Hybrid Dual-Heterogeneous Knowledge Distillation Network for Anomaly Detection in Retinal OCT Images

Muhao Xu, Hua Wei, Zihan Nie, Xueying Zhou, Baochen Fu, Hongmei Yan, Yi Wan and Weiye Song

***Abstract*— Unsupervised medical anomaly detection aims to identify abnormal images by training exclusively on normal samples, thereby enabling the detection of disease-related irregularities without the need for large-scale labeled datasets. Current knowledge distillation-based methods typically detect anomalies by comparing feature discrepancies between teacher and student networks. However, because these methods employ an optimization strategy where the teacher and student architectures are highly similar, the student network's features tend to closely mirror those of the teacher, leading to an identity mapping issue. Moreover, the diversity of lesion types in retinal Optical Coherence Tomography (OCT) images further complicates anomaly detection. In this paper, we propose a novel hybrid dual-heterogeneous knowledge distillation network to overcome these challenges. Our approach consists of a teacher network with an encoder-only architecture and a student network that integrates an encoder with dual decoders. This heterogeneous design effectively mitigates the identity mapping problem, enhancing sensitivity to both structural and logical anomalies. Specifically, our Multi-Feature Model leverages convolutional and depthwise convolutional blocks to extract and integrate local features for structural anomaly detection, while the Mamba UpNet employs self-supervised learning to capture long-range dependencies and global anomaly patterns. Extensive experiments on two retinal OCT anomaly detection datasets demonstrate that our method achieves state-of-the-art performance, effectively handling diverse anomaly types. The source code is available at https://github.com/Xmh-L/HDHKD.**

***Index Terms*— Anomaly localization, Knowledge Distillation, Unsupervised learning, Dual-Heterogeneous**

This work was supported in part by the Youth Project of Natural Science Foundation of Shandong Province, China (ZR2023QC262); the National Natural Science Foundation of China (62205181); the Natural Science Foundation of Shandong Province (ZR2022QF017); and the Shandong Province Outstanding Youth Science Fund Project (Overseas) (2023HWYQ-023); the Key R&D Program of Shandong Province, China (2024CXGC010106); the Taishan Scholar Foundation of Shandong Province (tsqn202211038).

Muhao Xu, Zihan Nie, Baochen Fu, Hua Wei, Yi Wan and Weiye Song are with the Department of Mechanical Engineering, Key Laboratory of High Efficiency and Clean Mechanical Manufacture of Ministry of Education, Shandong University, Jinan 250061, China. Baochen Fu is also with the School of Software, Shandong University, Shandong, Jinan 250061, China. (Corresponding authors: Weiye Song, e-mail:songweiye@sdu.edu.cn)

Xueying Zhou is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China.

Hongmei Yan is with the Shandong Provincial Key Laboratory of Sensor Technology and High Precision Weighing Instruments, Jinzhong Group, Jinan 250061, China.

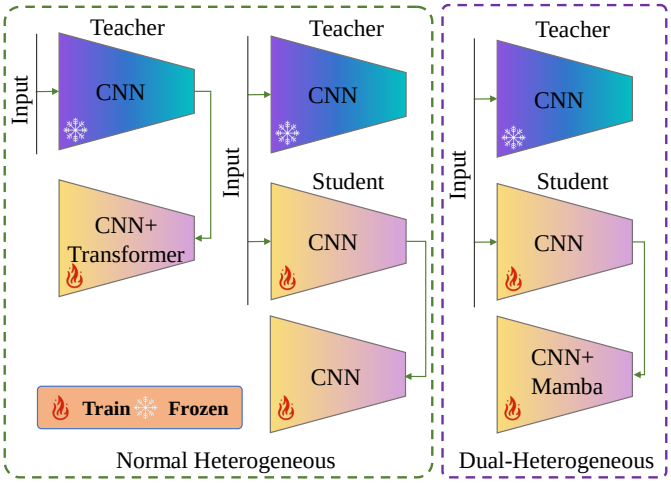

Fig. 1. Comparison of Knowledge Distillation Strategies: (a) Normal Heterogeneous: A frozen CNN-based teacher encoder supervises student architectures with heterogeneous decoders. Two variants are depicted: one employs a CNN+Transformer decoder, while the other utilizes a plain CNN encoder and decoder; both student decoders are trainable. (b) Dual-Heterogeneous: A frozen CNN-based teacher encoder transfers knowledge to a student model comprising a shared CNN encoder and two distinct decoders—a standard CNN and a CNN augmented with Mamba. Both decoders are jointly trained to enhance representational diversity.

