This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# PDD: Manifold-Prior Diverse Distillation for Medical Anomaly Detection

Xijun Lu[1], Hongying Liu[1,4*], Fanhua Shang[2,4*], Yanming Hui[2], Liang Wan[1,3,4]
[1]Medical School, Tianjin University
[2]School of Computer Science and Technology, Tianjin University
[3]School of Computer Software, Tianjin University
[4]The Key Research Center for Surface Monitoring and Analysis of Relics, SMARC
{xjunlu, hyliu2009, fhshang, ymhui, lwan}@tju.edu.cn

## Abstract

*Medical image anomaly detection faces unique challenges due to subtle, heterogeneous anomalies embedded in complex anatomical structures. Through systematic Grad-CAM analysis, we reveal that discriminative activation maps fail on medical data, unlike their success on industrial datasets, motivating the need for manifold-level modeling. We propose **PDD** (Manifold-Prior Diverse Distillation), a novel framework that unifies dual-teacher priors into a shared high-dimensional manifold and distills this knowledge into dual students with complementary behaviors. Specifically, frozen VMamba-Tiny and wide-ResNet50 encoders provide global contextual and local structural priors, respectively. Their features are unified through a **Manifold Matching and Unification (MMU)** module, while an **Inter-Level Feature Adaption (InA)** module enriches intermediate representations. The unified manifold is distilled into two students: one performs layer-wise distillation via **InA** for local consistency, while the other receives skip-projected representations through a **Manifold Prior Affine (MPA)** module to capture cross-layer dependencies. A diversity loss prevents representation collapse while maintaining detection sensitivity. Extensive experiments on multiple medical datasets demonstrate that **PDD** significantly outperforms existing state-of-the-art methods, achieving improvements of up to 11.8%, 2.9%, and 8.5% in terms of AUROC on HeadCT, BrainMRI, and ZhangLab datasets, respectively, and 3.4% in terms of F1 max on the Uni-Medical dataset, establishing new state-of-the-art performance in medical image anomaly detection. The implementation will be released at https://github.com/OxygenLu/PDD.*

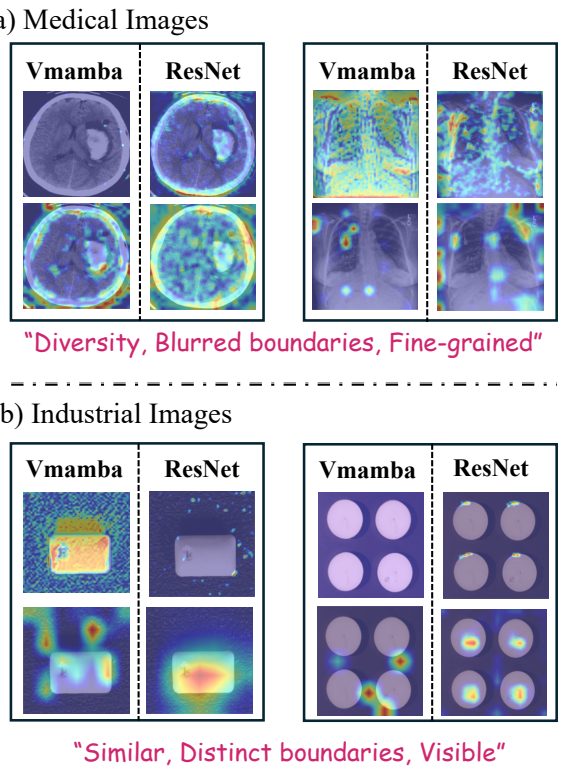



Figure 1. **Grad-CAM visualization of frozen Vmamba and ResNet across medical and industrial images.** Within each group, feature maps progress from low-dimensional to high-dimensional feature representations (top to bottom). At the same feature dimension, the dispersed and aggregated activation patterns of the two backbones form complementary prior information.

## 1. Introduction

Medical image anomaly detection has attracted increasing attention as a crucial component for early disease screening and computer-aided diagnosis [4, 20, 23, 47]. Unlike con-

*Corresponding authors

ventional anomaly detection on natural or industrial images, medical anomalies often exhibit subtle, heterogeneous, and weakly contrasted patterns embedded in highly complex anatomical structures [16, 21]. These characteristics make unsupervised anomaly detection (UAD) particularly challenging, since models must learn a compact and stable manifold of normal anatomy solely from healthy samples, without access to annotated abnormal regions [2].

Recent teacher–student frameworks have shown promising performance in anomaly detection, especially in industrial inspection tasks [3, 6, 8, 37]. However, as shown in Fig. 1, Grad-CAM visualizations reveal a striking contrast: on industrial datasets such as MVTec, Grad-CAM produces clean and highly localized heatmaps, while on medical datasets (e.g., BrainMRI, HeadCT), the heatmaps become diffuse, noisy, and anatomically inconsistent [4, 20]. This observation reveals a fundamental difference: industrial defects are typically texture-driven and spatially localized, while medical anomalies are structural deviations distributed across anatomical hierarchies with subtle, context-dependent boundaries [8, 10, 14, 16]. As a result, single-stream feature extractors are insufficient to learn a complete and anatomically coherent normal manifold.

These findings motivate the need for a principled approach that can integrate multiple complementary priors to learn a robust and unified normal manifold for medical images [7, 26, 41, 42]. CNNs excel at capturing fine-grained local textures [5, 11, 25, 29, 33], whereas sequence models such as Mamba capture long-range dependencies and global structural patterns [12, 15, 24, 27, 28, 36, 40]. However, directly fusing their features does not guarantee alignment between their heterogeneous manifolds, nor does it ensure that downstream student networks preserve representation diversity—an important property for reliably detecting subtle abnormal deviations [12, 44, 46].

