ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2026.0151

Implicit multi-teacher distillation for unsupervised industrial anomaly detection

Lanlan WU[1,2], Gaoming YANG[1,2,3]✉, Xiujun WANG[4], Ji ZHANG[5]✉, Zhi LIU[6]

[1]*State Key Laboratory of Digital Intelligent Technology for Unmanned Coal Mining, Anhui University of Science and Technology, Huainan 232001, China*

[2]*School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China*

[3]*The First Hospital of Anhui University of Science and Technology, Huainan 232001, China*

[4]*School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China*

[5]*School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba 4350, Australia*

[6]*Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan*

Abstract: Industrial anomaly detection is important for intelligent manufacturing since defects can degrade product reliability. Existing methods based on knowledge distillation typically use a single teacher, limiting feature diversity and robustness. Multi-teacher methods often require explicit feature alignment across different network architectures, which may introduce semantic conflicts and weaken anomaly responses. We propose implicit multi-teacher distillation (IMTD) for industrial anomaly detection. IMTD employs three complementary teacher networks, namely ResNet18, ResNet34, and Wide-ResNet50, to capture texture, structural, and semantic representations from different feature spaces. Each teacher is paired with an independent decoder trained only on normal samples, enabling implicit knowledge transfer without cross-architecture feature alignment. A false-negative-aware maximum response aggregation (FN-MRA) module is further introduced to preserve the strongest anomaly evidence during fusion. Experiments on MVTec anomaly detection (MVTec AD), visual anomaly benchmark (VisA), and metal part defect dataset (MPDD) show that IMTD consistently outperforms recent state-of-the-art methods in both image-level and pixel-level anomaly detection. The results demonstrate the effectiveness of combining complementary teachers, implicit reconstruction, and maximum-response aggregation under anomaly-scarce conditions.

Key words: Industrial anomaly detection; Knowledge distillation; Multi-teacher ensemble; Implicit reconstruction; False-negative-aware maximum response aggregation (FN-MRA)

1 Introduction

Industrial anomaly detection is essential for intelligent manufacturing, where undetected defects can compromise product quality and production reliability. Under anomaly-scarce conditions, only normal samples are available for training, making reliable anomaly detection particularly challenging.

✉ Gaoming YANG, gmyang@aust.edu.cn
✉ Ji ZHANG, Ji.Zhang@usq.edu.au
Lanlan WU, https://orcid.org/0009-0000-2142-2999
CLC number: TP391

Received: May 20, 2026; Revision accepted: Sept. 9, 2026;
Crosschecked: Oct. 9, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd. This is an open access article distributed under the terms of the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

This setting presents three challenges. First, industrial defects span texture, structure, and semantic levels, limiting the representational capability of a single backbone. Second, explicit feature alignment across different network architectures is often unstable since feature maps differ in spatial resolution, channel dimension, and semantic representation. Such alignment may distort the normal feature distribution. Third, anomaly responses are usually sparse and localized. Averaging-based fusion may weaken strong responses and increase false-negative rates.

Existing methods can be broadly divided into three categories. Methods based on reconstruction and feature modeling, including multi-scale feature fusion (Ma ZQ et al., 2025), gradient-driven anomaly synthesis (Chen et al., 2025), masked vision Transformer reconstruction (Luo W et al., 2025), and deep feature clustering (Wang et al., 2025), mainly rely on a single backbone and therefore provide limited representational

diversity. Distillation-based methods, including reverse distillation (Tien et al., 2023), efficient local-descriptor alignment (Batzner et al., 2024), hybrid feature-reconstruction distillation (Xu et al., 2025), adaptive discrepancy optimization (Li N et al., 2025), and multimodal teacher–student memory designs (Liu et al., 2025), are constrained by the representational capacity of a single teacher. Some methods also require explicit feature alignment, which may become unreliable under anomaly-scarce conditions. Multi-teacher and multi-branch approaches, including weighted teacher schemes (Ma Y et al., 2023), meta-learned fusion (Zhang HL et al., 2023), bidirectional reverse distillation (Gu et al., 2024), and dynamic loss-weighted aggregation (Hendria et al., 2025), improve feature diversity. However, most of them rely on averaging or soft fusion, which may suppress strong localized anomaly evidence.

To address these challenges, we propose implicit multi-teacher distillation (IMTD), a unified framework for unsupervised industrial anomaly detection. IMTD integrates three complementary teacher networks, teacher-specific implicit reconstruction, and the false-negative-aware maximum response aggregation (FN-MRA) module. This design enhances representation capacity, avoids explicit feature alignment, and preserves anomaly responses in feature fusion.

The main contributions of this work are summarized as follows:

1. Multi-manifold representation. We extend distillation-based anomaly detection from a single-manifold paradigm to a multi-manifold formulation. Three teachers with different architectures learn complementary normal feature manifolds covering texture, structural, and semantic information. This design alleviates the representational limitations of single-backbone methods.

2. Implicit knowledge transfer via teacher-specific reconstruction. We propose a teacher-specific reconstruction mechanism. Each teacher connects to an independent normalization-decoder branch trained only on normal samples. Knowledge transfer is achieved through reconstruction consistency instead of explicit cross-architecture feature alignment.

3. FN-MRA. We propose a pixel-wise maximum-response fusion strategy combined with Gaussian smoothing. The method preserves strong localized anomaly evidence and reduces anomaly-signal attenuation during ensemble fusion.

2 Related works

To address anomaly detection in a highly imbalanced setting, recent research has evolved from single-backbone modeling toward multi-source fusion and knowledge distillation. In this section, we review the literature from three perspectives.

2.1 Single-backbone anomaly detection methods

Vanilla industrial anomaly detection methods (Chen et al., 2025; Ma ZQ et al., 2025; Wang et al., 2025) model the normal data distribution using a single backbone. They improve anomaly detection through multi-scale feature fusion, synthetic anomaly generation, and feature-space clustering. More recently, vision Transformer (ViT)-based methods (Jeong et al., 2023; Luo W et al., 2025; Wahid et al., 2025) leverage global self-attention and masked reconstruction techniques. Meanwhile, diffusion-based methods (Wei et al., 2025; Zhang H et al., 2025) adopt denoising mechanisms to characterize the distribution of normal samples.

Despite these advances, all aforementioned methods depend on only one representation space. ViT-based models excel at capturing the global context but lose local texture details, while diffusion models suffer from heavy computation and over-smoothing defects. Hence, they cannot effectively handle anomalies covering low-, middle-, and high-level features.

2.2 Multi-source information fusion and anomaly modeling

To address the drawbacks stemming from single representation learning, recent research has focused on information fusion and multi-source anomaly modeling.

Multi-source fusion methods combine complementary information from different feature spaces, views, or modalities to improve anomaly detection. Representative approaches include interaction-aware feature aggregation (Ahmad and Rehman, 2025) and semantic discrepancy modeling (Zhong et al., 2026). In addition, multimodal methods jointly exploit visual, structural, and temporal information through spatio-temporal modeling or cross-modal knowledge transfer (Cheng et al., 2025; Luo JZ et al., 2025; Xu et al., 2026).

