Pattern Recognition 178 (2026) 113261

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Few-shot medical anomaly detection through centroid consultation back and test-time self-calibration

Zihan Nie [a,b], Muhao Xu [a,b], Yuan Cui [a,b], Hua Wei [a,b], Wei Yi [a,b], Sijie Niu [c], Yi Wan [a,b], Xunbin Wei [d,e,f], Weiye Song [a,b,*]

[a] School of Mechanical Engineering, Shandong University, No. 27 Shandong University South Road, Jinan, 250061, Shandong, China
[b] Key Laboratory of High Efficiency and Clean Mechanical Manufacture, Shandong University Ministry of Education, Qianfoshan Campus, Shandong University, 17923 Jingshi Road, Jinan, 250061, Shandong, China
[c] Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, No. 336, West Road of Nanxinzhuang, Jinan, 250022, Shandong, China
[d] Institute of Medical Technology and Cancer Hospital, Peking University, No. 38, Xueyuan Road, Haidian District, Beijing, 100191, Beijing, China
[e] Institute of Advanced Clinical Medicine and Biomedical Engineering Department, Peking University, No. 38, Xueyuan Road, Haidian District, Beijing, 100191, Beijing, China
[f] Peking University International Cancer Institute, No. 38, Xueyuan Road, Haidian District, Beijing, 100191, Beijing, China

ARTICLE INFO

Keywords:
Anomaly detection
Medical image analysis
Few-shot learning

ABSTRACT

Accurate detection of anomalies in medical images is of paramount importance for the early diagnosis and effective treatment of diseases. Nevertheless, the task is often impeded by the scarcity of labeled data, especially in specialized medical domains, which necessitates the development of few-shot learning methods that can generalize well from limited examples. Despite the strong representation capacity of recent pretrained models, existing few-shot anomaly detection methods still struggle in medical scenarios, as patch-level representations are often dominated by stable anatomical structures or repetitive textures rather than subtle pathological deviations. As a result, anomaly-relevant cues are suppressed at the representation level, giving rise to background-driven false positives even before anomaly scoring is applied. Moreover, the anomaly classification scores may be disproportionately influenced by localized prominent noise or artifacts, rather than accurately reflecting the true extent of abnormalities across the entire image, resulting in unreliable detection results.

To address these challenges, we propose a novel few-shot medical image anomaly detection framework that incorporates two innovative Plug-and-Play modules: a Centroid Consultation Back (CCB) module, which refines patch-level representations by introducing intermediate semantic anchors that provide global contextual feedback, suppressing background-dominated responses while enhancing weak but consistent anomaly cues; and a Test-Time Self-Calibration (TSC) module, which calibrates anomaly scores at the decision level using test-time distribution statistics, ensuring that image-level predictions reflect the relative severity of abnormalities without modifying the underlying representations or requiring additional training.

Evaluated on multi-modality datasets, our framework achieves state-of-the-art performance, with average AUROCs of 82.16% (classification) and 96.50% (localization), offering a robust solution for clinical use. Homepage and code: https://wylab.sdu.edu.cn/syym/Few_shot_Medical_Anomaly_Detection_through_Centroi.htm.

1. Introduction

Medical image anomaly detection assumes a pivotal role in the early diagnosis and treatment of diseases, where the accurate identification of abnormal regions enables doctors to detect lesions at an earlier stage, thereby enhancing treatment efficacy and patient survival rates [1]. Nevertheless, the scarcity of medical image data presents a formidable challenge, given the close association with patient privacy and the low proportion of abnormal data, which makes collecting large-scale annotated data difficult [2]. As a result, this scarcity of annotated data circumscribes the application of traditional supervised learning approaches, which generally necessitate substantial amounts of labeled samples for reliable model training [3].

In this context, unsupervised learning methods have emerged, which do not require labeled data and can learn anomaly features directly from normal data. However, unsupervised methods also face numerous

* Corresponding author.
E-mail addresses: 202414371@mail.sdu.edu.cn (Z. Nie), ujnmhxu@hotmail.com (M. Xu), yuancui@sdu.edu.cn (Y. Cui), weihua@sdu.edu.cn (H. Wei), yi_wei@foxmail.com (W. Yi), sjniu@hotmail.com (S. Niu), wanyi@sdu.edu.cn (Y. Wan), xwei@bjmu.edu.cn (X. Wei), songweiye@sdu.edu.cn (W. Song).

https://doi.org/10.1016/j.patcog.2026.113261
Received 17 October 2025; Received in revised form 22 January 2026; Accepted 7 February 2026
Available online 6 March 2026

challenges, which require a large amount of normal data to train the model to ensure that the model learns effective feature representations[4]. Furthermore, the training process often requires extensive computing resources, which not only increases hardware costs but also limits their scalability in real-world medical scenarios [5].

In recent years, few-shot learning have gradually gained attention in medical image anomaly detection, which aims to effectively train models using limited labeled data to improve model generalization [6]. Current few-shot approaches have evolved along three main directions: feature representation methods that model normal feature distributions, reconstruction-based approaches that identify anomalies through reconstruction errors, and distillation frameworks that leverage teacher-student discrepancies. While these methods have shown promising results, they collectively face significant challenges in medical applications.

Specifically, under such limited supervision, models often struggle to learn representations that align with clinically meaningful anomaly cues, as subtle pathological variations are easily overshadowed by dominant anatomical structures in medical images [7], making it difficult to distinguish subtle pathological features from complex background noise, leading to high false positive rates [8]. In medical images, such background noise typically corresponds to normal but visually dominant anatomical structures or repetitive textures, which tend to overshadow subtle pathological variations in the learned representation space under limited supervision. As a result, anomaly-relevant cues can be suppressed at the representation level, causing background-driven false positives to emerge even before any anomaly scoring is applied. Moreover, existing methods are highly susceptible to local imaging artifacts and anatomical variations, which can further distort anomaly scores at the decision stage and result in unreliable detection outcomes [9].The sparse supervision in few-shot settings further exacerbates these issues, as decision boundaries tend to lie within the noise manifold, making them vulnerable to being dilated by acquisition artifacts into an irreducible false-positive regime.

To address these challenges, we propose a novel framework incorporating two complementary Plug-and-Play modules. Inspired by recent advances in vision-language models for anomaly detection [10,11], we leverage the potential of multimodal learning for improving anomaly detection performance. Our framework explicitly targets two sources of misalignment in few-shot medical anomaly detection: representation-level dominance of anomaly-irrelevant structures and decision-level distortion of anomaly scores under test-time variability. For the challenge of distinguishing subtle pathological features from background noise, we propose the Centroid Consultation Back (CCB) module, which refines patch-level representations by introducing intermediate semantic anchors that provide global contextual feedback. This design enables the model to better capture clinically relevant pathological patterns while effectively suppressing background interference. Furthermore, to mitigate the adverse effects of local noise and artifacts on anomaly scoring, we introduce the Test-Time Self-Calibration (TSC) module. Rather than modifying representations, this module calibrates anomaly scores at the decision level using test-time distribution statistics, ensuring that anomaly scores accurately reflect the true abnormality level across the entire image, rather than being disproportionately influenced by localized artifacts. In addition, we compute training loss by integrating both global and local features, which are compared against textual features, thereby enhancing the overall precision and reliability of the anomaly detection process.

Our framework was extensively experimented on six medical image datasets spanning five imaging modalities. The results demonstrate that the CCB module significantly reduces false positives and improves the detection of subtle anomalies, while the TSC module ensures the accuracy and consistency of anomaly scores. Our framework achieves state-of-the-art performance in both anomaly classification and localization with an average AUROC of 82.16% and 96.50% respectively, significantly outperforming existing methods. This demonstrates the significant advantages of our framework in few-shot anomaly detection in medical images, providing a robust and efficient solution for clinical anomaly detection.

The key contributions are:

- We propose a novel framework for few-shot anomaly detection in medical images. This framework integrates global and local features and compares them with textual features to optimize the computation of anomaly maps, significantly improving the accuracy and reliability of anomaly detection.
- We introduce the Centroid Consultation Back (CCB) module to refine patch-level representations by suppressing anomaly-irrelevant anatomical structures and enhancing subtle pathological cues under limited supervision.
- We propose the Test-Time Self-Calibration (TSC) module, which calibrates anomaly scores at inference time to correct decision-level distortions caused by local noise and artifacts.
- Extensive experiments on multiple medical image datasets demonstrate that this framework achieves an average AUROC of 82.16% on anomaly classification and 96.50% on anomaly localization, significantly outperforming existing methods.

2. Related works

Medical image anomaly detection methods mainly fall into three paradigms: feature representation, image reconstruction, and knowledge distillation, each addressing the unsupervised setting-learning solely from normal data-through different mechanisms [12].

2.1. Feature representation-based methods

Feature representation-based approaches detect anomalies by modeling the distribution of normal features and identifying deviations. PatchCore [13] improves efficiency via feature subsampling but may miss subtle pathological patterns in low-density regions, which is critical for medical imaging.

Memory-guided methods [14] introduce prototype memories to model normal pattern diversity, reducing over-representation and improving anomaly separation, while MOCCA [15] captures multi-scale normality through hierarchical distributions but remains sensitive to inter-patient anatomical variation.

Despite their effectiveness, these methods often struggle to jointly preserve fine-grained pathological details and global contextual consistency, limiting robustness across heterogeneous anatomical presentations.

2.2. Reconstruction-based methods

Reconstruction-based methods identify anomalies through elevated reconstruction errors when decoding normal patterns. Autoencoder-based approaches [16] preserve structural information but are sensitive to imaging artifacts, while GAN-based inpainting methods [17] struggle with diffuse or boundary-less pathologies.