To address these challenges, we propose a novel Manifold-Prior Diverse Distillation (**PDD**) framework that unifies dual-teacher priors into a shared high-dimensional manifold and distills this unified knowledge into dual students with complementary learning behaviors. Specifically, a frozen VMamba-Tiny encoder and a frozen wide-ResNet50 encoder provide global contextual and local structural priors, respectively. Their multi-scale features are aligned through the **Manifold Matching and Unification (MMU)** module, while the **Inter-Level Feature Adaption (InA)** module fuses intermediate features from both teachers. This unified manifold is then distilled into two structurally identical but functionally diverse students: Student 1 performs layer-wise distillation from the fused teacher features via **InA**, focusing on local consistency, while Student 2 receives skip-projected latent representations from the unified manifold through the **Manifold Prior Affine (MPA)** module, enabling it to capture cross-layer contextual dependencies. A diversity loss ensures agreement on normal structures while allowing diversity in detecting potential abnormalities.

Our contributions can be summarized as follows:

- We propose a novel dual-teacher architecture that leverages complementary representations from heterogeneous backbones (VMamba-Tiny for global contextual priors and wide-ResNet50 for local structural priors), addressing the limitations of single-stream feature extractors in medical anomaly detection.
- We introduce a dual-teacher manifold matching and unification module (**MMU**) that integrates global contextual priors from VMamba-Tiny and local structural priors from wide-ResNet50 into a cohesive high-dimensional anatomical manifold.
- We propose a dual-student diverse distillation strategy that enhances representation stability by combining local distillation (via **InA**), cross-layer manifold projection (via **MPA**), and dual-student consistency regularization, achieving significant improvements over existing methods on multiple medical datasets.

## 2. Related Work

### 2.1. Visual anomaly detection

Visual anomaly detection can be framed as identifying deviations from learned normal sample patterns. In recent years, numerous effective, domain-specific anomaly detection models have emerged in both the medical and industrial fields. CutPaste [22], a method developed for industrial image analysis, introduced an auxiliary task (image cropping and rotation) to simulate anomalies, pioneering the application of self-supervised learning in anomaly detection. For instance, f-AnoGAN [35], which is based on adversarial generative networks, identifies anomalies using discriminators operating on feature vectors in latent space; however, its iterative optimization process is inefficient, and the generator is prone to mode collapse.

In addition, recent efforts have focused on developing unified models for anomaly detection. For example, UniAD [44] employs a patch-based reconstruction method optimized for industrial applications; however, its performance degrades in other domains and requires a large number of normal samples for training. UniVAD [9] utilizes pre-trained Grounding DINO [31] and SAM [19], achieving training-free anomaly detection and demonstrating satisfactory performance on cross-domain datasets.

In the field of medical image analysis, SQUID [41] and SimSID [42], proposed by Tiange Xiang et al., represent successive works. These models incorporate a memory storage module and transformer positional encoding, proving effective for medical images with fixed structures, such as X-rays, but ineffective at decoupling anatomical struc-

tural information during dimensionality reduction processes like CT (3D to 2D).

### 2.2. Methods Based on Knowledge Distillation

In recent years, knowledge distillation based anomaly detection models have emerged as a promising method for balancing computational efficiency with model performance. The traditional knowledge distillation (KD) paradigm involves transferring knowledge from a teacher network to a student network for the identification and localization of anomalous samples. Typically, teacher and student networks adopt similar or identical architectures, resulting in insufficient representation diversity when processing out-of-distribution samples. RD4AD [6] pioneered a new paradigm termed reverse knowledge distillation, in which the student network learns directly from the teacher's high-level features, progressively capturing multi-scale information across the teacher's layers to enable the efficient separation of normal from anomalous samples. Inspired by RD4AD [6], Skip-TS [30] employs skip connections and multi-scale feature sampling between the teacher and student networks, thereby mitigating the challenge of normal pattern reconstruction prevalent in medical datasets—a limitation identified in RD4AD. More recently, CFRG [43] has been introduced, leveraging heterogeneous teacher-student networks to extract multi-scale features, utilizing recovery networks to reconstruct normal patterns from anomalous features, and enabling fine-grained anomaly localization through segmentation networks. Our method introduces a novel reverse knowledge distillation framework, aiming to enhance the consistency of the student networks to achieve more effective separation of normal from anomalous patterns.

Some studies have found that pre-trained networks on large datasets such as ImageNet-1k can serve as backbones for anomaly detection. Models like SAM, Mamba, and DINO leverage their ability to extract features across images, enabling direct anomaly detection on samples. However, this computational efficiency comes at the cost of reduced feature extraction capability. UniVAD, as previously mentioned, requires explicit mask labeling for its target objects. Consequently, its performance on datasets lacking such annotations or characterized by fine-grained normal regions—such as BrainMRI [34] and CheXray [39]—exhibits a significant gap compared to domain-specific models.

However, existing methods typically employ single-teacher architectures and focus on homogeneous feature spaces, limiting their ability to capture diverse representations of normal patterns. Our work addresses these limitations by introducing a dual-teacher dual-student framework that leverages heterogeneous backbones and unified manifold spaces to enable more diverse and robust normal pattern learning.

## 3. Methodology

### 3.1. Manifold-Unified Reverse Distillation

As shown in Fig. 2, we propose **PDD**, a manifold-unified reverse distillation framework that employs dual-teacher and dual-student architecture to leverage complementary representations from heterogeneous backbones. The teachers consist of frozen VMamba-Tiny and frozen wide-ResNet50 encoders, which process input images in parallel and extract multi-scale intermediate features. These features are fused via the **InA** module (Fig. 2-(b)) to obtain enriched fused features $f_b^i$. The two teacher encoders compress input images into distinct high-dimensional manifold spaces: VMamba-Tiny learns sequential state-space representations while wide-ResNet50 captures spatial convolutional features. The **MMU** module (Fig. 2-(c)) performs geometric alignment between these heterogeneous manifolds, mapping semantically similar features to a unified manifold space. The aligned features are fed into two student networks: Student 1 distills features from **InA** via $\mathcal{F}_{E_u}^i$ for unified representations, while Student 2 incorporates multi-scale manifold features through MLP-based skip connections to $\mathcal{F}_{E_p}^i$, distilling both prior knowledge from the unified manifold and enriched features from **InA**. This dual-student design enables diverse reconstruction of normal samples while maintaining sensitivity to anomalies.