However, most existing methods rely on explicit feature alignment or shared embedding spaces, resulting in semantic inconsistency across complementary representations. Moreover, averaging and soft aggregation weaken anomaly responses and lower localization accuracy. These drawbacks inspire our implicit reconstruction and response-preserving fusion strategy.

2.3 Knowledge distillation for anomaly detection

Knowledge distillation has become a widely adopted paradigm for industrial anomaly detection under anomaly-scarce conditions. It detects anomalies by measuring discrepancies between teacher and student representations.

2.3.1 Single-teacher distillation

Single-teacher distillation uses one pre-trained network to supervise student models. Reverse distillation (Tien et al., 2023) boosts detection performance by reversing the teacher–student paradigm. Local descriptor alignment (Batzner et al., 2024) cuts computations while maintaining feature consistency. Hybrid distillation (Xu et al., 2025) combines feature alignment and reconstruction for improved robustness. Adaptive discrepancy optimization (Li N et al., 2025) dynamically adjusts distillation constraints based on data characteristics. Multimodal memory-driven teacher–student frameworks (Liu et al., 2025) further refine anomaly representations with auxiliary feature memory. Despite promising results, all these approaches suffer from the inherent bottleneck of a single teacher's representation capability.

2.3.2 Multi-teacher distillation

To improve representation capability, multi-teacher distillation employs multiple teacher networks with complementary

characteristics. Representative approaches include weighted teacher aggregation (Ma Y et al., 2023), meta-learning-based fusion (Zhang HL et al., 2023), bidirectional distillation (Gu et al., 2024), and dynamic teacher aggregation (Hendria et al., 2025).

However, these multi-teacher distillation methods face unique obstacles when fusing outputs from heterogeneous pre-trained backbones. Directly harmonizing mismatched feature statistics across different teacher networks requires extra learnable projections, which makes optimization unstable under anomaly-scarce training regimes.

In summary, existing methods still suffer from limited representation capability, unstable feature alignment, and anomaly signal attenuation during fusion. These limitations motivate the proposed IMTD framework, which addresses the three issues through complementary teacher networks, teacher-specific reconstruction, and response-preserving fusion.

3 Method

This section introduces the proposed IMTD framework. It consists of multi-teacher feature extraction, teacher-specific implicit reconstruction, and pixel-level anomaly response ensemble using the proposed FN-MRA.

3.1 Overall framework

The overall architecture of IMTD is shown in Fig. 1. It consists of three modules: multi-teacher feature extraction, teacher-specific implicit reconstruction, and pixel-level anomaly response ensemble. After data preprocessing, three complementary teacher backbones extract texture-, structure-, and semantic-level features. Each teacher is paired with an independent one-class bottleneck embedding (OCBE) layer and a symmetric decoder for feature reconstruction, avoiding explicit cross-architecture feature alignment. The OCBE module fuses multi-scale teacher features and compresses them into low-dimensional compact embeddings to filter out anomalous feature perturbations during distillation. Cosine similarity loss is used to supervise feature reconstruction. During inference, each branch generates a pixel-level anomaly score, and the proposed FN-MRA module combines these scores to preserve strong anomaly responses.

During training, each reconstruction branch learns only from normal samples, enabling it to model the corresponding normal feature distribution. During inference, anomalous inputs produce reconstruction discrepancies that serve as anomaly cues. The anomaly maps from different branches are fused after independent reconstruction, preserving complementary information captured by different teacher networks.

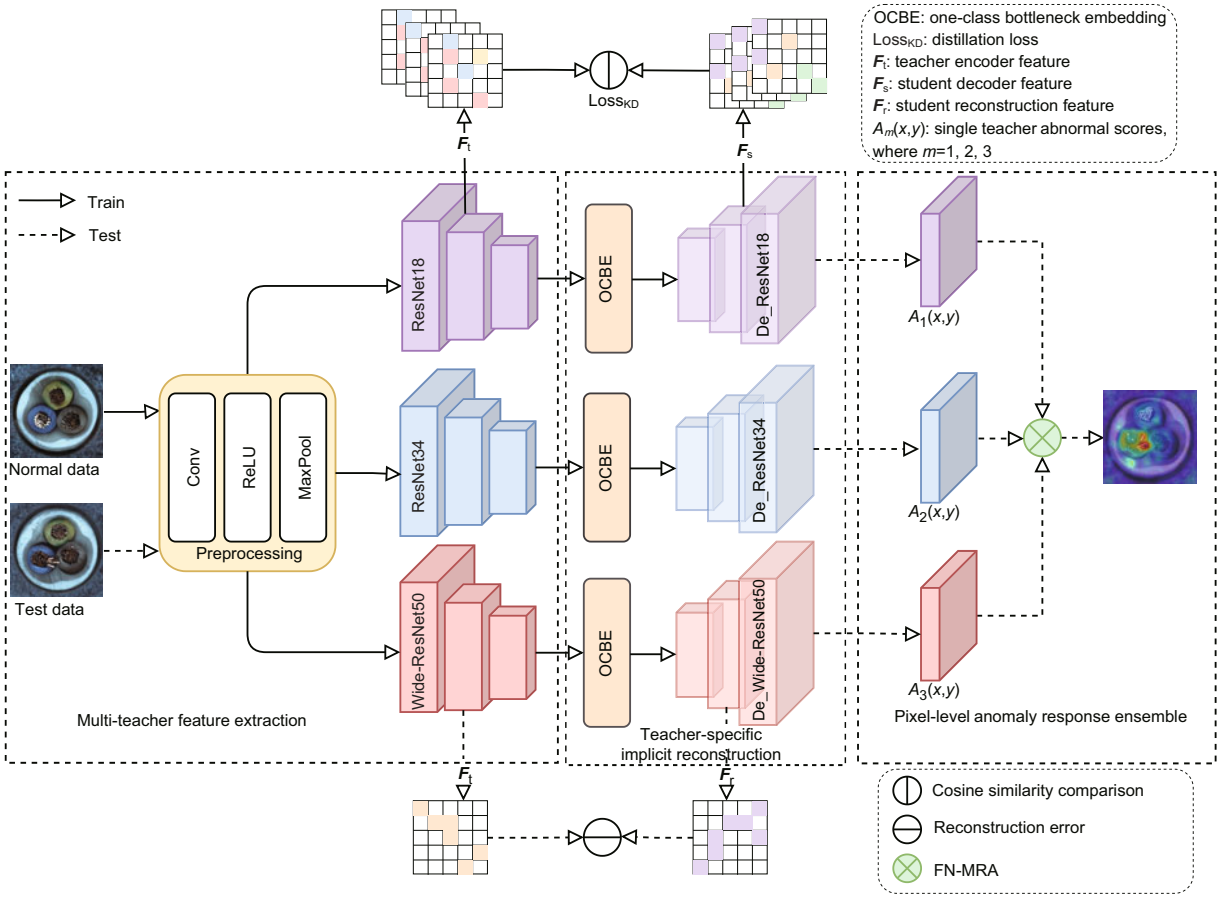


Fig. 1 Overview of the proposed IMTD framework. Three heterogeneous teacher networks (ResNet18, ResNet34, and Wide-ResNet50) extract complementary features and perform teacher-specific implicit reconstruction through independent decoder branches. Pixel-level anomaly maps generated by each branch are fused using FN-MRA to obtain the final anomaly localization result

3.2 Teacher–student architecture with representation learning

Industrial anomalies range from fine textures to structural and semantic defects, making it difficult for a single feature encoder to capture all anomaly patterns. As shown in Fig. 2, the input image is first processed by a shared preprocessing module and then fed into three complementary teacher networks: ResNet18, ResNet34, and Wide-ResNet50. ResNet18 emphasizes texture information, ResNet34 captures structural features, and Wide-ResNet50 extracts high-level semantic representations. Their complementary characteristics provide more comprehensive feature representations for industrial anomaly detection.