Recent works reconstruct semantic features instead of raw pixels to mitigate these issues. ADTR [18] employs transformers to restrict anomaly reconstruction while preserving normal representations. Moreover, Hetero-AE [19] introduces a heterogeneous auto-encoder architecture for medical anomaly detection. By using a convolutional neural network (CNN) as the encoder and a hybrid CNN-Transformer network as the decoder, Hetero-AE enables the model to capture intrinsic information from normal data while enlarging the difference on abnormal samples, addressing the challenges of pixel-wise reconstruction and overfitting.

Nevertheless, reconstruction-based methods remain limited by their reliance on global reconstruction metrics and their tendency to hallucinate plausible tissue patterns, which can obscure subtle anomalies.

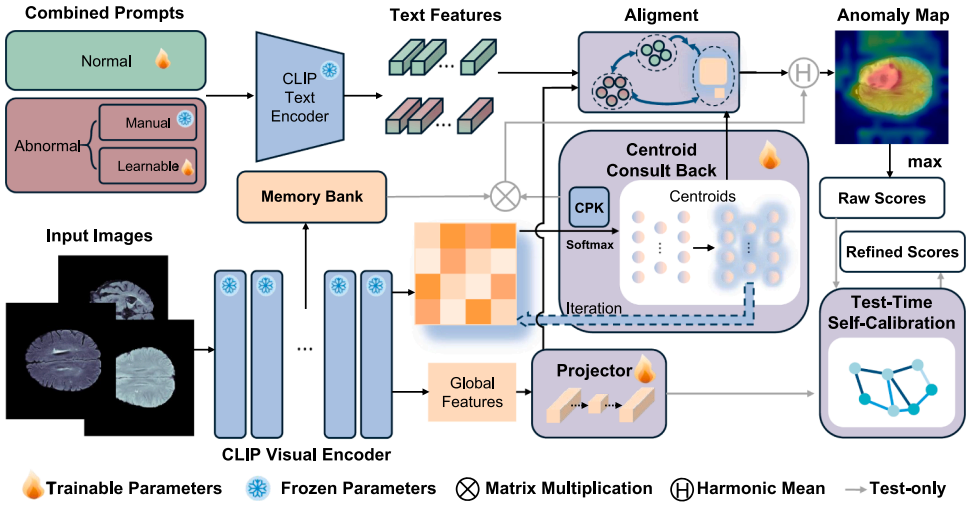


Fig. 1. Overview of the anomaly detection framework for few-shot medical images. CLIP-encoder extracts global CLS features and local patch features simultaneously from input images. The CCB module refines patch-level representations by introducing global semantic feedback, suppressing background-driven responses and enhancing subtle anomaly cues. During the test phase, the TSC module calibrates anomaly scores at the decision level using test-time distribution statistics, mitigating the influence of localized noise and artifacts. Global and local features are jointly aligned with textual representations to optimize both localization and classification objectives.

2.3. Distillation-based methods

Distillation-based approaches detect anomalies by amplifying discrepancies between teacher and student representations. RD4AD [20] exploits multi-scale discrepancies but inherit biases from domain-agnostic teachers and often suppress fine-grained lesion details.

IKD [21] further addresses overfitting by distilling informative features, while reverse distillation variants [22] introduce latent anomaly suppression to mitigate abnormal pattern leakage.

Overall, distillation-based methods often struggle to provide anatomy-aware guidance and stable anomaly scoring, leading to blurred lesion boundaries and elevated background false positives due to insufficient integration of global context and local pathology.

3. Method

3.1. Visual language model workflow

Our proposed few-shot medical image anomaly detection framework builds upon a pre-trained vision-language model (CLIP) [23] and integrates two Plug-and-Play modules that operate at different stages of the detection pipeline: the Centroid Consultation Back (CCB) module for representation refinement and the Test-Time Self-Calibration (TSC) module for anomaly score calibration.

Given a support set of K-shot normal images $\mathcal{D}_{\text{train}} = \{\mathbf{I}_i\}_{i=1}^{K}$, we extract multi-scale visual features using the CLIP image encoder with the V-V attention mechanism proposed in CLIP-Surgery [24], which enhances local feature sensitivity while preserving the original representation structure. For each image, the encoder outputs a global feature vector and patch-level features, which together form the basis for subsequent representation refinement and anomaly scoring:

$$\mathbf{F}_{\text{global}}, \mathbf{F}_{\text{local}} = \text{Encoder}_{\text{vis}}(\mathbf{I}) \tag{1}$$

To support few-shot inference, intermediate visual features extracted from normal samples are cached as reference statistics rather than explicit prototypes. For each input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, the encoder produces intermediate feature maps $\mathbf{F}^{(l)} \in \mathbb{R}^{h_l\times w_l\times d_l}$ from multipl e layers, which are stored in a memory bank $\mathcal{M}$:

$$\mathcal{M} = \{\mathbf{F}_i^{(l)} \mid i = 1, \ldots, K; l = 1, \ldots, L\} \tag{2}$$

The memory bank caches multi-scale visual features extracted from K normal training images and provides a lightweight normality prior at inference time. Rather than performing hard retrieval, we aggregate similarities over the cached normal patterns to obtain a stable normal-reference score that complements the image-text anomaly evidence.

Subsequently, the CCB module processes the patch-level features to produce refined representations that emphasize anomaly-relevant variations while suppressing dominant background structures:

$$\mathbf{F}_{\text{cent}} = \text{CCB}(\mathbf{F}_{\text{local}}) \tag{3}$$

For the text branch, learnable prompts for both normal and abnormal classes are encoded by the CLIP text encoder. Following PromptAD [7] and AdaCLIP [25], the normal prompts $\mathbf{P}_{\text{n}}$ are fully learnable, while abnormal prompts combine a fixed manually-designed component and a learnable component to enrich anomaly semantics.

Global and refined local visual features are jointly aligned with textual representations through a multi-objective learning scheme, enabling consistent optimization of image-level classification and pixel-level localization without introducing additional supervision. To align visual and textual features, we employ a multi-objective loss function consisting of three complementary components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{triplet}} + \mathcal{L}_{\text{global}} \tag{4}$$

where $\mathcal{L}_{\text{patch}}$ denotes the patch-level image-text contrastive loss that maximizes the similarity between local image features and normal text features while minimizing similarity to abnormal text features:

$$\mathcal{L}_{\text{patch}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(s_{pn}/\tau)}{\exp(s_{pn}/\tau)+\exp(s_{pa}/\tau)}\right) \tag{5}$$

where $s_{pn} = \langle \mathbf{F}_{\text{cent}}^{(i)}, \psi(\mathbf{P}_{\text{n}})\rangle$, $s_{pa} = \langle \mathbf{F}_{\text{cent}}^{(i)}, \psi(\mathbf{P}_{\text{a}})\rangle$, N is the number of samples, τ is the temperature parameter scaling the logits, and ψ is global feature projector consisting of two layers of fully connected residual branches with ReLU activation. $\mathcal{L}_{\text{triplet}}$ represents the triplet margin loss that enhances the discrimination between normal and abnormal features:

$$\mathcal{L}_{\text{triplet}} = \max(0, \|(\mathbf{F}) - \psi(\mathbf{P}_{\text{n}})\|_2^2 - \|(\mathbf{F}) - \psi(\mathbf{P}_{\text{a}})\|_2^2 + m) \tag{6}$$

where m is the margin value that controls the separation degree between positive and negative pairs.

$\mathcal{L}_{\text{global}}$ indicates the global feature alignment loss:

$$\mathcal{L}_{\text{global}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(s_{gn}/\tau)}{\exp(s_{gn}/\tau)+\exp(s_{ga}/\tau)}\right) \tag{7}$$

where $s_{gn} = \langle \mathbf{F}^{(i)}_{\text{global}}, \psi(\mathbf{P}_{\text{n}}) \rangle$, $s_{ga} = \langle \mathbf{F}^{(i)}_{\text{global}}, \psi(\mathbf{P}_{\text{a}}) \rangle$, $\mathbf{F}^{(i)}_{\text{global}}$ represents the global [CLS] token feature of the ith sample. This ensures consistency between global image features and text representations in both normal and abnormal semantic spaces.

3.2. Centroid consultation back (CCB) module

The Centroid Consultation Back (CCB) module is designed to refine patch-level representations under limited supervision by correcting representation-level dominance of anomaly-irrelevant structures. Rather than using centroids as explicit decision references for feature matching or anomaly scoring, as in prototype- or clustering-based methods, CCB treats centroids as intermediate semantic anchors that aggregate global contextual statistics and actively modulate local representations through a feedback mechanism. Specifically, the assignment-based aggregation groups patches according to the consistency of their feature responses rather than spatial proximity, allowing semantically coherent but spatially scattered anomaly cues to be jointly integrated into the same centroid representation. As a result, centroid features capture low-variance semantic modes under few-shot supervision, providing a stable global context that feeds back to local patches and improves anomaly separability at the representation level.

This consultation mechanism enables subtle pathological cues to be enhanced while suppressing visually dominant but anomaly-irrelevant background patterns. At its core, CCB decouples centroid estimation from anomaly decision. The centroids are not used as explicit reference embeddings for similarity-based classification; instead, they serve as context carriers that mediate information exchange between local patches and global feature statistics.

The technical workflow begins with center probability prediction using the Centroid Predictor Kernel (CPK), which is implemented as a lightweight 1×1 convolutional assignment predictor. CPK predicts assignment logits that indicate the affinity between each spatial location and a fixed set of latent centroid slots, from which centroid features are subsequently constructed via weighted aggregation of local features. These logits are optimized jointly with the backbone in an end-to-end manner.

$$\mathbf{P}_{\text{center}} = \text{CPK}(\mathbf{X}) \tag{8}$$

where $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ is the input feature map. To stabilize optimization under few-shot supervision, we employ Gumbel-Softmax sampling to obtain stochastic hard assignments:

$$\mathbf{Q} = \text{GumbelSoftmax}(\mathbf{P}_{\text{center}}, \gamma) \tag{9}$$

where γ denotes the assignment temperature that controls the sharpness of the stochastic routing.