### 3.2. Intra-Backbone Feature Fusion Strategy

We employ two heterogeneous backbone networks with large-scale pre-trained weights as teacher models: wide-ResNet50 and vanilla-VMamba-Tiny (VMamba-Tiny), denoted as Teacher 1 and Teacher 2, respectively. Both backbones have demonstrated strong performance on ImageNet-1K, and their pre-trained weights exhibit powerful image feature extraction capabilities. As shown in Fig. 1, Grad-CAM visualizations reveal distinct feature representations at shallow and deep layers for both ResNet and Mamba backbones. The complementary nature of these representations suggests that fusing shallow features from wide-ResNet50 and VMamba-Tiny can significantly enrich the prior knowledge at each network layer, thereby obtaining a more powerful feature extractor.

To this end, we design a lightweight inter-backbone feature fusion adapter, termed InA (Inter-Level Feature Adaption). Formally, we define the input image as $x_n \in \mathbb{R}^{C\times H\times W}$, which is reshaped and fed into the two teacher encoders. The VMamba backbone, as shown in Fig. 2, consists of 4 blocks denoted as $\mathcal{F}_{D_m}^i$ for $i \in [1, 4]$. The frozen vanilla-VMamba-Tiny pre-trained weights output a feature map $f_m^i \in \mathbb{R}^{h\times w\times c}$ between each layer, where $h$, $w$, and $c$ represent the height, width, and number of channels, respectively. Similarly, the wide-ResNet50 branch consists of 4 blocks, denoted as $\mathcal{F}_{D_c}^i$ for $i \in [1, 4]$, which output fea-

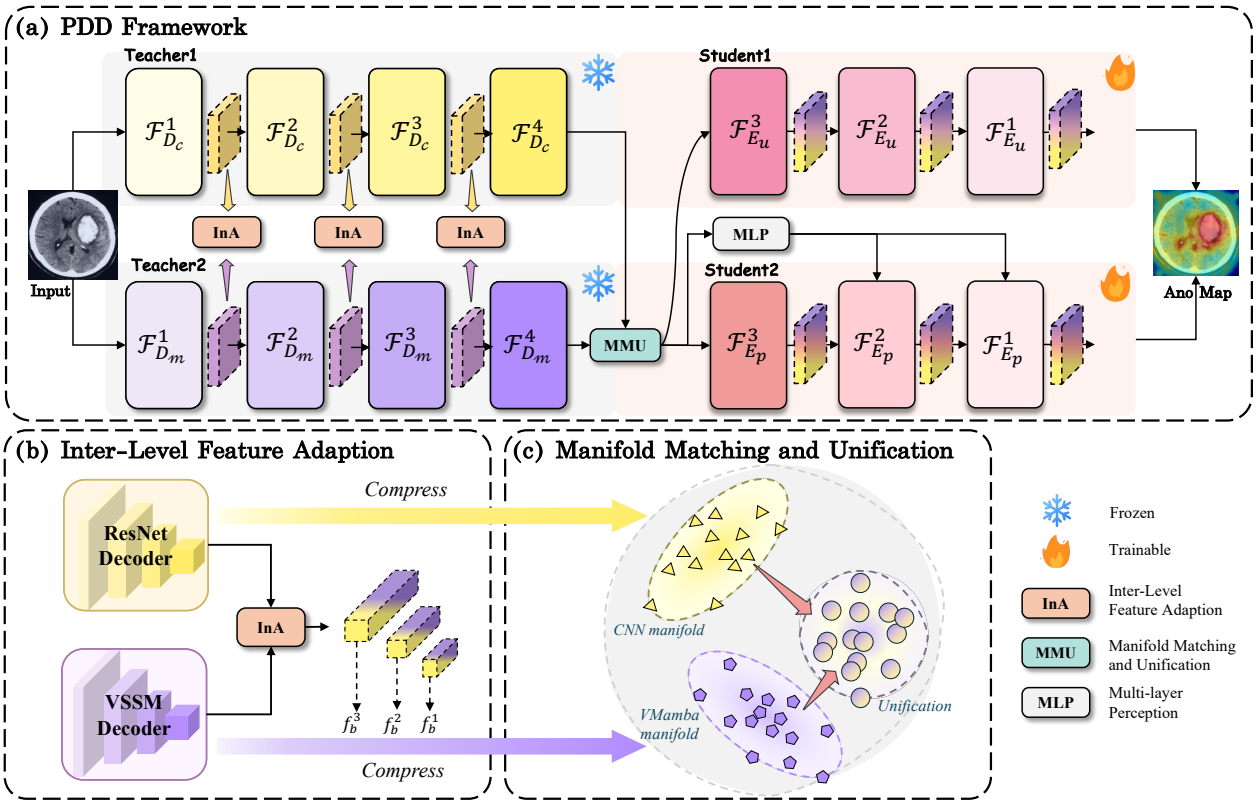



Figure 2. **Overview of the proposed PDD framework.** The framework employs a dual-teacher and dual-student architecture. The teachers consist of frozen VMamba-Tiny and frozen wide-ResNet50 encoders, whose intermediate features are fused via the InA module (shown in (b)) to obtain $f_b^i$. The two teacher encoders compress input images into distinct high-dimensional manifold spaces, which are then aligned through the MMU module. The aligned features are fed into two student networks: Student 1 distills features from InA via $\mathcal{F}_{E_u}^i$, while Student 2 incorporates multi-scale manifold space features from the unified manifold through MLP-based skip connections to $\mathcal{F}_{E_p}^i$, distilling both prior and InA features. This enables diverse reconstruction of normal samples and effective separation of anomalies.

tures $f_c^i \in \mathbb{R}^{h' \times w' \times c'}$ between layers.

The shallow feature fusion process between the two backbones is performed as follows:

$$S(f_m^i, f_c^i) = \left(\frac{h'}{h}, \frac{w'}{w}\right), \quad \tilde{f}_m^i = \mathcal{U}(f_m^i, S), \tag{1}$$

$$f_b^i = \mathcal{F}_{\text{InA}}\left(f_m^i, f_c^i\right) = \tilde{f}_m^i + f_c^i, \tag{2}$$

where $S(\cdot)$ is a scaling function that computes the spatial scaling factors, $\mathcal{U}(\cdot)$ is an upsampling function that performs 2D bilinear interpolation to scale the Mamba feature map to match the convolutional feature map dimensions, and $\mathcal{F}_{\text{InA}}(\cdot)$ implements the feature fusion at shallow layers, producing the fused feature $f_b^i$.