In the following formulation, $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ denotes the input image, $T_t(\cdot)$ represents the t^{th} teacher encoder, and $S_t(\cdot)$ denotes the corresponding decoder. The total number of teacher networks is denoted by M, while k indicates the feature scale index. The teacher feature map and reconstructed feature map are denoted by $\boldsymbol{f}_t^k$ and $\hat{\boldsymbol{f}}_t^k$, respectively. Here, C_t^k denotes the number of channels, and $H_t^k \times W_t^k$ represents the spatial resolution of the feature map. Furthermore, $\psi_t(\cdot)$ denotes the batch normalization operation, and $\mathcal{M}_t$ represents the normal feature manifold induced by teacher T_t.

3.2.1 Multi-teacher encoder model

Industrial anomalies may appear at different abstraction levels, including texture-level defects, structural deformations, and semantic inconsistencies. A single backbone usually emphasizes only one type of representation, thereby limiting anomaly observability. To address this limitation, IMTD employs three differentiated teacher encoders: ResNet18, ResNet34, and Wide-ResNet50.

Given an input image $\boldsymbol{x}$, the teacher feature map of the t^{th} teacher network is defined as

$$\boldsymbol{f}_t^k = T_t^k(\boldsymbol{x}), \quad t \in \{1,2,3\}, \tag{1}$$

where $T_t^k(\cdot)$ denotes the feature extraction operation at the k^{th} layer of teacher T_t, and we have

$$\boldsymbol{f}_t^k \in \mathbb{R}^{C_t^k \times H_t^k \times W_t^k}, \tag{2}$$

which represents the corresponding multi-scale feature map.

From a manifold-learning perspective, each teacher defines an independent normal feature manifold:

$$\mathcal{M}_t = \left\{ f_t^k \mid \boldsymbol{x} \sim p_{\text{normal}}(\boldsymbol{x}) \right\}, \tag{3}$$

where $p_{\text{normal}}(\boldsymbol{x})$ represents the normal sample distribution. The union of differentiated manifolds enlarges the observable representation space for anomaly detection:

$$\mathcal{M}_{\text{union}} = \bigcup_{t=1}^{M} \mathcal{M}_t. \tag{4}$$

Consequently, anomalies can be captured across complementary feature manifolds, improving robustness to different defect types. All teacher encoders remain frozen during training to preserve stable feature representations.

3.2.2 Teacher-specific implicit student decoder

Conventional multi-teacher distillation methods usually enforce explicit feature alignment between teachers and students. However, explicit alignment may distort feature manifolds when different teacher networks produce incompatible

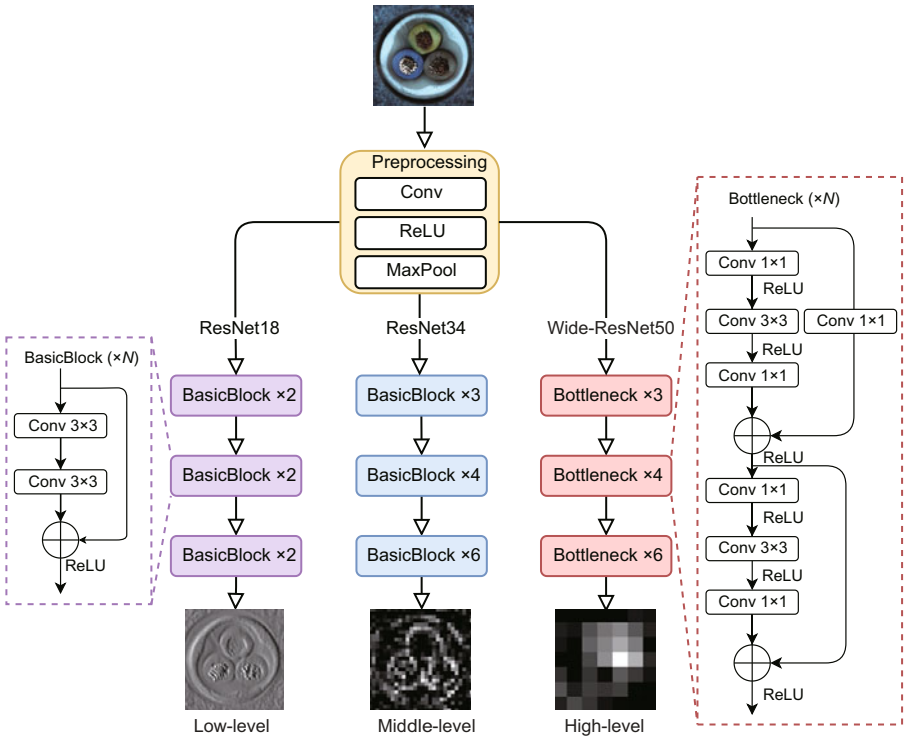


Fig. 2 Multi-teacher encoder architecture. ResNet18, ResNet34, and Wide-ResNet50 are used to extract complementary low-, middle-, and high-level features from preprocessed input images, respectively

representations. To address this issue, IMTD adopts a teacher-specific implicit reconstruction strategy. Each teacher is paired with a dedicated one-class bottleneck embedding layer and a decoder.

For teacher T_t, the extracted feature map is first normalized as

$$\boldsymbol{z}_t^k = \psi_t\left(\boldsymbol{f}_t^k\right), \tag{5}$$

where $\boldsymbol{z}_t^k$ denotes the normalized latent representation.

The corresponding decoder then reconstructs the normalized feature:

$$\hat{\boldsymbol{f}}_t^k = s_t^k\left(\boldsymbol{z}_t^k\right), \tag{6}$$

where $s_t^k(\cdot)$ denotes the decoder associated with teacher T_t.

Unlike explicit feature matching, the proposed formulation learns a manifold-preserving reconstruction mapping:

$$s_t : \mathcal{M}_t \to \mathcal{M}_t. \tag{7}$$

Since only normal samples are used for training, each decoder learns the normal feature manifold $\mathcal{M}_t$. During inference, anomalous features deviate from this manifold and produce large reconstruction discrepancies.

This design avoids explicit cross-teacher feature alignment, reducing semantic conflicts among complementary representations. Each decoder operates on its own manifold, preserving independent feature structures. Consequently, IMTD maintains complementary feature diversity throughout the distillation process.

3.3 False-negative-aware maximum response fusion

After teacher-specific reconstruction, anomaly maps are independently computed for each teacher branch. The key challenge then becomes how to effectively fuse differentiated anomaly responses without suppressing localized defect evidence.