Given the current assignment matrix, centroid features are computed as weighted aggregations of local features and iteratively refined:

$$\mathbf{C}^{(t+1)} = \text{Self-Attention}\left(\frac{\mathbf{Q}^{(t)}\mathbf{X}}{\sum \mathbf{Q}^{(t)}}\right) \tag{10}$$

where $\mathbf{C}^{(t)} \in \mathbb{R}^{B\times M\times C}$ denotes the centroid features at iteration t.

This stochastic assignment avoids reliance on fixed initialization or warm-up strategies, while enabling flexible aggregation of local features. The iterative refinement is not intended to form stable clusters, but to progressively integrate global contextual information into the centroid representations.

After refinement, the globally-enhanced centroid features are broadcast back to the spatial domain to correct local representations:

$$\mathbf{F}_{\text{cent}} = \mathbf{X} + f_{\text{recon}}(\mathbf{Q}^{(T)}\mathbf{C}^{(T)}) \tag{11}$$

where f_{recon} is a lightweight reconstruction projection. This residual consultation step injects global semantic context into local features without altering their spatial structure, yielding representation-level correction rather than explicit re-clustering. The consultation mechanism enables subtle pathological cues to be enhanced while suppressing visually dominant but anomaly-irrelevant background patterns.

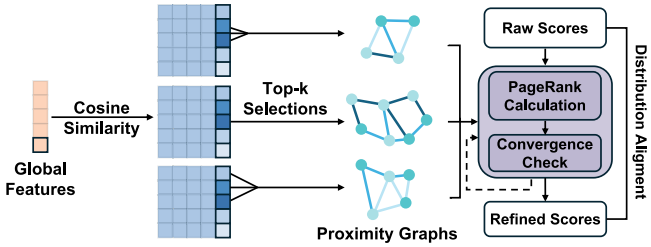


Fig. 2. Workflow of the Test-Time Self-Calibration (TSC) module. TSC operates exclusively on anomaly scores at inference time and does not modify visual representations. Initial scores are calibrated via similarity-aware propagation to suppress localized noise while preserving global score ordering.

The detailed procedure of the Centroid Consultation Back (CCB) module is summarized in Algorithm 1.

Algorithm 1 Centroid Consultation Back (CCB).

Require: Input feature $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, centroids M, iterations T, temperature γ
Ensure: Enhanced feature $\mathbf{Y} \in \mathbb{R}^{B\times C\times H\times W}$

1. Center Initialization
$\mathbf{P} \leftarrow \text{Conv}1\times 1(\mathbf{X})$ ▷ Predict center probabilities
$\mathbf{Q} \leftarrow \text{GumbelSoftmax}(\mathbf{P}, \tau = \gamma)$ ▷ Hard assignment
$\mathbf{C} \leftarrow \text{WeightedAverage}(\mathbf{X}, \mathbf{Q})$ ▷ Initial centroids
2. Iterative Refinement
for $t = 1$ **to** T **do**
  $\mathbf{S} \leftarrow \text{Similarity}(\mathbf{C}, \mathbf{X})$ ▷ Cosine similarity
  $\mathbf{Q} \leftarrow \text{Softmax}(\mathbf{S}/\gamma)$ ▷ Update assignments
  $\mathbf{C} \leftarrow \text{WeightedAverage}(\mathbf{X}, \mathbf{Q})$ ▷ Update centroids
end for
3. Global Context Integration
$\mathbf{C}_{\text{refined}} \leftarrow \text{SelfAttention}(\mathbf{C})$ ▷ Enhance centroids with global context
4. Feature Reconstruction
$\mathbf{X}_{\text{recon}} \leftarrow \text{Broadcast}(\mathbf{C}_{\text{refined}}, \mathbf{Q})$ ▷ Map back to pixel space
$\mathbf{Y} \leftarrow \mathbf{X} + \text{Conv}1\times 1(\mathbf{X}_{\text{recon}})$ ▷ Residual connection
return $\mathbf{Y}$

3.3. Test-time self-calibration (TSC)

Inspired by calibration approaches such as MUSC [9] and Batch-Normalization [26], which optimize anomaly scores using multiple test samples simultaneously without explicitly requiring supervision or retraining, the Test-Time Self-Calibration (TSC) module also leverages a multiple samples inference setup.

The TSC module (Fig. 2) addresses the tendency of localized noise or artifacts to disproportionately inflate image-level anomaly scores in few-shot settings. Unlike graph-based smoothing or kNN similarity propagation methods that typically rely on neighborhood-based score aggregation across similar samples, TSC formulates calibration as a constrained score redistribution process that preserves the global ordering of the original scores while attenuating isolated extreme responses. The optimization is fully training-free and operates only during inference.

The calibration process begins by constructing a similarity graph based on global visual representations:

$$\mathbf{W}_{ij} = \frac{\mathbf{F}^{(i)T}_{\text{global}}\mathbf{F}^{(j)}_{\text{global}}}{\|\mathbf{F}^{(i)}_{\text{global}}\|\|\mathbf{F}^{(j)}_{\text{global}}\|} \tag{12}$$

where $\mathbf{F}^{(i)}_{\text{global}} \in \mathbb{R}^d$ denotes the [CLS] feature of the ith test sample. These similarities are used only to define score proximity, rather than to propagate features or semantic labels.

Instead of relying on fixed k-nearest neighbors or heuristic thresholds, we adopt an adaptive neighborhood strategy that balances local

connectivity and global sparsity:

$$k_i = \max\left(k, \sum_{j=1}^{N} \mathbb{I}(\mathbf{W}_{ij} > \theta) - 1\right) \tag{13}$$

where $\theta = \text{percentile}(\mathbf{W}, \beta)$ controls graph sparsity. This adaptive selection avoids excessive smoothing while ensuring sufficient contextual support for calibration.

The adjacency matrix is constructed as:

$$\mathbf{W}_{ij}^{\text{adaptive}} = \begin{cases} \mathbf{W}_{ij} & \text{if } j \in \mathcal{N}(i) \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

To calibrate anomaly scores, we employ a PageRank-style iterative update:

$$\mathbf{s}^{(t+1)} = \alpha \mathbf{P}\mathbf{s}^{(t)} + (1-\alpha)\mathbf{s}^{(0)} \tag{15}$$

where $\mathbf{s}^{(0)}$ denotes the initial anomaly scores. This procedure redistributes anomalous evidence, mitigating the impact of isolated extreme values without enforcing global smoothing. The restart term $(1-\alpha)\mathbf{s}^{(0)}$ anchors the propagation to the original anomaly evidence, preventing the iterative process from drifting toward spurious consensus under noisy similarity neighborhoods and ensuring convergence to a stable, order-preserving fixed point.

For robustness, we compute calibrated scores over a multi-k set $\mathcal{K}$ and aggregate them via a weighted average:

$$\mathbf{s}^{\text{opt}} = \sum_{k \in \mathcal{K}} w_k \mathbf{s}^{(k)} \tag{16}$$

Finally, a distribution-preserving normalization rescales the calibrated scores:

$$\hat{s}_i = \left(\frac{s_i^{\text{opt}} - \min_j s_j^{\text{opt}}}{\max_j s_j^{\text{opt}} - \min_j s_j^{\text{opt}}}\right)(\max_j s_j^{(0)} - \min_j s_j^{(0)}) + \min_j s_j^{(0)} \tag{17}$$

This step preserves the global ordering and statistical range of the original scores, allowing TSC to function as a score calibration procedure that stabilizes image-level predictions under test-time variability.

To further justify the choice of the PageRank-based update in TSC, we compare it with alternative score-level propagation operators implemented within the same calibration framework, including kNN averaging, Laplacian smoothing, and label propagation, under identical similarity graphs and initial anomaly scores.

The results, reported in Table S-3 of the Supplementary Material, demonstrate that the PageRank (random walk with restart) operator consistently achieves higher image-level AUROC under this controlled setting. This comparison directly validates our design choice: when employed as a score-level calibration operator, the restart-based propagation provides a more reliable mechanism for redistributing anomaly evidence, effectively mitigating isolated extreme responses while preserving globally informative anomaly ordering.

3.4. Testing

The testing phase of our framework is designed to leverage both the learned representations and the stored normal patterns for robust anomaly detection. The process begins with feature extraction, where test images are encoded through the CLIP backbone to obtain multi-scale visual features, which are subsequently enhanced by the CCB module.

In the few-shot setting, all test samples are treated as unlabeled, and no abnormal samples are accessed or used for calibration. Specifically, we first compute an image-level anomaly score based on image–text alignment, which reflects the degree of semantic deviation from normal textual descriptions:

$$s_{\text{anomaly}} = \text{sim}(\mathbf{F}_{\text{img}}, \mathbf{F}_{\text{text}}) \tag{18}$$

where $\mathbf{F}_{\text{img}}$ and $\mathbf{F}_{\text{text}}$ denote the global visual and textual features, respectively.

A key component of the testing pipeline is the use of the normal memory bank as a reference for estimating normality, rather than as a direct anomaly decision criterion.

$$s_{\text{memory}} = \frac{1}{|\mathcal{M}|} \sum_{\mathbf{F}_{\text{mem}} \in \mathcal{M}} \text{sim}(\mathbf{F}_{\text{cent}}, \mathbf{F}_{\text{mem}})e \tag{19}$$

This score reflects the overall consistency of the test sample with normal patterns and serves as a stabilizing normality prior, rather than a nearest-neighbor or prototype matching score.