### 3.3. Manifold Space Unification of Heterogeneous Backbones

The input image $x_n$ is processed by the two teacher encoders in parallel, and through their respective encoding processes, the image is compressed into high-dimensional manifold spaces, as illustrated in Fig. 2-(c). Due to their distinct architectural inductive biases, the two teacher encoders naturally embed input images into different high-dimensional manifold spaces. The VMamba-Tiny encoder, with its state-space modeling mechanism, learns representations in a sequential state-space manifold $\mathcal{M}_m \subset \mathbb{R}^{d_m}$, while the wide-ResNet50 encoder, with its convolutional structure, captures features in a spatial convolutional manifold $\mathcal{M}_c \subset \mathbb{R}^{d_c}$, where $d_m$ and $d_c$ denote the dimensions of the respective manifolds. These heterogeneous manifolds encode complementary semantic information but exist in different geometric spaces, making direct feature interaction challenging.

The deep features at the tail of the backbones contain rich high-level semantic information that is crucial for understanding the overall image semantics. To fully leverage the high-level semantic features from both teacher networks and unify the heterogeneous manifolds into a common space, we propose the **M**anifold **M**atching and **U**nification (MMU) module. The MMU module is responsible for fus-

ing the high-level semantic features from both teacher models, thereby strengthening the modeling capability for overall image semantic consistency.

Formally, let $f_m^i \in \mathbb{R}^{h\times w\times c}$ and $f_c^i \in \mathbb{R}^{h'\times w'\times c'}$ denote the tail features extracted from the VMamba and ResNet encoders at layer $i$, respectively. The MMU module first adapts the Mamba features through a channel-wise adaptation pathway. The MMU consists of a $1\times1$ dilated convolution and a $3\times3$ standard convolution, both activated by the GeLU function and connected in sequence, with a residual connection. The adapted Mamba features are computed as:

$$f_{m^c}^i = \text{Res}\left([C^3 \circ \mathcal{G}(\text{BN}(C^1(f_m^i)))], C^1(f_m^i)\right), \quad (3)$$

where $\mathcal{G}$ denotes the GeLU activation function, $C^k$ represents a convolutional layer with kernel size $k$, BN denotes batch normalization, and $\text{Res}(\cdot,\cdot)$ implements the residual connection by element-wise addition. The $1\times1$ convolution $C^1$ performs channel-wise adaptation, while the $3\times3$ convolution $C^3$ captures spatial context, enabling the module to model both channel-wise relationships and spatial dependencies.

The unified features are then obtained by fusing the adapted Mamba features with the ResNet features:

$$f_t^i = \mathcal{F}_{\text{MMU}}\left(f_m^i, f_c^i\right) = \tilde{f}_{m^c}^i + f_c^i, \quad (4)$$

where $\tilde{f}_{m^c}^i$ denotes the spatially aligned version of $f_{m^c}^i$ to match the spatial dimensions of $f_c^i$ (if necessary), and $\mathcal{F}_{\text{MMU}}(\cdot,\cdot)$ implements the feature fusion. This unified feature $f_t^i$ combines the complementary semantic information from both heterogeneous manifolds, creating a common representation space that preserves the strengths of both architectures while enabling effective feature interaction for the subsequent student networks.

### 3.4. Normal Pattern Representation Diversification

As shown in Fig. 3-(a), to achieve diverse reconstruction of normal patterns while maintaining sensitivity to anomalies, we design a dual-student architecture with three complementary representation strategies. The unified features $f_t^i$ from the MMU module serve as the input to both student networks, enabling them to learn different aspects of normal patterns.

**Representation Strategy 1: Basic Student Network Distillation.** The first student network, denoted as Student 1, receives features from the unified manifold space $f_t^i$ and learns to reconstruct the fused features from the InA module. Formally, let $\mathcal{F}_{E_u}^i$ denote the features extracted from Student 1 at layer $i$. The distillation loss is defined as:

$$\mathcal{L}_{\text{kr}} = \sum_{i=1}^{L} \|f_b^i - \mathcal{F}_{E_u}^i\|_2^2, \quad (5)$$

where $L$ is the number of layers, and $f_b^i$ represents the fused features from the InA module at layer $i$. This MSE-based loss ensures that Student 1 learns to capture the cross-backbone fused knowledge encoded in the InA features.

**Representation Strategy 2: Prior Knowledge-Guided Reconstruction.** The second student network, Student 2, incorporates prior knowledge from the unified manifold space through the Manifold Prior Affine module. The prior knowledge is extracted from the unified manifold features $f_t^i$ via an MLP-based affine transformation:

$$z_p^i = \text{MLP}(f_t^i) = W_p^i \cdot f_t^i + b_p^i, \quad (6)$$

where $W_p^i$ and $b_p^i$ are learnable weight matrix and bias vector, respectively. The transformed prior knowledge $z_p^i$ is then injected into Student 2 at each layer through skip connections, enabling the network to leverage both the prior knowledge from the unified manifold and the enriched features from InA. The optimization objective combines MSE loss and cosine similarity:

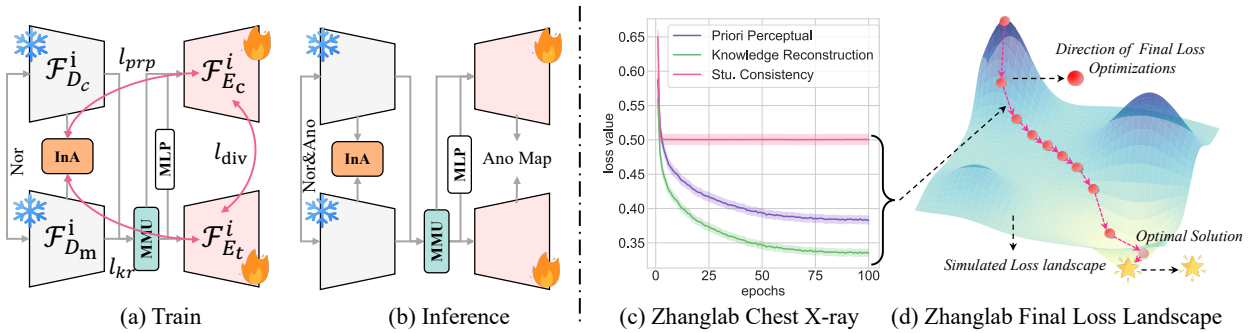
$$\mathcal{L}_{\text{prp}} = \sum_{i=1}^{L}\left[\alpha\|f_b^i - \mathcal{F}_{E_p}^i\|_2^2 + \beta\left(1 - \frac{f_b^i \cdot \mathcal{F}_{E_p}^i}{\|f_b^i\|_2\|\mathcal{F}_{E_p}^i\|_2}\right)\right], \quad (7)$$

where $\mathcal{F}_{E_p}^i$ denotes the features from Student 2 at layer $i$, $\alpha$ and $\beta$ are weighting factors, and the cosine similarity term encourages angular alignment between the InA features and Student 2 features.