For the t^{th} teacher, the anomaly score at spatial location (i,j) is defined using cosine distance:

$$A_t(i,j) = 1 - \frac{(\boldsymbol{f}_t(i,j))^{\mathrm{T}}\hat{\boldsymbol{f}}_t(i,j)}{\left\|\boldsymbol{f}_t(i,j)\right\|\left\|\hat{\boldsymbol{f}}_t(i,j)\right\|}, \tag{8}$$

where $A_t(i,j)$ denotes the anomaly score produced by teacher T_t, while $\boldsymbol{f}_t(i,j)$ and $\hat{\boldsymbol{f}}_t(i,j)$ denote the original and reconstructed feature vectors at spatial location (i,j), respectively. $\|\cdot\|$ denotes the L2 norm of the feature vector.

To preserve dominant anomaly evidence, IMTD adopts the FN-MRA strategy:

$$A(i,j) = \max_{t\in\{1,2,\ldots,M\}} A_t(i,j). \tag{9}$$

Unlike averaging-based aggregation, FN-MRA preserves the strongest anomaly activation across distinct manifolds.

To suppress isolated noise and improve spatial continuity, Gaussian smoothing is further applied to the fused anomaly map:

$$A_{\text{smooth}}(i,j) = (A * G_\sigma)(i,j), \tag{10}$$

where $*$ denotes the convolution operation, and G_σ represents the Gaussian kernel, which is defined as

$$G_\sigma(u,v) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{u^2+v^2}{2\sigma^2}\right). \tag{11}$$

In our implementation, a 5×5 Gaussian kernel with a standard deviation of $\sigma = 1.0$ is adopted. Gaussian filtering enhances regional continuity while preserving dominant anomaly peaks generated by FN-MRA. The resulting anomaly map therefore achieves improved localization consistency and reduced fragmented responses.

3.4 Training objective and optimization strategy

The proposed IMTD framework replaces explicit teacher–student feature alignment with implicit feature reconstruction. Since only normal samples are used for training, each reconstruction branch learns a compact normal feature manifold.

For the t^{th} teacher network, the extracted feature map $\boldsymbol{f}_t^k$ is first normalized through the one-class bottleneck embedding layer $\psi_t(\cdot)$ and then reconstructed by the corresponding decoder $s_t(\cdot)$. The reconstructed feature is denoted as $\hat{\boldsymbol{f}}_t^k$.

To measure the discrepancy between the teacher and reconstructed feature, IMTD adopts cosine similarity loss:

$$\mathcal{L}_{\cos}^{t,k} = 1 - \frac{\left(\boldsymbol{f}_t^k\right)^{\mathrm{T}}\hat{\boldsymbol{f}}_t^k}{\left\|\boldsymbol{f}_t^k\right\|\left\|\hat{\boldsymbol{f}}_t^k\right\|}. \tag{12}$$

The reconstruction loss for the t^{th} teacher is computed by aggregating all feature scales:

$$\mathcal{L}_t = \sum_{k=1}^{K}\mathcal{L}_{\cos}^{t,k}, \tag{13}$$

where K denotes the number of feature scales.

Finally, the overall optimization objective of IMTD is defined as the average loss across all teacher branches:

$$\mathcal{L}_{\text{total}} = \frac{1}{M}\sum_{t=1}^{M}\mathcal{L}_t. \tag{14}$$

The training and inference procedures are summarized in Algorithms 1 and 2, respectively.

3.5 Computational complexity and scalability analysis

The computational cost of IMTD consists of forward passes through three frozen teachers, lightweight decoder reconstruction, and pixel-level max fusion. Let C_{T} denote the

Algorithm 1 Training procedure of IMTD

Input: normal training dataset $\mathcal{D}$; frozen teacher encoders $\{T_1, T_2, \ldots, T_M\}$; teacher-specific decoders $\{S_1, S_2, \ldots, S_M\}$.
Output: trained decoders.

for each training epoch **do**
  for each normal sample $\boldsymbol{x} \in \mathcal{D}$ **do**
    for each teacher T_t **do**
      Extract features: $\boldsymbol{f}_t^k = T_t^k(\boldsymbol{x})$
      Normalize features: $\boldsymbol{z}_t^k = \psi_t(\boldsymbol{f}_t^k)$
      Reconstruct features: $\hat{\boldsymbol{f}}_t^k = s_t^k(\boldsymbol{z}_t^k)$
      Compute cosine reconstruction loss: $\mathcal{L}_t = \sum_{k=1}^{K}\mathcal{L}_{\cos}^{t,k}$
    end for
    Compute total loss: $\mathcal{L}_{\text{total}} = \frac{1}{M}\sum_{t=1}^{M}\mathcal{L}_t$
    Update decoder parameters using adaptive moment estimation (Adam)
  end for
end for

Algorithm 2 Inference procedure of IMTD

Input: test image $\boldsymbol{x}$; teacher encoders $\{T_1, T_2, \ldots, T_M\}$; trained decoders $\{S_1, S_2, \ldots, S_M\}$.

Output: final anomaly score A_{smooth}.

1: **for** each teacher T_t **do**

2: Extract features: $\boldsymbol{f}_t^k = T_t^k(\boldsymbol{x})$

3: Normalize features: $\boldsymbol{z}_t^k = \psi_t(\boldsymbol{f}_t^k)$

4: Reconstruct features: $\hat{\boldsymbol{f}}_t^k = s_t^k(\boldsymbol{z}_t^k)$

5: Compute anomaly map:

$$A_t(i,j) = 1 - \frac{(\boldsymbol{f}_t(i,j))^{\mathrm{T}} \hat{\boldsymbol{f}}_t(i,j)}{\|\boldsymbol{f}_t(i,j)\| \|\hat{\boldsymbol{f}}_t(i,j)\|}$$

6: **end for**

7: Perform FN-MRA fusion:

$$A(i,j) = \max_{t=1,2,\ldots,M} A_t(i,j)$$

8: Apply Gaussian smoothing:

$$A_{\text{smooth}}(i,j) = (A * G_\sigma)(i,j)$$

9: Return A_{smooth}

cost of a single teacher forward pass. The total inference complexity is

$$\mathcal{O}(3C_{\mathrm{T}} + 3C_{\mathrm{D}}), \tag{15}$$

where C_{D} denotes the decoder cost.

Since teacher networks are frozen, inference remains efficient. The decoder networks are shallow and introduce only a small number of additional parameters. Memory usage increases linearly with the number of teacher branches. Since no feature concatenation is performed, intermediate tensor expansion is avoided, and the pixel-wise fusion incurs negligible computational overhead.

4 Experimental results

This section presents the experimental setup and evaluation of IMTD. It introduces the datasets, evaluation metrics, and implementation details. Quantitative comparisons, ablation studies, and qualitative results are then presented.

4.1 Experimental setup

4.1.1 Datasets

We evaluate IMTD on three widely adopted industrial anomaly detection benchmarks.

The MVTec anomaly detection (MVTec AD) dataset (Bergmann et al., 2019) contains 15 categories of industrial objects and textures. It provides only normal samples for training and includes diverse real-world defects in the test set. MVTec AD is the most widely used benchmark for industrial anomaly detection.