Subsequently, the memory-based similarity is then fused with the direct anomaly score derived from image-text feature comparison using harmonic mean integration:

$$s_{\text{fused}} = \frac{2 \cdot s_{\text{anomaly}} \cdot s_{\text{memory}}}{s_{\text{anomaly}} + s_{\text{memory}}} \tag{20}$$

The harmonic fusion balances text-guided anomaly evidence with the normality prior, preventing isolated background responses from dominating the final score.

The framework simultaneously generates pixel-level anomaly maps for precise localization while producing image-level classification scores. The final anomaly score is obtained through a two-step process: first extracting the maximum value from the anomaly map as the initial score, then refining it through the TSC module:

$$\hat{s}_{\text{final}} = \text{TSC}\big(\max(\mathbf{S}_{\text{map}})\big) \tag{21}$$

where $\mathbf{S}_{\text{map}}$ represents the pixel-wise anomaly score map, and $\hat{s}_{\text{final}}$ denotes the final calibrated anomaly score.

4. Experimental setup

4.1. Datasets

All experiments are carried out on the BMAD benchmark [27], which includes five different medical imaging modalities and anatomical regions: BrainMRI [28–30], LiverCT [31,32], retinal OCT [33,34], chest X-ray [35], and digital histopathology [36]. The detailed information of each subset is listed in Table S-1 and Section 1 of the Supplementary Materials.

4.2. Competing methods and baselines

We compare our method with representative state-of-the-art anomaly detection (AD) approaches under different training assumptions. Specifically, we consider two categories of baselines: (i) *few-normal-shot* methods, including CLIP [8], MedCLIP [37], WinCLIP [38], SimpleNet [39], MSFlow [40], URD [41], MHKD [42], INCTRL [43] and INP-Former [44] (INCTRL and INP-Former only support image-level anomaly detection, and thus we report its image-level AUROC only); and (ii) *fully unsupervised* methods that utilize all normal data, including CFlowAD [45], RD4AD [46], PatchCore [13], MKD [2], Dinomaly [47], FSR [48], E2AD [49] and EDC [50] (E2AD and EDC only support image-level anomaly detection, and thus we report its image-level AUROC only).

The experimental results for the baseline methods are primarily derived from Huang et al. [51] who collated and supplemented the experimental results of BMAD [27], while the results for SimpleNet [39], msflow [40], URD [41], MHKD [42], INCTRL [43], INP-Former [44], Dinomaly [47], FSR [48], E2AD [49] and EDC [50] are obtained through our experiments under the same benchmark settings.

In addition, we discuss several CLIP-based medical AD methods with different supervision assumptions. MadCLIP [52] adopts a stronger few-shot supervised setting by using both normal and abnormal samples during training (e.g., 4 normal + 4 abnormal), which differs from the few-normal assumption in our work; its results are included for reference and

interpreted accordingly. We also qualitatively discuss AnoCLIP [53], whose official implementation is not publicly available.

It is notable that the comparison methods used in the paper are primarily based on single-inference, where each image is processed independently, which is different from the ohers that depend on multiple samples inference to optimize anomaly scores. In our approach, though the TSC module improves performance by calibrating across a set of samples, when only a single test sample is available, the CCB module and the overall framework continue to operate effectively, providing robust anomaly detection.

4.3. Evaluation metric

We adopt the Area Under the Receiver Operating Characteristic Curve (AUROC) as the evaluation criterion for both image-level anomaly detection and pixel-level anomaly localization. AUROC summarizes the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across all possible decision thresholds.

4.4. Implementation details

All experiments were executed on a single NVIDIA A100 40 GB GPU with CUDA 12.2 and PyTorch 2.6. We adopt the CLIP with ViT-B/16+ vision transformer, pre-trained on the 400 M image-text pairs of LAION-400M [54]. The encoder contains 86 M parameters, employs a patch size of 15×15 pixels, and produces 768-dimensional patch embeddings. All weights are frozen; no gradient updates are applied.

The general and object-customized text prompts are provided in Section 2 of the Supplementary Material. All images are resized to 240×240 and normalized using ImageNet [55] statistics.

For the CCB module, the Gumbel–Softmax temperature is set to $\gamma = 0.1$ to balance exploration and exploitation during stochastic centroid assignment, and the number of centroids is fixed to $M = 20$ across all experiments. For the TSC module, we employ multiple neighborhood scales $\mathcal{K} = \{11, 13, 15\}$ for adaptive graph construction. The similarity threshold is set to the 70th percentile ($\beta = 0.7$) to ensure sparse yet meaningful connectivity, and the PageRank restart probability is fixed to $\alpha = 0.9$ to balance local consistency and global score fidelity. Ablation studies for these hyperparameters are reported in Section 4.4 and the Supplementary Material.

Unless otherwise specified, the three loss terms are equally weighted during training. We empirically observe that the proposed framework is insensitive to moderate variations in loss weighting. An additional analysis on loss-weight robustness and gradient magnitudes is provided in Tables S-7 and S-8 of the Supplementary Material.

4.5. Computational efficiency

We analyze the computational efficiency of the proposed framework by comparing its inference time with representative baseline methods reported in Table 1. As summarized in Table S-4, the proposed method maintains competitive per-image inference latency, while introducing only a modest overhead compared to backbone-only inference, despite incorporating both representation refinement (CCB) and test-time score calibration (TSC).

We further analyze the computational cost and scalability of the proposed TSC module. As summarized in Tables S-5 and S-6 in the Supplementary Material, the proposed framework exhibits stable and predictable runtime behavior. Specifically, the dataset-level calibration introduced by TSC scales approximately linearly with the test-set size, while the per-image feature extraction and anomaly map generation remain unchanged.

5. Results and discussion

5.1. Evaluation on image and pixel level tasks

We comprehensively compare our framework against state-of-the-art methods under few-normal (few-shot) and full-normal settings (Tables 1 and 3). Our method shows significant progress across six medical imaging datasets, especially in data-scarce few-shot scenarios.

As shown in Table 1, our method consistently outperforms existing approaches under the 4-shot setting on both image-level and pixel-level tasks. The improvement is particularly pronounced on datasets with subtle or diffuse anomalies, reflecting the complementary roles of CCB in refining local representations and TSC in stabilizing image-level anomaly scores. Notably, while some methods benefit from domain overlap on specific datasets (e.g., ChestXray), our framework maintains robust performance across modalities, indicating stronger generalization under limited supervision.

Fig. 3 shows that the proposed method produces compact and accurate anomaly maps with minimal background activation across modalities. This behavior is consistent with the design of CCB, which suppresses visually dominant but anomaly-irrelevant structures while preserving coherent lesion boundaries.

In all visualizations, there are almost no false positives outside the abnormal region, demonstrating the high specificity of our method in distinguishing abnormal from normal tissue, which is particularly important for clinical diagnosis, as it reduces the workload for doctors interpreting images and avoids unnecessary further examinations due to false positives.

However, to better understand the limitations of the proposed method, we further analyze representative failure cases on datasets with pixel-level annotations. Fig. 4 presents representative false positive and false negative cases. False positives mainly arise from strong structural variations, while false negatives correspond to extremely subtle or low-contrast lesions, reflecting intrinsic challenges of medical anomaly detection under limited supervision.

Since the proposed TSC module calibrates image-level anomaly scores at test time, we further examine whether the method overly depends on the availability of the entire test set. Specifically, we randomly partition the BrainMRI test set into 1, 2, and 3 non-overlapping subsets of equal size, using a stratified strategy to preserve the normal-to-abnormal ratio across subsets, and apply the complete inference pipeline independently to each subset. Table 2 shows that reducing test-set size results in only marginal performance variation, indicating that TSC does not rely on specific test-set distributions.

Fig. 5 illustrates that high anomaly scores arise from coherent semantic deviations rather than isolated intensity extremes. This observation supports effective image-level discrimination by showing that abnormality is determined by globally consistent semantic patterns instead of local grayscale outliers.

Under the full-normal setting (Table 3), the proposed method achieves performance comparable to or exceeding representative unsupervised baselines across multiple datasets. This result indicates that the framework does not rely on extreme data scarcity to be effective, and remains robust when more normal training data are available. Together with the few-shot results, this suggests that the proposed design generalizes across different data availability regimes.

5.2. CCB visual analysis

Fig. 6 provides a qualitative illustration of the effect of the CCB module on feature representations. Compared with the original CLIP features, CCB effectively suppresses scattered background responses while enhancing spatially coherent lesion structures, resulting in more anatomically consistent feature maps. This observation supports the role of CCB as a representation-level refinement module that operates on patch-level features prior to anomaly scoring. Additional qualitative

Table 1
AUROC(%) Performance under Few-normal (4-shot) Settings.