**Representation Strategy 3: Student Diversity Optimization.** To prevent both student networks from collapsing into a single representation mode and to enable diverse reconstruction of normal patterns, we introduce a diversity constraint that encourages different representation strategies at different feature scales. Specifically, we apply an inverted cosine similarity constraint that encourages dissimilarity at low-dimensional feature spaces (to capture diversity) while maintaining similarity at high-dimensional feature spaces (to ensure consistency). The diversity loss is defined as:

$$\mathcal{L}_{\text{div}} = \sum_{i=1}^{L_{\text{low}}} \max\left(0, \frac{\mathcal{F}_{E_u}^i \cdot \mathcal{F}_{E_p}^i}{\|\mathcal{F}_{E_u}^i\|_2\|\mathcal{F}_{E_p}^i\|_2} - \tau_{\text{low}}\right) - \sum_{i=L_{\text{low}}+1}^{L} \min\left(0, \frac{\mathcal{F}_{E_u}^i \cdot \mathcal{F}_{E_p}^i}{\|\mathcal{F}_{E_u}^i\|_2\|\mathcal{F}_{E_p}^i\|_2} - \tau_{\text{high}}\right), \quad (8)$$

where $L_{\text{low}}$ denotes the number of low-dimensional layers, $\tau_{\text{low}}$ and $\tau_{\text{high}}$ are threshold parameters. The first term penalizes high cosine similarity in low-dimensional layers (encouraging diversity), while the second term penalizes low cosine similarity in high-dimensional layers (ensuring consistency). This design enables the framework to capture the




(a) Train (b) Inference (c) Zhanglab Chest X-ray (d) Zhanglab Final Loss Landscape

Figure 3. **Visualization of PDD pipeline and loss functions.** Figure (a) illustrates the training process of PDD, which is trained exclusively on normal samples. The training simultaneously optimizes three loss functions: $l_{prp}$, $l_{kr}$, and $l_{div}$. Figure (c) shows the loss function curves over 100 epochs on the ZhangLab Chest-Xray dataset. The combined direction of the three curves is represented in the simulated loss landscape in Figure (d), with arrows indicating the direction of loss descent. In the final 100 epochs, the training converges near the optimal solution, demonstrating that PDD effectively learns features of normal samples and can separate anomalies.

Table 1. Comparison with six baselines on four medical datasets. Evaluation metric: AUROC (%). **Bold**: best; underline: second best.

| **Method**→ **Dataset** ↓ | **f-AnoGAN**[35] MIA'19 | **CutPaste**[22] CVPR'21 | **RD4AD**[6] CVPR'22 | **SQUID**[41] CVPR'23 | **SIMSID**[42] TPAMI'24 | **Skip-TS**[30] TIM'24 | **Ours(PDD)** |
|---|---|---|---|---|---|---|---|
| **HeadCT** | 82.6 | 73.0 | 74.3 | 75.4 | 74.9 | 85.7 | **97.5** |
| **Zhanglab** | 75.5 | 73.3 | 87.5 | 87.6 | 91.1 | 79.2 | **94.0** |
| **BrainMRI** | 77.1 | 67.0 | 80.9 | 74.7 | 81.5 | 88.2 | **96.7** |
| **CheXpert** | 65.8 | 65.5 | 71.9 | 78.1 | **79.7** | 68.7 | 79.1 |

diversity on medical imaging datasets, including variations in structure, density, and imaging protocols, while maintaining the ability to effectively separate anomalous patterns from normal ones.

The overall optimization objective combines the three loss functions with learnable weighting factors:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{kr}}\mathcal{L}_{\text{kr}} + \lambda_{\text{prp}}\mathcal{L}_{\text{prp}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \tag{9}$$

where $\lambda_{\text{kr}}$, $\lambda_{\text{prp}}$, and $\lambda_{\text{div}}$ are the weighting parameters that balance the contributions of knowledge distillation loss, prior knowledge loss, and diversity loss, respectively. These weights are either set as hyperparameters or learned during training to optimize the performance of the framework.

# 4. Experiments

## 4.1. Dataset and Experimental Setup

**Datasets.** We evaluate the performance of our model on multiple medical imaging datasets covering diverse anatomical regions and imaging modalities, including chest X-ray images, brain MRI images, liver CT images, and retinal OCT images.

ZhangLab Chest X-ray [18]. This dataset contains chest X-ray images with healthy (normal) and pneumonia (anomaly) cases. Following the experimental protocol , we use 1,249 healthy images from the official training set for model training, and evaluate on a test set consisting of 234 normal and 390 abnormal images.

CheXpert [17]. This large-scale chest X-ray dataset provides a comprehensive benchmark for automated chest X-ray interpretation. We train our model on 4,499 healthy images from the official training set and test on a balanced test set containing 250 normal and 250 abnormal images.

HeadCT [38]. This dataset consists of 200 CT images for brain hemorrhage detection. The training set contains 60 normal images, while the test set comprises 140 images including 40 normal samples and 100 abnormal brain hemorrhage samples.

BrainMRI [34]. This dataset contains 252 MRI images for brain tumor detection. The training set includes 58 normal images, and the test set consists of 194 images with 40 normal samples and 154 abnormal samples.