The visual anomaly benchmark (VisA) dataset (Zou et al., 2022) contains 12 categories and 10 821 high-resolution images. It includes cluttered backgrounds, small defects, and subtle appearance variations, making anomaly detection more difficult.

The metal part defect dataset (MPDD) (Jezek et al., 2021) contains painted metal parts collected from real industrial environments. It is widely used to evaluate anomaly detection under complex inspection conditions.

4.1.2 Evaluation metrics

We adopt three widely used evaluation metrics: image-level AUROC (I-AUROC), pixel-level AUROC (P-AUROC), and per-region overlap (P-AUPRO). I-AUROC measures the ability to distinguish normal and abnormal images. P-AUROC evaluates pixel-level localization accuracy. P-AUPRO measures region-level overlap while penalizing excessive false positives. Together, these metrics assess global discrimination, fine-grained localization, and spatial consistency. This multi-perspective evaluation aligns with our theoretical claim that IMTD models complementary normal manifolds across scales.

4.1.3 Implementation details

The proposed IMTD framework was implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU. ImageNet-pretrained teacher encoders were kept frozen, while only the one-class bottleneck embedding layers and decoder modules were optimized. The Adam optimizer was used with an initial learning rate of 0.005. The batch size was 16, and training was performed for 200 epochs using cosine similarity loss. All images were resized to 256×256. Random horizontal flipping and random rotation were applied to normal training samples. A fixed random seed of 111 was used throughout the experiments. For fair comparison, all retrainable methods were trained under identical settings. The results for non-retrainable methods were taken from the original papers. The same preprocessing and evaluation protocol was adopted for all compared methods.

4.2 Comparison experiments

4.2.1 Results on MVTec AD

Table 1 reports the image-level and pixel-level results on MVTec AD. IMTD achieves 98.81% I-AUROC, 97.98% P-AUROC, and 94.47% P-AUPRO, outperforming the compared methods on average. Several methods achieve near-perfect performance on simple categories, such as Carpet, Grid, and Leather. However, their performance drops in more challenging categories, such as Cable, Capsule, and Transistor. In contrast, IMTD maintains stable performance across both texture and object categories. This improvement indicates that complementary teacher networks provide more robust feature representations.

Compared with existing anomaly detection methods based on reverse distillation (Deng and Li, 2022), discrepancy modeling (Cai et al., 2024), and contrastive language–image pretraining (CLIP)-based vision-language models (Cao et al., 2024), IMTD models complementary normal feature manifolds through multiple teacher networks. The reconstruction residuals (Zhang ZL et al., 2023) amplify subtle feature deviations and produce clearer anomaly responses. In addition, FN-MRA preserves strong anomaly cues during fusion and reduces the attenuation caused by averaging-based aggregation. The improved localization performance benefits from complementary teacher representations and the FN-MRA module. The former enhances feature representation, while the latter preserves strong anomaly responses during fusion.

Table 1 Comparison of I-AUROC/P-AUROC/P-AUPRO across different methods on MVTec AD

Class	I-AUROC (%)/P-AUROC (%)/P-AUPRO (%)						
	GLASS (Chen et al., 2025)	AdaCLIP (Cao et al., 2024)	RD4AD (Deng and Li, 2022)	MuSc (Li XR et al., 2024)	MMR (Zhang ZL et al., 2023)	DAF (Cai et al., 2024)	IMTD
Carpet	97.99/99.16/95.30	**100.00**/**99.72**/–	99.27/96.63/97.17	99.87/99.45/**97.58**	99.61/98.75/96.63	94.10/98.66/94.73	**100.00**/99.28/97.32
Grid	94.06/90.33/72.62	95.15/98.87/–	**100.00**/98.50/97.13	98.74/98.16/93.90	**100.00**/99.00/96.54	**100.00**/98.45/93.14	99.40/**99.51**/**97.73**
Leather	**100.00**/99.27/99.28	96.73/91.24/–	**100.00**/98.91/99.00	**100.00**/**99.71**/98.74	98.50/97.21/98.60	**100.00**/99.68/98.70	**100.00**/99.42/**99.31**
Tile	**100.00**/96.82/**97.89**	**100.00**/**99.50**/–	99.00/93.20/91.00	**100.00**/97.90/94.64	98.27/95.59/90.21	**100.00**/99.33/96.65	99.13/96.02/91.40
Wood	99.12/87.26/**97.48**	98.12/**97.53**/–	98.91/95.05/90.92	98.50/97.23/94.50	99.00/94.79/88.93	99.65/97.12/93.45	**99.76**/95.46/92.25
Bottle	97.85/94.66/91.92	**100.00**/98.30/–	**100.00**/98.13/96.50	99.92/98.48/96.10	98.70/96.27/96.00	99.29/**98.89**/94.87	**100.00**/98.70/**96.71**
Cable	83.66/86.07/75.54	97.06/91.75/–	98.30/95.58/92.54	**98.99**/95.76/89.62	97.85/95.43/87.27	94.38/96.98/88.11	98.60/**98.10**/**93.30**
Capsule	87.48/91.20/81.51	98.41/93.53/–	96.86/96.30/96.50	96.64/**98.95**/95.57	96.90/98.02/94.50	90.87/96.56/83.36	**98.71**/98.83/**96.70**
Hazelnut	98.43/94.19/93.39	97.26/96.44/–	**100.00**/97.00/93.52	99.61/**99.37**/92.23	**100.00**/98.50/91.20	**100.00**/98.92/96.34	**100.00**/99.00/**96.39**
Metal nut	**100.00**/99.37/94.54	98.34/98.91/–	**100.00**/93.27/91.81	96.68/86.12/89.32	99.10/95.92/88.64	99.95/**99.64**/**96.59**	**100.00**/97.74/91.50
Pill	**99.29**/**99.35**/97.55	98.96/98.22/–	96.90/94.42/96.73	96.39/97.47/**98.01**	98.70/98.43/96.18	98.31/98.22/90.46	97.10/98.32/97.90
Screw	**99.97**/95.14/97.39	92.10/95.07/–	97.32/99.40/98.44	82.25/98.76/94.40	92.34/99.46/97.73	92.95/99.06/95.30	98.27/**99.60**/**98.72**
Toothbrush	99.54/98.54/**95.54**	98.61/97.66/–	96.44/96.53/94.82	**100.00**/**99.53**/95.47	98.05/98.34/93.00	**100.00**/98.30/88.40	98.92/99.00/93.81
Transistor	**99.74**/**97.63**/**95.02**	97.58/95.87/–	96.50/87.91/76.47	99.50/91.37/77.18	95.00/90.21/79.32	93.92/89.75/75.35	97.50/92.26/77.63
Zipper	99.71/**99.64**/**99.01**	**99.95**/97.18/–	92.68/95.14/94.40	99.21/98.40/94.84	97.67/97.99/95.00	99.89/99.16/96.46	94.75/98.44/96.37
Average	97.12/95.24/92.27	97.88/96.65/–	98.15/95.73/93.80	97.75/97.11/93.47	97.98/96.93/92.65	97.55/97.91/92.13	**98.81**/**97.98**/**94.47**

The best results are highlighted in bold, and the second-best results are underlined. The "–" symbol indicates that the corresponding metric was not reported in the original paper and could not be reproduced because the source code was unavailable. The same convention applies to the following tables

4.2.2 Results on VisA

Table 2 summarizes the results on VisA. IMTD achieves 95.83% I-AUROC, 98.72% P-AUROC, and 92.87% P-AUPRO. It outperforms the compared semantic-matching methods (Jeong et al., 2023; Sun et al., 2025) and reconstruction-based methods (Cai et al., 2024; Chen et al., 2025) on average.