I. Introduction

Unsupervised anomaly detection (UAD) [1]–[3] has attracted increasing attention in medical imaging, as it alleviates the need for costly and time-consuming annotations that require clinical expertise. This is particularly important for optical coherence tomography (OCT) [4], [5], a non-invasive volumetric modality widely used in ophthalmological diagnostics [6], [7], where annotation is especially challenging for rare or complex diseases and more demanding than for natural images [8], [9]. Existing UAD methods can be broadly divided into reconstruction-based and representation-based approaches. Reconstruction-based methods [10], [11] learn to reproduce normal samples but often suffer from limited reconstruction accuracy, which reduces their ability to distinguish subtle anomalies. Representation-based methods [12], [13] focus on learning compact embeddings of normal data, but they are usually less sensitive to global structural anomalies, a limitation that is particularly critical for OCT

images, where both local retinal layers and long-range tissue organization need to be considered.

Recently, knowledge distillation-based methods [14], [15] have demonstrated strong potential in anomaly detection. During training, a teacher network guides the student network to learn feature representations exclusively from normal samples. By transferring knowledge from a teacher trained on normal data, these approaches effectively capture the distribution of normal features. In testing, anomalies are identified by measuring discrepancies between the representations produced by the teacher and the student. However, a fundamental limitation arises when the student network shares an identical or highly similar architecture with its teacher. As discussed in RD4AD [16], this structural homogeneity can lead to isomorphic mapping, where the student tends to replicate the teacher's internal representations rather than learning discriminative deviations. Such identity-like behavior weakens anomaly sensitivity and reduces the model's ability to distinguish pathological regions from normal structures. Subsequent studies, including Hetero-AE [17] and EfficientAD [18], demonstrated that introducing architectural or functional heterogeneity between the teacher and student can effectively mitigate this issue by promoting representational diversity and preventing direct feature alignment. Furthermore, retinal OCT images present a unique combination of fine-grained local structures and long-range global dependencies. Conventional architectures often favor one aspect—either local texture abnormalities or global morphological deviations—resulting in unbalanced detection performance. Convolutional networks are adept at modeling localized structural disruptions but limited in capturing extended dependencies, whereas Mamba-like or global modeling designs better handle large-scale contextual changes such as diffuse layer disorganization or overall retinal thinning. Therefore, a framework that simultaneously prevents identity mapping and effectively integrates both local and global cues is essential for achieving accurate and clinically meaningful anomaly detection in retinal OCT imaging.

To address these challenges, we propose a novel hybrid dual-heterogeneous knowledge distillation network for anomaly detection in retinal OCT images. Fig. 1 illustrates the knowledge distillation strategies considered, including the proposed dual-heterogeneous design. Our framework consists of a teacher with an encoder-only architecture and a student comprising a ResNet-based encoder with dual heterogeneous decoders. Heterogeneity between the teacher and student mitigates identity mapping, while the additional heterogeneity between the student's encoder and decoders further regularizes feature learning and enhances anomaly sensitivity. Specifically, the dual-decoder structure is tailored to capture complementary cues: a convolutional upsampling branch extracts fine-grained local structural anomalies, whereas the Mamba UpNet models long-range dependencies and global patterns through selective scanning, which is particularly suited for the layered continuity of OCT images. The outputs are fused via the Multi-Feature Model to exploit complementary strengths, resulting in enhanced detection performance across both structural and global anomalies. This dual-heterogeneous design not only alleviates the risk of trivial feature replication but also improves the robustness, sensitivity, and accuracy of anomaly detection in retinal OCT imaging.

Our contributions can be summarized as follows:

- We introduce a hybrid dual-heterogeneous knowledge distillation network for anomaly detection in retinal OCT images, which effectively mitigates the identity mapping problem through the innovative design of the student network's encoder-decoder architecture.
- We introduce the Multi-Feature Model, which utilizes convolutional layers and depthwise convolutional blocks to enable robust feature extraction and interaction, rendering it particularly effective for detecting structural anomalies.
- We introduce the Mamba UpNet integrates self-supervised mechanisms to capture long-range dependencies and enhance the detection of logical anomalies.

II. Related Work

A. Feature Representation-based Methods

Feature representation-based anomaly detection methods focus on extracting meaningful feature representations from data and using these representations for anomaly detection. By leveraging deep learning, they are capable of effectively capturing complex structures and non-linear relationships, making them particularly well-suited for high-dimensional and unstructured data. DeepSVDD [19] introduced a deep one-class classification framework that learns a compact representation of normal data in a latent space. Anomalies are identified based on their distance from the center of the learned hypersphere. While this approach effectively captures the essential characteristics of normal data, it assumes that the normal data distribution can be compactly represented by a hypersphere. This assumption may not hold for complex or multi-modal data distributions, potentially limiting the method's applicability in such cases. Furthermore, the model may exhibit reduced sensitivity to anomalies that are subtle or located near the boundary of the normal data distribution. PaDiM [20] proposed a statistical approach for anomaly detection, modeling feature representations using multivariate Gaussian distributions. By calculating the Mahalanobis distance between test samples and the learned normal distribution, PaDiM enhances detection performance. However, the reliance on Gaussian assumptions limits its effectiveness in handling complex, non-Gaussian data distributions.