Uni-Medical [45]. This multimodal dataset integrates three medical imaging benchmarks from BMAD [1], encompassing brain, liver, and retinal imaging modalities. The dataset includes 7,500 training and 3,715 test images for the brain class, 1,542 training and 1,493 test images for the liver class, and 4,297 training and 1,805 test images for the retinal class.

**Metrics.** We evaluate our model at the image level using three evaluation metrics: Area Under the Receiver Operating Characteristic curve (AU-ROC), Average Precision

Table 2. Comparison with SoTA methods on **Uni-Medical** dataset for multi-class anomaly localization with AU-ROC / AP / F1_max metrics. **Bold** numbers denote the best (SoTA) results in each row, and underlined numbers denote the second-best results.

| **Method→** **Category ↓** | **RD4AD** [6] CVPR'22 | **UniAD** [44] NeurIPS'22 | **SimpleNet** [32] CVPR'23 | **DeSTSeg** [46] CVPR'23 | **DiAD** [13] AAAI'24 | **MambaAD** [12] NeurIPS'24 | **Ours(PDD)** |
|---|---|---|---|---|---|---|---|
| **brain** | 82.4/94.4/91.5 | 89.9/97.5/92.6 | 82.3/95.6/90.9 | 84.5/95.0/92.1 | 93.7/98.1/95.0 | **94.2/98.6**/94.5 | 90.9/96.7/**96.7** |
| **liver** | 55.1/46.3/64.1 | 55.1/46.3/64.1 | 55.8/47.6/60.9 | **69.2/60.6**/64.7 | 59.2/55.6/60.9 | 63.2/53.1/64.7 | 59.7/50.5/**70.5** |
| **retinal** | 89.2/86.7/78.5 | 84.6/79.4/73.9 | 88.8/87.6/78.6 | 88.3/83.8/79.2 | 88.3/86.6/77.7 | **93.6**/88.7/86.6 | **93.6/92.7/88.9** |
| **Mean** | 75.6/75.8/78.0 | 78.5/75.2/76.6 | 75.6/76.9/76.8 | 80.7/79.8/78.7 | 80.4/**80.1**/77.8 | **83.7/80.1**/82.0 | 81.4/80.0/**85.4** |

(AP), and maximum F1 score (F1 max). Higher values indicate better performance for all these metrics.

**Implementation Details.** All the input images are resized to $256 \times 256$ pixels. We trained our model using the Adam optimizer with an initial learning rate of $2 \times 10^{-3}$, and the learning rate is updated using a cosine annealing schedule. All experiments are conducted on a single RTX A6000 GPU.

As illustrated in Fig. 3-(a), during the training phase, our model is trained exclusively on normal samples. The input images are processed through the InA, MMU, and MLP modules, and the model is optimized using the loss functions described in the previous sections ($\mathcal{L}_{\text{kr}}$, $\mathcal{L}_{\text{prp}}$, and $\mathcal{L}_{\text{div}}$). During the inference phase, the trained model can perform anomaly detection on unknown samples (either normal or abnormal) by computing anomaly scores based on the reconstruction errors and feature discrepancies, as shown in Fig. 3.

## 4.2. Comparative Experiment

We compare PDD against 6 state-of-the-art baselines: f-AnoGAN [35], CutPaste [22], RD4AD [6], SQUID [41], SIMSID [42], and Skip-TS [30]. As shown in Table 1, PDD achieves state-of-the-art AUROC on three out of 4 datasets—97.5% on HeadCT, 94.0% on Zhanglab, and 96.7% on BrainMRI—surpassing the best baseline by 11.8, 2.9, and 8.5 percentage points, respectively. On CheXpert, PDD remains competitive with the leading method SIMSID (79.1% vs. 79.7%).

On the Uni-Medical dataset (Table 2), PDD attains the highest F1 max across all three categories: 96.7% (brain), 70.5% (liver), and 88.9% (retinal), while also achieving the best AP of 92.7% on the retinal category. Overall, PDD surpasses the strongest competitor MambaAD by 3.4 percentage points on mean F1 max.

## 4.3. Ablation Study

To validate the effectiveness of each component in our proposed PDD framework, we conduct comprehensive ablation studies on the ZhangLab Chest-Xray and BrainMRI datasets. Table 3 presents the ablation results for different distillation paradigms and modular components.

Table 3. Ablation on distillation paradigm and modular components on **ZhangLab**. Metrics: AUROC / AUPR / F1 max (%).

| **Model** | InA w. MMU (2t1s) | InA w. MMU (2t2s) | RD (1t1s) | MPA | **AUROC↑** | **AUPR↑** | **F1↑** |
|---|---|---|---|---|---|---|---|
| M1 | ✗ | ✗ | ✓ | ✗ | 81.5 | 85.5 | 84.8 |
| M2 | ✗ | ✗ | ✓ | ✓ | 82.0 | 86.1 | 85.7 |
| M3 | ✓ | ✗ | ✗ | ✗ | 90.8 | 95.7 | 89.7 |
| M4 | ✓ | ✗ | ✗ | ✓ | 92.9 | 96.6 | 90.3 |
| Ours | ✗ | ✓ | ✗ | ✓ | **94.0** | **99.0** | **96.6** |

Table 4. Ablation on student consistency supervision and anomaly map computation strategies on **BrainMRI**. Metrics: AUROC / AUPR / F1 max (%).

| **Model** | $L_{div}$ | $cos(t1,s1)$ $+cos(t2,s2)$ | $cos(s1,s2)$ | **AUROC↑** | **AUPR↑** | **F1↑** |
|---|---|---|---|---|---|---|
| M1 | ✓ | ✗ | ✓ | 32.50 | 81.23 | 93.02 |
| M2 | ✗ | ✗ | ✓ | 70.46 | 97.26 | 95.65 |
| M3 | ✗ | ✓ | ✗ | 93.41 | 98.81 | 95.93 |
| Ours | ✓ | ✓ | ✗ | **96.67** | **99.51** | **97.56** |

Table 5. Ablation on $\tau_{\text{low}}$ and $\tau_{\text{high}}$ across datasets. Metric: AUROC (%).