VisA contains cluttered backgrounds and subtle defects. Single-view models often generate overly smooth anomaly maps. IMTD benefits from complementary receptive fields and implicit manifold modeling across teachers. The ensemble of teachers captures fine textures and global semantics simultaneously. This behavior reflects the complementary feature representations learned by the teacher networks, which help capture anomalies across different feature levels.

Table 2 Comparison of different models on VisA

Method	I-AUROC (%)	P-AUROC (%)	P-AUPRO (%)
GLASS (Chen et al., 2025)	92.33	94.77	89.76
AdaCLIP (Cao et al., 2024)	80.79	93.18	–
DAF (Cai et al., 2024)	94.74	98.57	90.38
WinCLIP (Jeong et al., 2023)	74.13	77.24	71.99
AnomalyAny (Sun et al., 2025)	91.50	98.10	92.60
IMTD	**95.83**	**98.72**	**92.87**

4.2.3 Results on MPDD

Table 3 reports the results on MPDD. IMTD achieves the best P-AUROC (98.72%) and competitive I-AUROC and P-AUPRO scores. Reconstruction-based baselines (Cai et al., 2024; Chen et al., 2025) show painted metal surfaces with heterogeneous structures. These methods struggle with subtle texture shifts and macroscopic defects.

IMTD benefits from independent decoder design. Each teacher reconstructs stable normal features, enabling each student branch to learn a compact representation of normal patterns. Deviations from this manifold are amplified during inference. The results demonstrate improved robustness under small-sample and class-imbalanced settings.

Table 3 Comparison of different models on MPDD

Method	I-AUROC (%)	P-AUROC (%)	P-AUPRO (%)
RD4AD (Deng and Li, 2022)	92.85	97.54	95.24
AdaCLIP (Cao et al., 2024)	82.71	95.86	–
DAF (Cai et al., 2024)	88.47	97.64	90.65
MMR (Zhang ZL et al., 2023)	94.16	97.67	94.05
DeSTSeg (Zhang X et al., 2023)	93.66	94.47	86.84
IMTD	**94.57**	**98.72**	**95.37**

4.3 Ablation studies

4.3.1 Effect of the teacher configuration

Table 4 presents the ablation results for different teacher combinations. Single-teacher configurations achieve reasonable performance, while dual-teacher settings further improve the results by combining complementary features. The three-teacher configuration achieves the best performance, reaching 98.81% I-AUROC and 94.47% P-AUPRO. These results demonstrate that three complementary teacher networks provide more comprehensive feature representations than one or two teachers.

4.3.2 Effect of the loss function

Compared with cosine similarity loss, the combined objectives of CosSim+MSE and CosSim+SSIM show slower

Table 4 Teacher network ablation results

Teacher network	I-AUROC (%)	P-AUROC (%)	P-AUPRO (%)
ResNet18	97.02	96.91	92.57
ResNet34	97.46	97.09	92.61
Wide-ResNet50	97.32	96.60	92.91
ResNet18+ResNet34	98.09	97.33	93.14
ResNet18+Wide-ResNet50	98.12	96.83	93.07
ResNet34+Wide-ResNet50	98.27	97.15	93.27
IMTD	**98.81**	**97.98**	**94.47**

convergence and less stable optimization. As shown in Fig. 3, I-AUROC drops to 77.29% and 63.90%, respectively. This degradation is likely due to the different optimization objectives of the loss functions. Cosine similarity aligns semantic feature directions, whereas MSE and SSIM focus on reconstruction fidelity and local structural similarity. Their combination may reduce the quality of the learned feature representations, resulting in inferior anomaly detection performance.

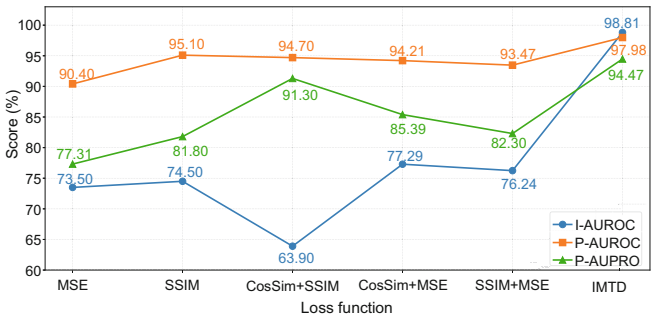


Fig. 3 Performance comparison under different reconstruction objectives

4.3.3 Effect of the fusion strategy

As shown in Table 5, multi-teacher fusion consistently outperforms the single-teacher setting, demonstrating the benefit of complementary teacher representations. Among all fusion strategies, FN-MRA achieves the best performance, indicating that preserving the strongest anomaly responses is more effective than average or weighted fusion.

Table 5 Ablation outcomes by the fusion method

Fusion method	I-AUROC (%)	P-AUROC (%)	P-AUPRO (%)
Single-teacher (no fusion)	97.27	96.87	92.70
Average fusion	98.01	97.43	93.55
Weighted fusion	<u>98.13</u>	<u>97.51</u>	<u>94.06</u>
FN-MRA	**98.81**	**97.98**	**94.47**

4.4 Efficiency and scalability analysis

Table 6 presents an efficiency and scalability analysis of different teacher configurations. I-AUROC is improved as additional teacher branches are introduced. Among the single-teacher models, ResNet34 achieves the best overall performance. Dual-teacher configurations further improve detection accuracy. The proposed IMTD with three teacher branches achieves the highest I-AUROC of 98.81% while maintaining real-time performance at 63.65 frames/s (FPS) with an inference time of 15.71 ms. Although the computational cost increases moderately, the overall deployment overhead remains practical. These results indicate that the proposed IMTD provides an effective trade-off between detection accuracy and computational efficiency.

4.5 Visualization

Fig. 4 illustrates the anomaly responses of different teacher networks. Different teachers exhibit distinct responses to different anomaly regions, indicating their complementary anomaly perception capabilities. FN-MRA preserves the strongest responses from each branch, resulting in more complete anomaly localization.

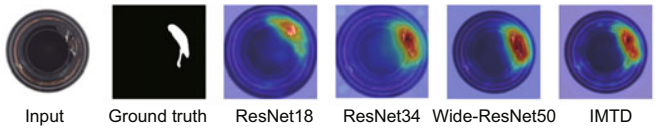


Fig. 4 Visualization of complementary anomaly responses from different teacher networks

Fig. 5 presents qualitative anomaly localization results on MVTec AD. The proposed IMTD produces heatmaps that closely match the ground-truth regions and accurately highlights small defects. Compared with other methods, it produces more complete anomaly maps while reducing false responses in normal regions.