Shot	Method	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
		Image	Image	Image	Image	Pixel	Image	Pixel	Image	Pixel
few-normal	CLIP [8]	63.48	70.74	98.59	74.31	93.44	56.74	97.20	84.54	95.03
	MedCLIP [37]	**75.89**	**84.06**	81.39	76.87	90.91	60.65	94.45	66.58	88.98
	WinCLIP [38]	67.49	70.00	97.89	66.85	94.16	67.19	96.75	88.83	**96.68**
	SimpleNet [39]	51.10	64.32	63.80	47.20	69.20	57.00	46.90	55.30	55.41
	msflow [40]	63.57	60.22	87.34	45.70	67.40	68.89	69.89	86.70	92.68
	URD [41]	63.47	52.05	51.71	43.28	62.03	61.55	62.87	70.60	64.30
	MHKD [42]	51.68	55.09	61.14	64.62	95.38	57.55	95.57	50.76	83.27
	INP-Former [44]	40.49	52.39	49.76	47.59	47.41	62.54	42.38	40.72	40.85
	INCTRL [43]	62.89	71.42	51.78	71.20	/	61.45	/	61.41	/
	ours	69.06	75.38	**99.01**	**86.70**	**95.73**	**69.89**	**97.93**	**92.91**	95.85

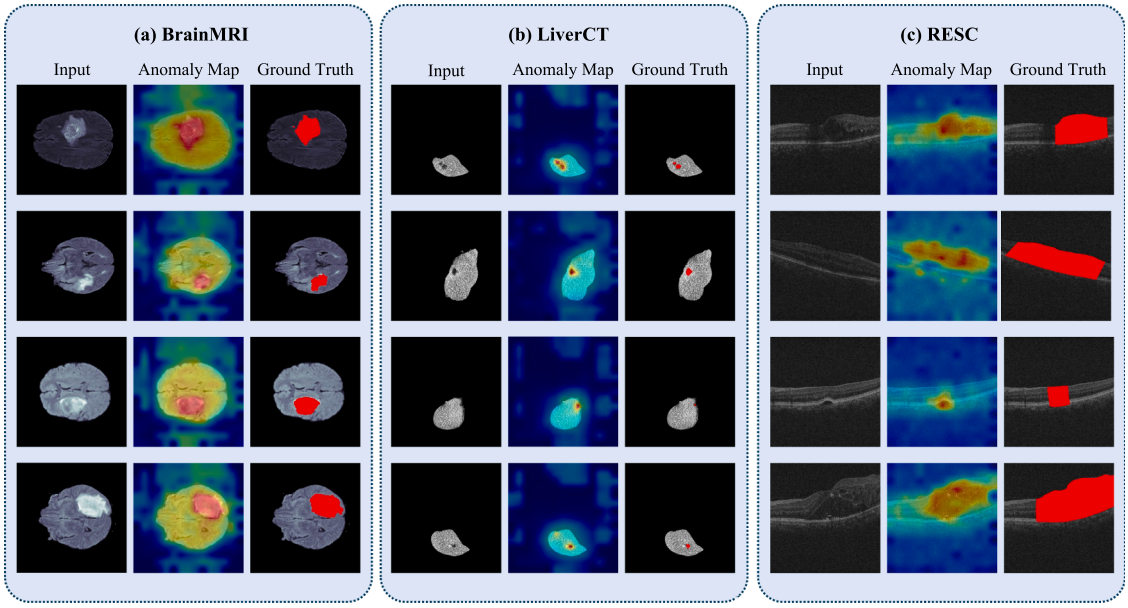


Fig. 3. Visualization of anomaly localization across modalities: (a) BrainMRI (b) LiverCT (c) RESC. Each row shows the input image, the anomaly map generated by our method, and the corresponding ground truth.

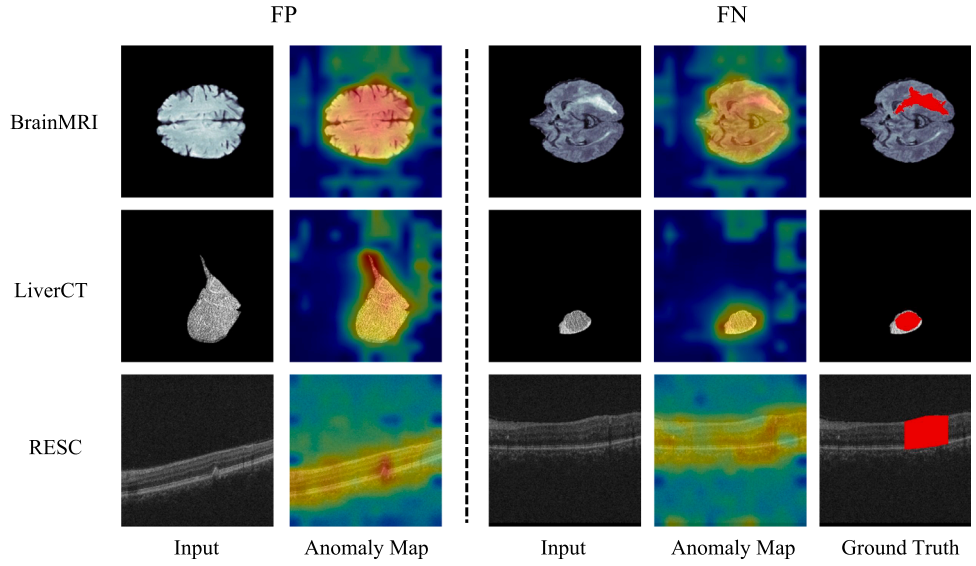


Fig. 4. Representative failure cases on localization datasets. Each row shows one false positive (FP) and one false negative (FN) example from BrainMRI, LiverCT, and OCT (RESC), including the input image, the predicted anomaly map, and the ground truth (for FN).

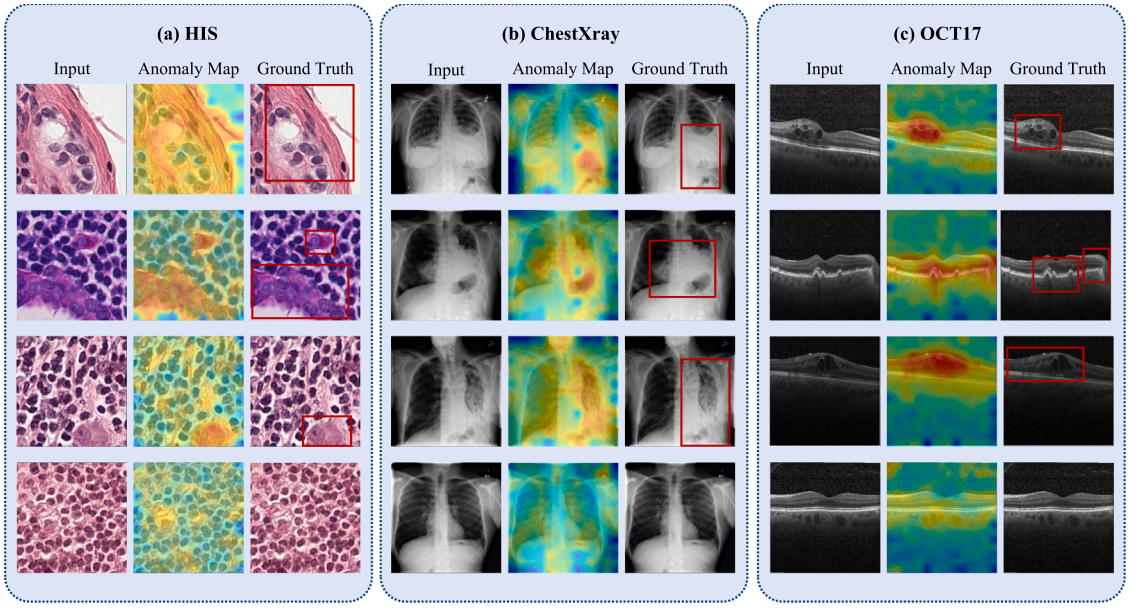


Fig. 5. Visualization of heatmaps on anomaly classification across modalities: (a) HIS (b) ChestXray (c) OCT17. Each row shows the input image, the anomaly map generated by our method, and the corresponding ground truth, where the area within the red frame is abnormal, and the images without the red frame are normal samples.

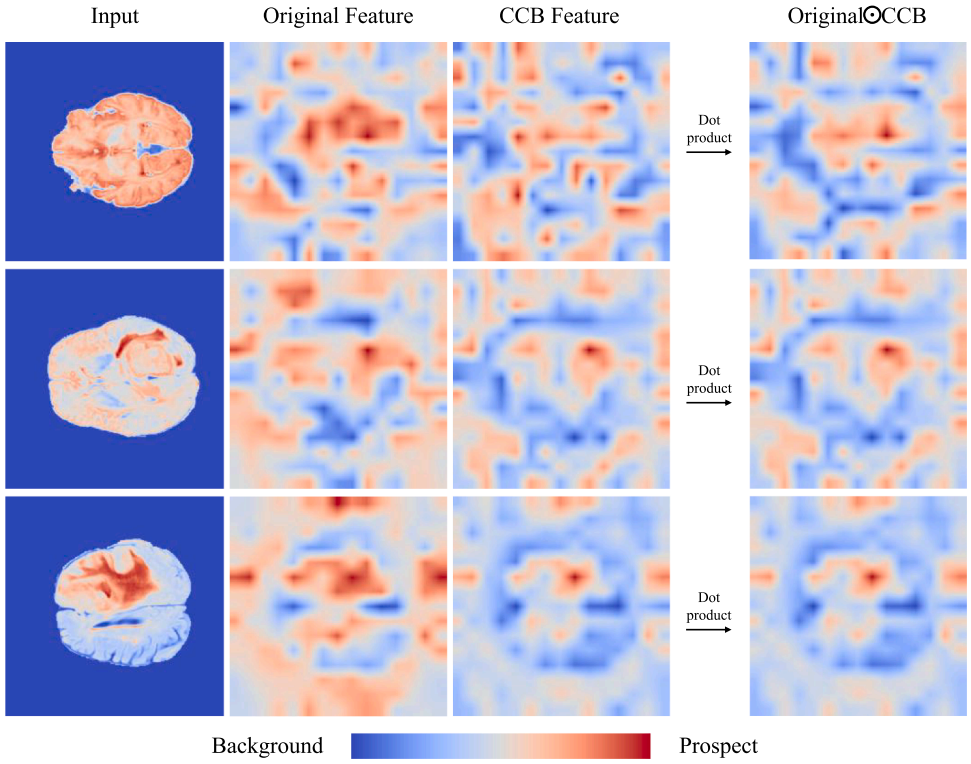


Fig. 6. Visual comparison of original features and CCB-enhanced features. From left to right: the input image, original features (CLIP output), features after processing the CCB module, and the dot product response of the original and CCB features.