| | **ZhangLab** | | | **HeadCT** | | | **BrainMRI** |
|---|---|---|---|---|---|---|---|
| $\tau_{\text{high}}$ | 0.40 | 0.30 | 0.75 | 0.30 | 0.30 | 0.75 | 0.75 |
| $\tau_{\text{low}}$ | 0.65 | 0.75 | 0.40 | 0.65 | 0.75 | 0.40 | 0.40 |
| AUROC | 94.52 | **96.18** | 95.92 | 96.92 | **97.31** | 96.15 | **93.30** |

**Ablation on Distillation Paradigms and Modular Components.** We evaluate different architectural configurations in Table 3, where $(2t1s)$ and $(2t2s)$ denote two teachers with one/two students, respectively. Starting from a vanilla RD(1t1s) baseline (Model 1: 81.5% / 85.5% / 84.8%), introducing the dual-teacher architecture with InA and MMU (Model 3) yields a substantial gain of 9.3 percentage points in AUROC, validating the benefit of complementary representations. Adding MPA (Model 4) further improves AUROC to 92.9%, and upgrading to a dual-student design (our final model) achieves the best performance of 94.0% / 99.0% / 96.6%. These results con-

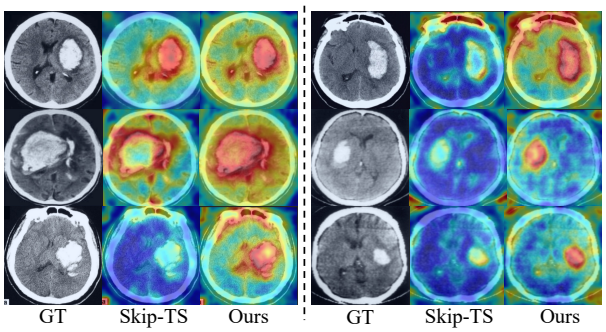



Figure 4. Anomaly localization comparison between PDD and Skip-TS on the HeadCT dataset.

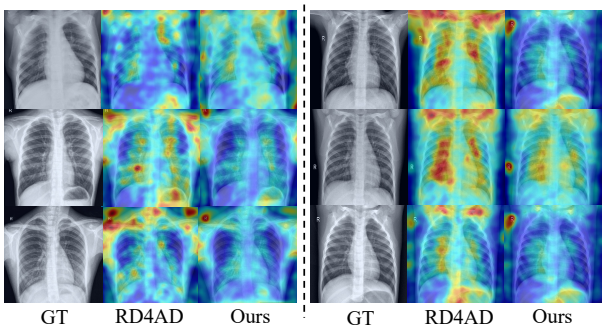



Figure 5. Anomaly localization comparison between PDD and RD4AD on the ZhangLab dataset. PDD produces significantly fewer false positives on normal samples, demonstrating stronger specificity.

firm that: (1) dual-teacher architecture with InA and MMU effectively leverages complementary representations; (2) dual-student design significantly enhances diverse normal pattern learning; and (3) MPA provides useful prior knowledge for further improvement.

**Ablation on Anomaly Localization.** We investigate the effects of training-time student consistency supervision and inference-time anomaly map computation on BrainMRI, as shown in Table 4. Three components are examined: diversity loss $L_{div}$ between two students during training, teacher-student cosine similarity $cos(t1, s1) + cos(t2, s2)$ for inference, and student-student cosine similarity $cos(s1, s2)$ for inference. Using only $cos(s1, s2)$ with $L_{div}$ (Model 1) leads to poor AUROC of 32.50%, while removing $L_{div}$ (Model 2) raises it to 70.46%, suggesting that $L_{div}$ forces divergent student representations that harm student-student based localization. Switching to teacher-student alignment (Model 3) significantly boosts AUROC to 93.41%. Our final model combines $L_{div}$ with teacher-student alignment, achieving the best performance of 96.67% / 99.51% / 97.56%. This demonstrates that: (1) teacher-student alignment is essential for effective anomaly localization; (2) $L_{div}$ prevents student collapse into identical representations; and (3) the combination of diversity supervision and teacher-student alignment achieves optimal results.

**Hyper-parameter Sensitivity.** We further conduct hyper-parameter ablations on the diversity thresholds $(\tau_{\text{low}}, \tau_{\text{high}})$ and the loss weights $(\lambda_{\text{kr}}, \lambda_{\text{prp}}, \lambda_{\text{div}})$. As shown in Table 5, PDD achieves consistently strong AUROC across datasets under different $(\tau_{\text{low}}, \tau_{\text{high}})$ combinations, indicating that the performance is not overly sensitive to these thresholds and the method is robust to moderate parameter changes. We set $\tau_{\text{low}} = 0.30$ and $\tau_{\text{high}} = 0.75$ in all experiments. For the overall objective, we use $\lambda_{\text{kr}} = \lambda_{\text{prp}} = 0.02$ and $\lambda_{\text{div}} = 0.5$, which provides a good balance between distillation, prior-guided reconstruction, and diversity regularization.

### 4.4. Anomaly Localization

Our anomaly localization is achieved during inference by aggregating the cosine similarity between each teacher-student pair, formulated as $\cos(t_1, s_1) + \cos(t_2, s_2)$.

Figure 4 presents qualitative anomaly localization results on HeadCT, comparing PDD with Skip-TS [30]. Each column shows ground truth (left), Skip-TS predictions (middle), and PDD predictions (right). Our method achieves more accurate localization, particularly for lesions with irregular boundaries and subtle pathological patterns that are often overlooked by Skip-TS.

Figure 5 compares anomaly localization between PDD and RD4AD on the ZhangLab dataset. As shown, RD4AD tends to produce spurious activations on normal samples, while PDD maintains cleaner predictions with significantly fewer false positives, indicating stronger specificity in distinguishing normal regions from anomalies.

## 5. Conclusion

We propose **PDD** (Manifold-Prior Diverse Distillation), a novel framework for medical image anomaly detection. By unifying complementary priors from VMamba-Tiny and Wide-ResNet50 via **MMU** and **InA** modules, and employing a dual-student architecture with **MPA**-based prior injection and diversity regularization, PDD enables diverse reconstruction of normal patterns while maintaining sensitivity to subtle anomalies. Experiments across multiple benchmarks show that PDD achieves state-of-the-art performance, confirming the efficacy of manifold-level prior modeling.