4.6 Failure case analysis

As shown in Fig. 6, IMTD may still produce false-positive responses in repetitive texture regions and incomplete localization for subtle low-contrast defects. These examples suggest that repetitive textures and weak defect contrast remain challenging. Future work will focus on improving fine-grained texture representation and boundary refinement to further enhance localization accuracy.

5 Conclusions

This paper proposes IMTD for industrial anomaly detection under anomaly-scarce conditions. IMTD employs three complementary teacher networks to learn texture-, structure-, and semantic-level representations from multiple feature manifolds. Each teacher is paired with an independent

Table 6 Efficiency and scalability analysis of different teacher configurations

Configuration	Number of parameters ($\times 10^6$)	GFLOPs	GPU memory (MB)	Inference time (ms)	Frame rate (FPS)	I-AUROC (%)
ResNet18	27.60	5.49	125.81	4.74	210.91	97.02
ResNet34	44.49	10.34	192.09	5.33	187.71	97.46
Wide-ResNet50	161.08	38.94	673.23	9.74	102.64	97.32
ResNet18+ResNet34	72.09	15.83	301.28	8.39	119.13	98.09
ResNet18+Wide-ResNet50	188.68	44.43	783.67	13.10	76.34	98.12
ResNet34+Wide-ResNet50	205.57	49.27	849.70	14.31	69.87	<u>98.27</u>
IMTD	233.17	54.76	958.01	15.71	63.65	**98.81**

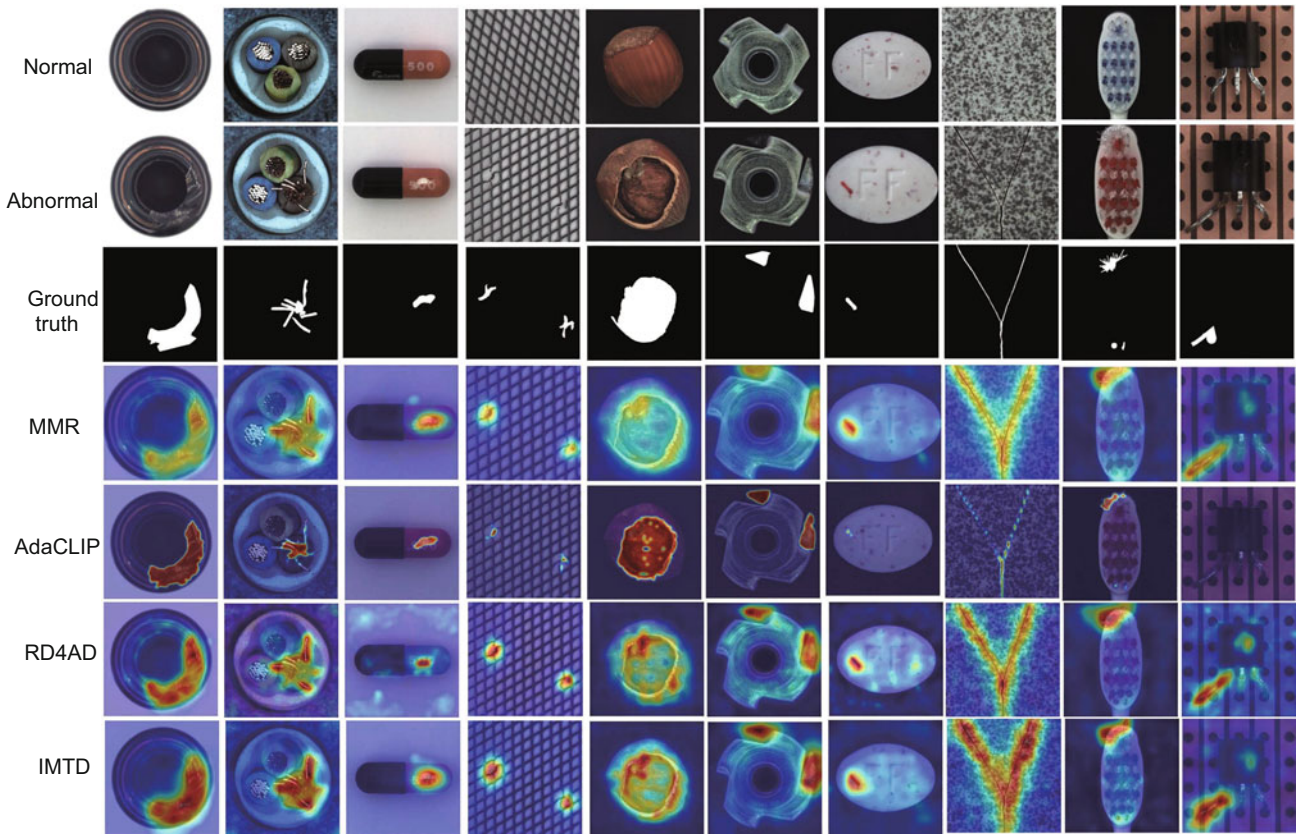


Fig. 5 Qualitative comparison of anomaly localization results on MVTec AD. The figure presents input images, ground-truth masks, and anomaly maps generated by different methods

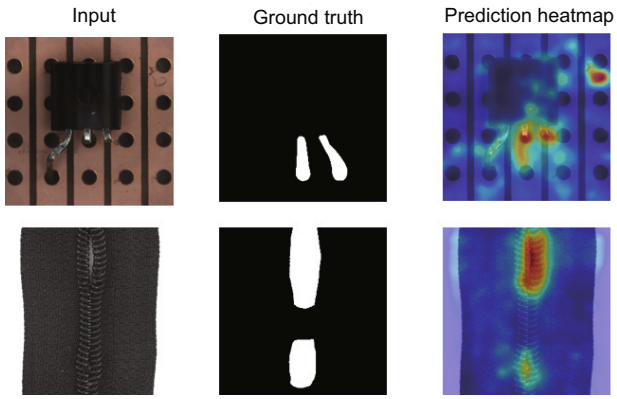


Fig. 6 Failure cases of the proposed IMTD

reconstruction decoder to model normal feature distributions, thereby avoiding explicit cross-architecture feature alignment. The FN-MRA strategy further preserves the strongest anomaly evidence during fusion. Extensive experiments on MVTec AD, VisA, and MPDD demonstrate the effectiveness of IMTD in both image-level and pixel-level anomaly detection. Qualitative results, ablation studies, and efficiency analysis further support its effectiveness and practicality. Future work will extend IMTD to multimodal industrial inspection scenarios and investigate lightweight distillation and model compression for practical deployment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 52374155), the Anhui Provincial Natural Science Foundation (No. 2308085MF218), the Natural Science Research Project of Colleges and Universities in Anhui Province (No. 2022AH040113), and the Medical Special Cultivation Project of Anhui University of Science and Technology (No. YZ2023H2B007).