Table 2
Effect of test set partitioning on image-level performance on BrainMRI.

Test Set Partition	Samples per Subset	Image-level AUROC (%)
1 subset	3715	86.70
2 subsets	1858	86.43
3 subsets	1239	85.89

analysis of feature responses before and after CCB processing is shown in Fig. S-2 of the Supplementary Material.

Fig. 7 further visualizes the effect of CCB on patch-level feature distributions using t-SNE on BrainMRI. Compared with the original CLIP features, CCB produces more structured and compact feature manifolds, where abnormal patches are less entangled with dominant anatomical patterns. Notably, it can be observed that normal features tend to form several distinct clusters, although the boundaries are not always sharply

Table 3
AUROC(%) Performance under full-normal settings.

Setting	Method	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
		Image	Image	Image	Image	Pixel	Image	Pixel	Image	Pixel
full-normal-shot	CFlowAD [45]	54.54	71.44	85.43	73.97	93.52	49.93	92.78	74.43	93.75
	RD4AD [46]	66.59	67.53	97.24	89.38	96.54	60.02	95.86	87.53	96.17
	Patchcore [13]	69.34	75.17	98.56	**91.55**	96.97	60.40	96.58	91.50	96.39
	MKD [2]	**77.74**	**81.99**	96.62	81.38	89.54	60.39	96.14	88.97	86.60
	Dinomaly [47]	70.47	74.68	98.34	90.99	**97.33**	73.32	97.65	94.17	95.81
	FSR [48]	68.81	72.88	72.32	91.22	96.21	57.50	97.72	94.32	93.49
	EDC [50]	58.91	73.86	94.65	82.48	/	63.21	/	89.53	/
	E2AD [49]	60.50	69.00	76.43	65.20	/	54.64	/	71.76	/
	ours	71.37	80.56	**99.73**	91.27	96.25	**85.90**	**98.77**	**96.88**	**96.73**

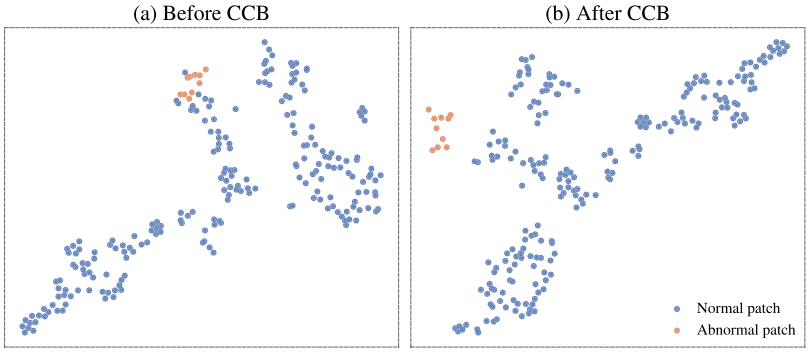


Fig. 7. t-SNE visualization of patch-level features on the BrainMRI dataset before and after applying the CCB module. (a) Original CLIP features without CCB. (b) Features refined by the CCB module.

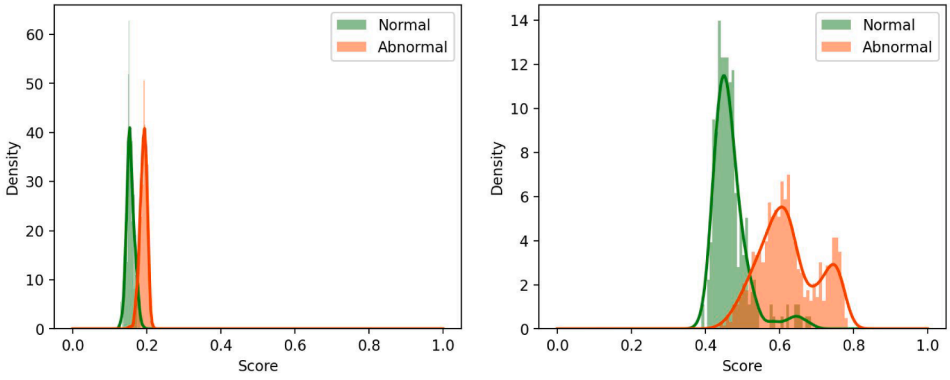


Fig. 8. Histogram of anomaly score distribution. (a) Original score; (b) TSC-optimized score. Orange: normal samples; Green: anomaly samples.

defined. This clustering pattern suggests that the CCB module enhances the model's ability to distinguish between different normal regions, such as cortex and white matter, and abnormalities, leading to more accurate anomaly detection by focusing on subtle pathological features rather than being overwhelmed by dominant anatomical structures.

5.3. TSC score calibration analysis

Fig. 8 illustrates the effect of TSC on anomaly score distributions. Before calibration, normal and abnormal samples exhibit heavily overlapping score modes, making threshold selection unstable. After TSC, scores are reorganized through similarity-based propagation, leading to clearer separation between normal and abnormal samples without sacrificing sensitivity.

5.4. Ablation study

Table 4 reports the ablation results under the few-shot setting and highlights the complementary roles of the proposed components. Using the pretrained vision–language backbone alone yields limited performance, indicating that direct feature transfer is insufficient for robust medical anomaly detection.

Table 4
Component contribution in few-shot settings.

$\mathcal{F}_{VLM}$	$\mathcal{CCB}$	$\mathcal{TSC}$	$\mathcal{GLOBAL}$	AUROC (Image)	AUROC (Pixel)
✓				79.33	92.13
✓	✓			79.54	94.54
✓		✓		84.23	92.13
✓	✓	✓		85.36	94.54
✓	✓	✓	✓	**86.70**	**95.73**

Introducing CCB improves pixel-level AUROC by refining local representations and suppressing background-dominated responses, which also leads to more spatially consistent anomaly activations. This representation-level correction provides a cleaner and more reliable basis for subsequent image-level scoring.

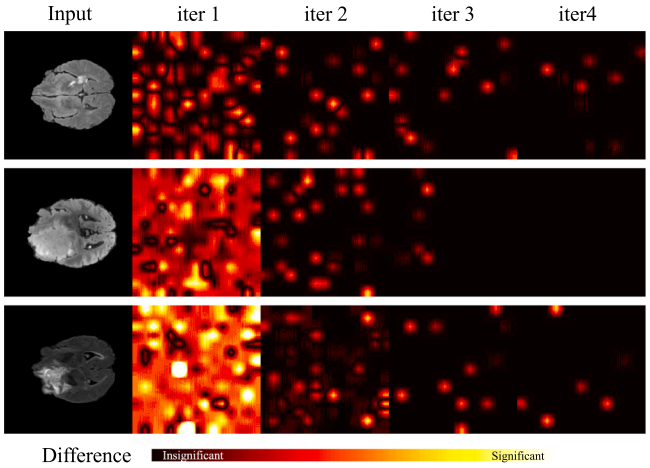


Fig. 9. Effect of CCB module across different numbers of iterations. From left to right: original image, differences after the first iteration, after the second iteration, after the third iteration, and after the fourth iteration. Brighter colors indicate greater differences, showing significant effects after the first iteration with minimal impact from subsequent iterations.

Table 5
Results under different M (AUROC%).

M	Image	Pixel
5	85.98	95.85
10	84.20	95.73
20	**86.70**	**95.73**
25	86.07	95.54
30	85.09	95.59

Table 6
AUROC(%) performance under different $\mathcal{K}$ settings.

$\mathcal{K}$	HIS	ChestXray	OCT17	BrainMRI	LiverCT	RESC	Average
1 3 5	66.91	73.13	98.85	85.62	69.71	91.86	81.01
3 5 7	69.04	73.11	98.69	**87.24**	69.82	92.92	81.80
5 7 9	69.04	73.2	98.61	87.04	69.83	**93.05**	81.80
7 9 11	69.04	74.64	98.57	86.94	69.85	92.93	78.68
9 11 13	**69.06**	74.39	98.58	86.76	69.87	92.8	81.91
11 13 15	**69.06**	**75.38**	**99.01**	86.7	69.89	92.91	**82.16**
13 15 17	69.02	74.27	**99.01**	86.61	**69.90**	92.74	81.93
15 17 19	68.97	74.47	98.54	86.39	**69.90**	92.6	81.81

Table 7
AUROC(%) on BrainMRI under different γ settings.

γ	Image	Pixel
0.05	86.53	95.67
0.1	86.7	95.73
0.2	86.69	95.71
0.5	85.57	94.71

Building on these refined features, TSC further improves image-level AUROC by calibrating anomaly scores at the decision level, reducing the influence of isolated high responses and stabilizing global abnormality assessment. When combined, CCB and TSC consistently enhance both localization and classification performance, demonstrating their synergistic effect.

Incorporating global information yields the best overall results, suggesting that integrating local refinement, score calibration, and global context is critical for few-shot medical anomaly detection. The detailed architectural configurations for each ablation setting are provided in Figs. S-4–S-7 of the Supplementary Material.

Moreover, we analyze the impact of the number of centroids M in the CCB module on BrainMRI. As shown in Table 5, moderate values of M achieve the best trade-off between representation capacity and stability, with $M = 20$ yielding the highest image-level AUROC while maintaining strong pixel-level performance. Too few centroids limit tissue coverage, whereas excessive centroids introduce redundancy without further benefit.

Fig. 9 visualizes feature changes across CCB iterations. Most representation refinement occurs in the first iteration, while subsequent iterations produce marginal differences. This observation justifies using a single iteration for efficiency without compromising performance.