Despite its strong results, **PDD** can generate false positives on non-pathological artifacts such as device markers or metallic implants. These artifacts deviate visually from normal tissue but are clinically irrelevant. Mitigating this limitation through artifact-aware priors or clinical context integration is a key direction for future work.

## 6. Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276182), and Tianjin Natural Science Foundation (Nos. 24JCYBJC01230, 24JCYBJC01460, 24JCZDJC01190 ).

## References

[1] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053, 2024. 6

[2] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Evaluating normative representation learning in generative ai for robust anomaly detection in brain imaging. *Nature Communications*, 16(1):1624, 2025. 2

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 2

[4] Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. Medianomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, 102:103500, 2025. 1, 2

[5] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 2

[6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022. 2, 3, 6, 7

[7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 2

[8] Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Annan Shu, Guannan Jiang, and Lizhuang Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16401–16409, 2023. 2

[9] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15194–15203, 2025. 2

[10] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36:10721–10740, 2023. 2

[11] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9):2763–2775, 2023. 2

[12] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024. 2, 7

[13] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8472–8480, 2024. 7

[14] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng 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*, pages 11375–11385, 2024. 2

[15] Jiahao Huang, Liutao Yang, Fanwen Wang, Yinzhe Wu, Yang Nan, Weiwen Wu, Chengyan Wang, Kuangyu Shi, Angelica I Aviles-Rivero, Carola-Bibiane Schoenlieb, et al. Enhancing global sensitivity and uncertainty quantification in medical image reconstruction with monte carlo arbitrary-masked mamba. *Medical Image Analysis*, 99:103334, 2025. 2

[16] Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. Unsupervised anomaly detection in medical images using masked diffusion model. In *International Workshop on Machine Learning in Medical Imaging*, pages 372–381. Springer, 2023. 2

[17] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 6

[18] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018. 6

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[20] Ioannis Lagogiannis, Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Unsupervised pathology detection: a deep dive into the state of the art. *IEEE transactions on medical imaging*, 43(1):241–252, 2023. 1, 2

[21] Ioannis Lagogiannis, Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Unsupervised pathology detection: a deep dive into the state of the art. *IEEE transactions on medical imaging*, 43(1):241–252, 2023. 2

[22] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly de-

tection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021. 2, 6, 7

[23] Jiaxi Li, Di Lin, Hao Chen, Hongying Liu, Liang Wan, and Wei Feng. Causal inference via style bias deconfounding for domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 1

[24] Baijiong Lin, Weisen Jiang, Pengguang Chen, Shu Liu, and Ying-Cong Chen. Mtmamba++: Enhancing multi-task dense scene understanding via mamba-based decoders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[25] Chengde Lin, Xijun Lu, and Guangxi Chen. Ratlip: generative adversarial clip text-to-image synthesis based on recurrent affine transformations. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2346–2352. IEEE, 2024. 2

[26] Chengde Lin, Xuezhu Gong, Shuxue Ding, Mingzhe Yang, Xijun Lu, and Chengjun Mo. Stegot: Trade-offs in steganography via optimal transport. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. 2

[27] Junkang Liu, Fanhua Shang, Yuanyuan Liu, Hongying Liu, Yuangang Li, and YunXiang Gong. Fedbcgd: Communication-efficient accelerated block coordinate gradient descent for federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2955–2963, 2024. 2

[28] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pages 615–625. Springer, 2024. 2

[29] Junkang Liu, Fanhua Shang, Yuxuan Tian, Hongying Liu, and Yuanyuan Liu. Consistency of local and global flatness for federated learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3875–3883, 2025. 2

[30] Mingxuan Liu, Yunrui Jiao, Jingqiao Lu, and Hong Chen. Anomaly detection for medical images using teacher–student model with skip connections and multiscale anomaly consistency. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024. 3, 6, 7, 8

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2

[32] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20402–20411, 2023. 7

[33] Declan McIntosh and Alexandra Branzan Albu. Inter-realization channels: Unsupervised anomaly detection beyond one-class classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6285–6295, 2023. 2

[34] Sruti Rai, Siddhartha Chowdhury, Sampa Sarkar, Kinnar Chowdhury, and Kundan Prasad Singh. A hybrid approach to brain tumor detection from mri images using computer vision. *Journal of innovation in computer science and engineering*, 8(2):8–12, 2019. 3, 6

[35] T Schlegl, Philipp Seeböck, Sebastian M Waldstein, G Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks [j]. *Medical image analysis*, 54:30–44, 2019. 2, 6, 7

[36] Hui Sun, Long Lv, Pingping Zhang, Tongdan Tang, Feng Tian, Weibing Sun, and Huchuan Lu. Spatial-frequency enhanced mamba for multi-modal image fusion. *IEEE Transactions on Image Processing*, 34:7684–7696, 2025. 2

[37] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24511–24520, 2023. 2

[38] Mesut Toğaçar, Zafer Cömert, Burhan Ergen, and Ümit Budak. Brain hemorrhage detection based on heat maps, autoencoder and cnn architecture. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pages 1–5. IEEE, 2019. 6

[39] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017. 3

[40] Ziyi Wu, Tao Lu, Yanduo Zhang, and Xiaoyu Chai. Cmanet: A cnn-mamba aggregation network for face super-resolution. *Pattern Recognition*, 168:111859, 2025. 2

[41] Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23890–23901, 2023. 2, 6, 7

[42] Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Exploiting structural consistency of chest anatomy for unsupervised anomaly detection in radiography images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 6070–6081, 2024. 2, 6, 7

[43] Enquan Yang, Peng Xing, Hanyang Sun, Wenbo Guo, Yuanwei Ma, Zechao Li, and Dan Zeng. 3cad: A large-scale real-world 3c product dataset for unsupervised anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9175–9183, 2025. 3

[44] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 2, 7

[45] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2312.07495*, 2023. 6

[46] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3914–3923, 2023. 2, 7

[47] Wenbin Zuo, Hongying Liu, Huadeng Wang, Lingqi Zeng, Ningning Tang, Fanhua Shang, Liang Wan, and Jingjing Deng. Enhancing the impact of model performance gains for semi-supervised medical image segmentation. *Pattern Recognition*, 174:112889, 2026. 1