Author contributions

Lanlan WU designed the research, processed the data, and drafted the paper. Zhi LIU helped organize the paper. Xiujun WANG, Ji ZHANG, and Gaoming YANG revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The code is available at https://github.com/wu-lanlan/IMTD. The other data that support the findings of this study are available from the corresponding authors upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used DeepSeek to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Ahmad M, Rehman A, 2025. Multi-source information fusion for anomaly detection in smart grids using federated learning. *Chin*

J Inform Fus, 2(2):157-170. https://doi.org/10.62762/CJIF.2025.220738

Batzner K, Heckler L, König R, 2024. EfficientAD: accurate visual anomaly detection at millisecond-level latencies. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.127-137. https://doi.org/10.1109/WACV57701.2024.00020

Bergmann P, Fauser M, Sattlegger D, et al., 2019. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9584-9592. https://doi.org/10.1109/CVPR.2019.00982

Cai YX, Liang DK, Luo DL, et al., 2024. A discrepancy aware framework for robust anomaly detection. *IEEE Trans Ind Inform*, 20(3):3986-3995. https://doi.org/10.1109/TII.2023.3318302

Cao YK, Zhang JN, Frittoli L, et al., 2024. AdaCLIP: adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. Proc 18th European Conf on Computer Vision, p.55-72. https://doi.org/10.1007/978-3-031-72761-0_4

Chen QY, Luo HY, Lv CK, et al., 2025. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. Proc European Conf on Computer Vision, p.37-54. https://doi.org/10.1007/978-3-031-72855-6_3

Cheng H, Luo JX, Zhang XY, 2025. Multimodal industrial anomaly detection via UNI-modal and cross-modal fusion. *IEEE Trans Ind Inform*, 21(6):5000-5010. https://doi.org/10.1109/TII.2025.3552723

Deng HQ, Li XY, 2022. Anomaly detection via reverse distillation from one-class embedding. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9727-9736. https://doi.org/10.1109/CVPR52688.2022.00951

Gu ZH, Zhang JN, Liu L, et al., 2024. Rethinking reverse distillation for multi-modal anomaly detection. Proc 38th AAAI Conf on Artificial Intelligence, 38(8):8445-8453. https://doi.org/10.1609/AAAI.v38i8.28687

Hendria WF, Kim H, Seo D, 2025. Multi-model anomaly detection for industrial inspection with dynamic loss weighting and soft-hard features loss. *Neur Comput Appl*, 37(21):17031-17054. https://doi.org/10.1007/s00521-025-11367-3

Jeong J, Zou Y, Kim T, et al., 2023. WinCLIP: zero-/few-shot anomaly classification and segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19606-19616. https://doi.org/10.1109/CVPR52729.2023.01878

Jezek S, Jonak M, Burget R, et al., 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. Proc 13th Int Congress on Ultra Modern Telecommunications and Control Systems and Workshops, p.66-71. https://doi.org/10.1109/ICUMT54235.2021.9631567

Li N, Liu AJ, Zhu ZW, et al., 2025. Knowledge distillation-based anomaly detection via adaptive discrepancy optimization. *IEEE Trans Ind Inform*, 21(9):7176-7187. https://doi.org/10.1109/TII.2025.3574540

Li XR, Huang ZM, Xue F, et al., 2024. MuSc: zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. Proc Int Conf on Learning Representations.

Liu Y, Zhang CS, Dong XJ, et al., 2025. A multimodal industrial anomaly detection method based on mask training and teacher–student joint memory. *Eng Appl Artif Intell*, 161:112299. https://doi.org/10.1016/j.engappai.2025.112299

Luo JZ, Cai Y, Cai J, et al., 2025. STMBAD: spatio-temporal multimodal behavior anomaly detector for industrial control systems. *IEEE Trans Ind Inform*, 21(5):3746-3755. https://doi.org/10.1109/TII.2025.3528559

Luo W, Yao HM, Yu WY, et al., 2025. AMI-Net: adaptive mask inpainting network for industrial anomaly detection and localization. *IEEE Trans Autom Sci Eng*, 22:1591-1605. https://doi.org/10.1109/TASE.2024.3368142

Ma Y, Jiang X, Guan N, et al., 2023. Anomaly detection based on multi-teacher knowledge distillation. *J Syst Archit*, 138:102861. https://doi.org/10.1016/j.sysarc.2023.102861

Ma ZQ, Li JX, Wong WK, 2025. Patch distance based auto-encoder for industrial anomaly detection. *Expert Syst Appl*, 270:126537. https://doi.org/10.1016/j.eswa.2025.126537

Sun H, Cao YK, Dong H, et al., 2025. Unseen visual anomaly generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.25508-25517. https://doi.org/10.1109/CVPR52734.2025.02375

Tien TD, Nguyen AT, Tran NH, et al., 2023. Revisiting reverse distillation for anomaly detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.24511-24520. https://doi.org/10.1109/CVPR52729.2023.02348

Wahid JA, Ayoub M, Xu ML, et al., 2025. NN2ViT: neural networks and vision transformers based approach for visual anomaly detection in industrial images. *Neurocomputing*, 615:128845. https://doi.org/10.1016/j.neucom.2024.128845

Wang RX, Li Z, Zheng L, et al., 2025. Deep feature clustering for multi-class industrial image anomaly detection. *Knowl-Based Syst*, 311:113134. https://doi.org/10.1016/j.knosys.2025.113134

Wei CY, Han H, Xia Y, et al., 2025. TDAD: self-supervised industrial anomaly detection with a two-stage diffusion model. *Comput Ind*, 164:104192. https://doi.org/10.1016/j.compind.2024.104192

Xu MH, Zhu CP, Feng G, et al., 2025. Multitask hybrid knowledge distillation for unsupervised anomaly detection. *IEEE Trans Ind Inform*, 21(7):5666-5676. https://doi.org/10.1109/TII.2025.3556083

Xu MH, Qu LL, Nie ZH, et al., 2026. Towards reliable medical unsupervised anomaly detection: a benchmark and dataset for PET/CT with cross-modality knowledge distillation network. *Inform Fus*, 133:104341. https://doi.org/10.1016/j.inffus.2026.104341

Zhang H, Wang Z, Zeng D, et al., 2025. DiffusionAD: norm-guided one-step denoising diffusion for anomaly detection. *IEEE Trans Patt Anal Mach Intell*, 47(8):7140-7152. https://doi.org/10.1109/TPAMI.2025.3570494

Zhang HL, Chen DF, Wang C, 2023. Adaptive multi-teacher knowledge distillation with meta-learning. Proc IEEE Int Conf on Multimedia and Expo, p.1943-1948. https://doi.org/10.1109/ICME55011.2023.00333

Zhang X, Li SY, Li X, et al., 2023. DeSTSeg: segmentation guided denoising student-teacher for anomaly detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3914-3923. https://doi.org/10.1109/CVPR52729.2023.00381

Zhang ZL, Zhao ZB, Zhang XW, et al., 2023. Industrial anomaly detection with domain shift: a real-world dataset and masked multi-scale reconstruction. *Comput Ind*, 151:103990. https://doi.org/10.1016/j.compind.2023.103990

Zhong YZ, Guo B, Sun ZM, et al., 2026. MSDCIR-AD: unsupervised anomaly detection via multi-criteria semantic distances and constrained image reconstruction. *Comput Ind*, 178:104467. https://doi.org/10.1016/j.compind.2026.104467

Zou Y, Jeong J, Pemula L, et al., 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. Proc 17th European Conf on Computer Vision, p.392-408. https://doi.org/10.1007/978-3-031-20056-4_23