In Table 6, we analyze the effect of neighborhood size $\mathcal{K}$ in the TSC module. Intermediate values achieve the best average performance, indicating a balance between local consistency and global propagation, while extremely small or large neighborhoods degrade robustness across datasets.

Table 7 reports the sensitivity of CCB to the assignment temperature γ. Performance remains stable over a wide range, with only minor fluctuations for $\gamma \geq 0.2$, indicating robust hard assignment behavior, where larger values overly soften assignments and reduce centroid discrimination.

Fig. 10 shows that performance consistently improves as more training samples are provided. Fig. 10(a) shows the results of the whole model, where performance steadily increases with more shots. Fig. 10(b) presents the results without the CCB module, where performance also improves as the number of shots increases, but the enhancement is less pronounced. We adopt the 4-shot setting as a balance between data efficiency and detection accuracy, as it provides strong performance while minimizing the need for excessive labeled samples.

To examine whether the proposed modules depend on a specific backbone architecture, we conduct a backbone analysis using three representative feature extractors: ResNet50, ViT-B/16, and CLIP. For each backbone, we evaluate the base model as well as its variants with CCB, TSC, and their combination, under the same experimental protocol.

As shown in Table 8, both CCB and TSC consistently improve performance across different backbone choices. Notably, the gains brought by CCB mainly manifest at the pixel level, indicating its effectiveness in refining local representations, while TSC primarily improves image-level performance by calibrating global anomaly scores. When combined, CCB and TSC yield complementary improvements across all backbones, with the most significant gains observed on CLIP-based features, reflecting the strong alignment between vision and language representations in the medical anomaly detection task.

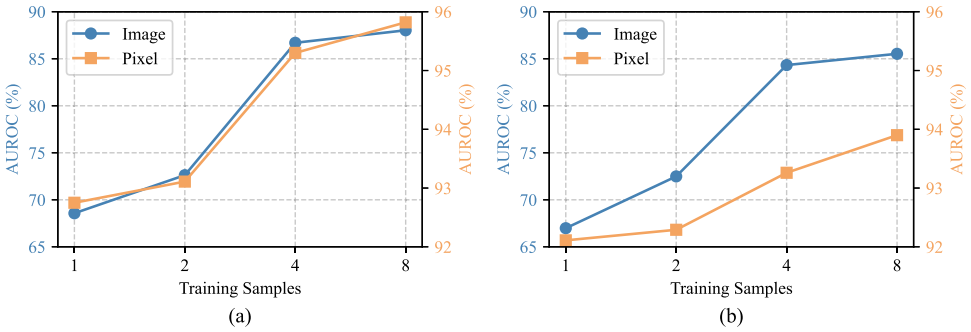


Fig. 10. Impact of training set size on performance: (a) the whole model and (b) model without CCB.

Table 8
Backbone analysis under different module configurations. Results are reported in terms of AUROC (%) on BrainMRI dataset.

Backbone	Base		+CCB		+TSC		+CCB+TSC	
	Image	Pixel	Image	Pixel	Image	Pixel	Image	Pixel
ResNet50	51.6	55.55	53.28	56.52	53.6	55.55	53.79	56.52
ViT-B/16	75.42	90.91	76.64	91.99	77.45	90.91	79.24	91.99
CLIP	79.33	92.13	79.54	94.54	84.23	92.13	85.36	94.54

Table 9
Performance on texture-oriented subsets of industrial anomaly detection benchmarks (image-level AUROC %).

Dataset	Category	Base	+CCB	+TSC	+CCB +TSC
MVTec AD	Grid	99.50	99.83	99.83	99.32
	Leather	100.0	100.0	100.0	100.0
	Carpet	100.0	100.0	100.0	100.0
	Wood	95.96	96.31	96.44	97.02
	Tile	100.0	100.0	100.0	100.0
VisA	Chewinggum	97.06	97.58	97.68	98.08
	Fryum	91.22	92.10	92.80	92.92

5.5. Generalization to texture-dominated industrial anomaly detection

To further clarify the applicability boundary of the proposed framework, we additionally evaluate our method on texture-oriented subsets of industrial anomaly detection benchmarks, including MVTec AD and VisA, as reported in Table 9. These subsets are selected because their anomaly patterns are primarily characterized by subtle texture irregularities, which are visually closer to medical imaging anomalies than geometry-driven industrial defects.

Across both benchmarks, we observe consistent performance improvements brought by CCB and TSC over the base configuration, while the combined model generally achieves the best results. This indicates that the proposed feature refinement and score calibration mechanisms are effective when anomalies manifest as distributed texture deviations rather than explicit structural defects.

Qualitative heatmap visualizations for these subsets are provided in S-3 in the Supplementary Material for further inspection, which serve as auxiliary evidence showing that the proposed framework extends naturally to texture-dominated scenarios, which aligns with the medical-focused motivation of this work.

6. Conclusion

In this work, we present a CLIP-based few-shot medical image anomaly detection framework that addresses two common failure modes under limited supervision: background-driven false positives at the representation level and unstable image-level anomaly scores caused by localized noise. By introducing the CCB module for feature refinement and the TSC module for score correction, the proposed method achieves strong and consistent performance across multiple medical imaging modalities. A key strength of the framework lies in the explicit separation between feature correction and score calibration. CCB enhances subtle pathological cues through global contextual consensus, while TSC stabilizes image-level decisions in a training-free manner. This design yields robust few-shot performance and remains stable across different backbones and hyper-parameter choices.

However, the framework still faces limitations, including occasional false positives from atypical anatomical structures, reduced sensitivity to extremely low-contrast lesions, and a mild dependence on test-set availability due to transductive calibration. Future work will explore incorporating lightweight anatomical priors and extending test-time calibration toward more flexible inference settings.

Overall, this work provides a practical and modular solution for few-shot medical image anomaly detection, offering reusable components for future medical AI research.

CRediT authorship contribution statement

Zihan Nie: Writing – original draft, Visualization, Validation, Software, Methodology; **Muhao Xu:** Writing – review & editing, Software, Methodology, Conceptualization; **Yuan Cui:** Writing – review & editing, Supervision, Funding acquisition; **Hua Wei:** Writing – review & editing, Supervision, Funding acquisition; **Wei Yi:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization; **Sijie Niu:** Writing – review & editing, Supervision; **Yi Wan:** Writing – review & editing, Supervision, Funding acquisition; **Xunbin Wei:** Writing – review & editing, Supervision; **Weiye Song:** Writing – review & editing, Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding sources

This work was supported by the Youth Project of Natural Science Foundation of Shandong Province, China [grant number ZR2023QC262]; the Natural Science Foundation of Shandong Province [grant number ZR2022QF017]; the Shandong Province Outstanding Youth Science Fund Project (Overseas) [grant number 2023HWYQ-023]; the National Natural Science Foundation of China [grant number 62205181]; the Taishan Scholar Foundation of Shandong Province [grant number tsqn202211038]; and the Key Technology Research and Development Program of Shandong Province [grant number 2024CXGC010106].

Acknowledgements

The authors have no acknowledgements to report.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.patcog.2026.113261.

