple yet effective, achieving significant improvements in both anomaly detection and localization across two commonly used benchmark datasets, providing a robust solution for practical anomaly detection tasks.

REFERENCES

[1] G. Luo, W. Xie, R. Gao, T. Zheng, L. Chen, and H. Sun, "Unsupervised anomaly detection in brain MRI: Learning abstract distribution from massive healthy brains," Computers in biology and medicine, vol. 154, p. 106610, 2023.

[2] C. Zhang, W. Dai, V. Isoni, and A. Sourin, "Automated anomaly detection for surface defects by dual generative networks with limited training data," IEEE Transactions on Industrial Informatics, vol. 20, no. 1, pp. 421-431, 2023.

[3] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," ACM Computing Surveys (CSUR), vol. 53, no. 6, pp. 1-26, 2020.

[4] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14372-14381.

[5] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20402-20411.

[6] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 8330-8339.

[7] J. Jiang et al., "Masked swin transformer unet for industrial anomaly detection," IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 2200-2209, 2022.

[8] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9592-9600.

[9] D. P. Kingma, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[10] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[11] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in International Conference on Pattern Recognition, 2021: Springer, pp. 475-489.

[12] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1088-1097.

[13] G. Hinton, "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531, 2015.

[14] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5008-5017.

[15] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4183-4192.

[16] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14902-14912.

[17] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 9737-9746.

[18] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9664-9674.

[19] X. Zhang, M. Xu, and X. Zhou, "RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16699-16708.

[20] T. D. Tien et al., "Revisiting reverse distillation for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 24511-24520.

[21] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," Pattern Recognition, vol. 112, p. 107706, 2021.

[22] M. Haselmann, D. P. Gruber, and P. Tabatabai, "Anomaly detection using deep learning based image completion," in 2018 17th IEEE international conference on machine learning and applications (ICMLA), 2018: IEEE, pp. 1237-1242.

[23] N.-C. Ristea et al., "Self-supervised predictive convolutional attentive block for anomaly detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 13576-13586.

[24] B. A. WANGZ and H. Sheikh, "Image qualityassessment: Fromerrorvisibilitytostructural similarity," IEEE TransactionsonImageProcessing, vol. 13, no. 4, p. 600G612, 2004.

[25] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv 2018," arXiv preprint arXiv:1807.02011, 2018.

[26] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences. arXiv 2020," arXiv preprint arXiv:2005.02357, 2005.

[27] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but differnet: Semi-supervised defect detection with normalizing flows," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 1907-1916.

[28] J. Song, K. Kong, Y.-I. Park, S.-G. Kim, and S.-J. Kang, "AnoSeg: Anomaly segmentation network using self-supervised learning," arXiv preprint arXiv:2110.03396, 2021.

[29] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," arXiv preprint arXiv: 2103.04257, 2021.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, 2009: Ieee, pp. 248-255.