References

[1] Z. Nie, M. Xu, et al., A review of application of deep learning in endoscopic image processing, J. Imaging 10 (11) (2024) 275.
[2] M. Salehi, N. Sadjadi, S. Baselizadeh, et al., Multiresolution knowledge distillation for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14902–14912.
[3] A.B. Nassif, M.A. Talib, et al., Machine learning for anomaly detection: a systematic review, IEEE Access 9 (2021) 78658–78700.
[4] U.A. Usmani, et al., A review of unsupervised machine learning frameworks for anomaly detection in industrial applications, in: K. Arai (Ed.), Intelligent Computing, Springer International Publishing, Cham, 2022, pp. 158–189.
[5] X. Xia, et al., GAN-based anomaly detection: a review, Neurocomputing 493 (2022) 497–535.
[6] Z. Wang, M. Li, R. Xu, L. Zhou, et al., Language models with image descriptors are strong few-shot video-language learners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, 35, Curran Associates, Inc., 2022, pp. 8483–8497.
[7] X. Li, Z. Zhang, X. Tan, et al., PromptAD: learning prompts with only normal samples for few-shot anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16838–16848.
[8] A. Radford, J.W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.
[9] X. Li, Z. Huang, F. Xue, Y. Zhou, MUSC: zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images, in: The Twelfth International Conference on Learning Representations, 2024.
[10] X. Xu, Y. Cao, H. Zhang, N. Sang, X. Huang, Customizing visual-language foundation models for multi-modal anomaly detection and reasoning, 2025. https://arxiv.org/abs/2403.11083.
[11] Y. Cao, X. Xu, Y. Cheng, C. Sun, Z. Du, L. Gao, W. Shen, Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection, IEEE Trans. Cybern. 55 (4) (2025) 1917–1929. https://doi.org/10.1109/TCYB.2025.3536165
[12] Y. Cao, X. Xu, J. Zhang, Y. Cheng, X. Huang, G. Pang, W. Shen, A survey on visual anomaly detection: challenge, approach, and prospect, (2024). arXiv preprint arXiv:2401.16402.
[13] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14318–14328.
[14] H. Park, J. Noh, B. Ham, Learning memory-guided normality for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[15] F.V. Massoli, F. Falchi, A. Kantarci, Ş. Akti, H.K. Ekenel, G. Amato, MOCCA: multilayer one-class classification for anomaly detection, IEEE Trans. Neural Netw. Learn. Syst. 33 (6) (2021) 2313–2323.
[16] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, C. Steger, Improving unsupervised defect segmentation by applying structural similarity to autoencoders, (2018). arXiv preprint arXiv:1807.02011.
[17] X. Yan, H. Zhang, X. Xu, X. Hu, P.-A. Heng, Learning semantic context from normal samples for unsupervised anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 35, 2021, pp. 3110–3118.
[18] Z. You, K. Yang, W. Luo, L. Cui, Y. Zheng, X. Le, ADTR: anomaly detection transformer with feature reconstruction, in: International Conference on Neural Information Processing, Springer, 2022, pp. 298–310.
[19] S. Lu, W. Zhang, H. Zhao, H. Liu, N. Wang, H. Li, Anomaly detection for medical images using heterogeneous auto-encoder, IEEE Trans. Image Process. 33 (2024) 2770–2782.
[20] H. Deng, 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.
[21] Y. Cao, Q. Wan, W. Shen, L. Gao, Informative knowledge distillation for image anomaly segmentation, Knowl. Based Syst. 248 (2022) 108846.
[22] G. Wang, Y. Zou, S. He, Y. Wang, R. Dai, Anomaly detection and localization via reverse distillation with latent anomaly suppression, IEEE Trans. Circuits Syst. Video Technol. 35 (2025) 9592–9607.
[23] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PmLR, 2021, pp. 8748–8763.
[24] Y. Li, H. Wang, Y. Duan, X. Li, Clip surgery for better explainability with enhancement in open-vocabulary tasks, arXiv e-prints (2023) 2304.
[25] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, G. Boracchi, AdaCLIP: adapting clip with hybrid learnable prompts for zero-shot anomaly detection, in: European Conference on Computer Vision, Springer, 2024, pp. 55–72.
[26] A. Li, C. Qiu, M. Kloft, P. Smyth, M. Rudolph, S. Mandt, Zero-shot anomaly detection via batch normalization, Adv. Neural Inf. Process. Syst. 36 (2023) 40963–40993.
[27] J. Bao, H. Sun, H. Deng, Y. He, Z. Zhang, X. Li, BMAD: benchmarks for medical anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4042–4053.
[28] U. Baid, S. Ghodasara, M. Bilello, et al., The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021. 2107.02314
[29] S. Bakas, H. Akbari, et al., Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, Sci. Data 4 (1) (2017) 170117.
[30] B.H. Menze, A. Jakab, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2015) 1993–2024.
[31] B. Landman, Z. Xu, et al., Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault-Workshop Challenge, 5, 2015, p. 12.
[32] P. Bilic, et al., The liver tumor segmentation benchmark (liTS), Med. Image Anal. 84 (2023) 102680.
[33] J. Hu, Y. Chen, et al., Automated segmentation of macular edema in OCT using deep neural networks, Med. Image Anal. 55 (2019) 216–227.
[34] D.S. Kermany, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.e9.
[35] X. Wang, Y. Peng, L. Lu, et al., ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[36] B. Ehteshami Bejnordi, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, JAMA 318 (22) (2017) 2199–2210.
[37] Z. Wang, Z. Wu, D. Agarwal, et al., MedCLIP: contrastive learning from unpaired medical images and text, 2022. 2210.10163
[38] J. Jeong, Y. Zou, T. Kim, et al., WinCLIP: zero-/few-shot anomaly classification and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19606–19616.
[39] Z. Liu, Y. Zhou, Y. Xu, 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.
[40] Y. Zhou, X. Xu, J. Song, F. Shen, H.T. Shen, MSFlow: multiscale flow-based framework for unsupervised anomaly detection, IEEE Trans. Neural Netw. Learn. Syst. 36 (2) (2025) 2437–2450. https://doi.org/10.1109/TNNLS.2023.3344118
[41] X. Liu, J. Wang, B. Leng, S. Zhang, Unlocking the potential of reverse distillation for anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 5640–5648.
[42] M. Xu, C. Zhu, G. Feng, S. Niu, Multitask hybrid knowledge distillation for unsupervised anomaly detection, IEEE Trans. Ind. Inf. 21(2025) 5666–5676 .
[43] J. Zhu, G. Pang, Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17826–17836.
[44] W. Luo, Y. Cao, H. Yao, X. Zhang, J. Lou, Y. Cheng, W. Shen, W. Yu, Exploring intrinsic normal prototypes within a single image for universal anomaly detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 9974–9983.
[45] D. Gudovskiy, S. Ishizaka, K. Kozuka, CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 98–107.
[46] H. Deng, X. Li, Anomaly detection via reverse distillation from one-class embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9737–9746.
[47] J. Guo, S. Lu, W. Zhang, F. Chen, H. Li, H. Liao, Dinomaly: the less is more philosophy in multi-class unsupervised anomaly detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 20405–20415.
[48] W. Luo, H. Yao, Z. Qiang, X. Zhang, W. Zhang, A feature shuffling and restoration strategy for universal unsupervised anomaly detection, Knowl. Based Syst. 332 (2025) 114874.
[49] P. Tang, X. Yan, X. Hu, K. Wu, T. Lasser, K. Shi, Anomaly detection in medical images using encoder-attention-2decoders reconstruction, IEEE Trans. Med. Imaging 44 (2025) 3370–3382.
[50] J. Guo, S. Lu, L. Jia, W. Zhang, H. Li, Encoder-decoder contrast for unsupervised anomaly detection in medical images, IEEE Trans. Med. Imaging 43 (3) (2023) 1102–1112.
[51] C. Huang, A. Jiang, J. Feng, Y. Zhang, X. Wang, Y. Wang, Adapting visual-language models for generalizable anomaly detection in medical images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11375–11385.
[52] M. Shiri, C. Beyan, V. Murino, MadCLIP: few-shot medical anomaly detection with CLIP, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 416–426.
[53] H. Deng, Z. Zhang, J. Bao, X. Li, AnoCLIP: text-guided zero-shot anomaly localization via self-supervised adaptation, in: 2025 IEEE International Conference on Multimedia and Expo (ICME), 2025, pp. 1–6. https://doi.org/10.1109/ICME59968.2025.11209270
[54] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, Laion-400m: open dataset of clip-filtered 400 million image-text pairs, (2021). arXiv preprint arXiv:2111.02114.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

Zihan Nie received the BS degree in the School of Mechanical, Electrical and Information Engineering, Shandong University, Shandong, China in 2024. She is a Master's degree candidate in School of Mechanical Engineering, Shandong University, China. Her research interests include artificial intelligence, deep learning, medical image processing and anomaly detection.

Muhao Xu received the MS degree in School of Information Science and Engineering, University of Jinan, Shandong, China in 2024. He is currently a PhD degree candidate at the School of Mechanical Engineering, Shandong University, Shandong, China. His research interests include medical image processing, anomaly detection and multimodal information mining.

Yuan Cui holds a BE from Jilin University, an MS from ETH Zurich, and a PhD from Uppsala University. She is currently a postdoctoral researcher at Shandong University. Her work mainly focuses on microscale measurements for biological applications using optical and microfluidic methods.

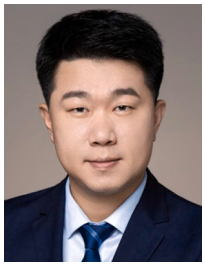
Hua Wei received the BSc degrees from Shandong University in 2016, and the PhD degree from the University of Chinese Academy of Sciences in 2022, and engaged in post doctoral research in Institute of Software Chinese Academy of Sciences from 2022 to 2024. She is currently a laboratory technician at the School of Mechanical Engineering, Shandong University. Her research interests include computational optical imaging and image quality enhancement techniques, development of optical coherence tomography equipment, and intelligent analysis of image data.

Wei Yi received the PhD degree in Mechanical Engineering and Automation from Shandong University, Jinan, China. He completed his postdoctoral research at Boston University, USA, where he served as a visiting scholar from 2018 to 2020. His work focused on the development and application of optomechanical systems for optical coherence tomography. He is currently Senior Experimentalist at Shandong University, with research interests in optical imaging and interdisciplinary applications of mechanical engineering.

Sijie Niu received BS and PhD Degrees from the school of Computer science at Liaocheng University and Nanjing University of Science and Technology in 2007 and 2016, respectively. He was a visiting scholar at Stanford University in 2014. Now he is a Post-doctoral with medical image analysis, UNC. He is currently a professor in the school of Information Science and Engineering, University of Jinan, China. His research interests include Pattern recognition, machine learning, image processing, and medical image analysis.

Yi Wan received the PhD degree from Shandong University, Jinan, China, in 2006. Since 2015, He is the Dean and Doctoral Supervisor of the School of Mechanical Engineering at Shandong University. His main research interests include Deep Learning, 3D printing, and control of robot. He is a project letter evaluation expert from the National Natural Science Foundation of China, as well as a peer reviewer for domestic and foreign journals such as International Journal of Advanced Manufacturing Technology, Materials Science Engineering C, Mechanical Systems and Signal Processing, Journal of South China University of Technology Natural Edition, and Journal of Southern Airlines.

Xunbin Wei received his BS degree from the University of Science and Technology of China in 1993, and his PhD degree from the University of California between 1993 and 1999. He was a postdoctoral fellow at Harvard Medical School from 1999 to 2001. He is currently a tenured professor at Peking University since 2019. His research interests include in vivo flow cytometry for early tumor detection, in vivo optical molecular imaging and near-infrared nano-optical probes, in vivo optical cell manipulation techniques, and laser medicine and phototherapy for Alzheimer's disease.



Weiye Song received the PhD degree from JiLin University, Changchun, China, in 2010. He is also a postdoctoral fellow at Boston University/Boston Medical Center and Harvard University/Massachusetts General Hospital. He is currently a professor at the School of Mechanical Engineering, Shandong University. His main research areas are opto electromechanical integrated medical equipment based on weak coherent optical imaging, spectroscopic analysis and other technologies, endoscopes, early medical diagnosis, intraoperative navigation equipment, etc.