To address these challenges, Glow [21] was explored for anomaly detection, utilizing normalizing flows to model the distribution of normal data. By evaluating the likelihood of test samples, anomalies can be detected. This method provides greater flexibility in modeling complex data distributions but requires careful tuning of the flow architecture to balance model performance and computational cost. In parallel, PatchCore [12] introduces an efficient patch-based approach, utilizing a greedy algorithm to select the most representative features from the normal dataset. This method strikes a balance between detection accuracy and computational efficiency by reducing memory consumption and accelerating the detection process. However, it may struggle to detect anomalies in

sparsely sampled or low-frequency regions of the feature space, as these subtle anomalies may be excluded during feature selection. Fine-tuning the number of retained features is also necessary to ensure optimal performance.

These methods have made significant advancements in the field of anomaly detection. However, they continue to face challenges in balancing detection accuracy, computational efficiency, and adaptability to complex data distributions.

B. The Reconstruction-based Methods

Reconstruction-based methods have become a prominent approach in anomaly detection, stemming from the assumption that models trained solely on normal data will struggle to reconstruct anomalous instances. Schlegl et al. demonstrated that models minimizing reconstruction errors for normal data will exhibit poor performance when tasked with reconstructing anomalies, leading to increased reconstruction errors for abnormal samples [22]. Autoencoders (AEs) are widely used in this domain to learn normal data distributions, with deviations from this norm signaling potential anomalies [23]. Despite their effectiveness, reconstruction-based methods face challenges when anomalies share characteristics with normal data, resulting in low reconstruction errors even for anomalous samples. To address this, Zong et al. introduced a hybrid autoencoder that combines both reconstruction and adversarial training to improve anomaly detection by leveraging unsupervised and adversarial learning mechanisms [24].

Furthermore, the issue of overfitting normal data can lead to poor generalization when confronted with novel types of anomalies. To mitigate this, adversarial training has been incorporated into reconstruction-based methods, encouraging models to focus on harder-to-reconstruct features, thereby improving their robustness [25].

C. Distillation-based Methods

Distillation-based methods are increasingly employed for anomaly detection, relying on a teacher-student framework where a pre-trained teacher network transfers knowledge to a student network. Anomalies are detected based on the discrepancies between the outputs of the two networks. Bergmann et al. used this framework by leveraging the uncertainty of student networks, employing these discrepancies as an anomaly score to identify outliers in data [14]. However, this approach can be limited when anomalies are subtle or when the output differences between the two models are insufficient to highlight these anomalies effectively. To address these limitations, Salehi et al. proposed a multi-resolution knowledge distillation approach, which uses the intermediate activations between a complex teacher model and a simpler student model. By capturing the differences in features across multiple resolutions, this approach enhances the ability of the student model to differentiate between normal and anomalous data, thus improving anomaly detection [15]. However, challenges arise when the test data significantly differs from the training data distribution, as this can hinder effective knowledge transfer, resulting in reduced generalizability.

Reverse distillation methods have also been proposed to combine distillation with reconstruction, as seen in the RD framework by Deng et al. By incorporating an encoder-decoder architecture, this method aims to improve anomaly detection by refining the feature learning process through the reconstruction of the input data [16]. While these techniques offer improvements, they still struggle with issues related to feature compression and the loss of detailed information, which are crucial for detecting subtle anomalies. Guo et al. addressed these issues with a template-based reconstruction method, THFR, which uses normal templates to guide the recovery process of abnormal instances [26]. This method improves anomaly detection by enhancing the quality of reconstruction. However, a significant challenge remains: if the normal template is not sufficiently representative of the test image, the reconstruction may degrade, potentially lowering detection accuracy. Moreover, this method may mask subtle anomalies by reconstructing them to appear normal, which emphasizes the need for a careful balance between the template guidance and the model's sensitivity to anomalies.

III. Proposed Method

A. Overview of the Framework

Retinal diseases exhibit highly diverse and often unpredictable morphological patterns. Even within a single diagnostic category, OCT manifestations can vary substantially across patients and disease stages, producing pronounced intra-class heterogeneity. Such variability increases the intrinsic sample complexity of the problem and implies that effective anomaly detectors must be sensitive to abnormalities that appear at multiple spatial scales and in a wide variety of forms. To address these requirements, we propose a Hybrid Dual-Heterogeneous Knowledge Distillation Network for anomaly detection in OCT retinal images. As illustrated in Fig. 2, the overall framework consists of a Teacher network and a Student network. The Teacher network follows an Encoder-Only architecture, while the Student network comprises a ResNet-based Encoder, two functionally distinct Decoders (Conv Upsample and Mamba UpNet), and a Multi-Feature Model. The dual-heterogeneous design effectively mitigates the commonly encountered "identity mapping" issue in conventional knowledge distillation, thereby significantly enhancing the detection of both structural and global anomalies. Given an input OCT image $X \in \mathbb{R}^{C\times H\times W}$, the Teacher network first extracts high-level feature representations $f_t \in \mathbb{R}^{C_T\times H_T\times W_T}$ via its encoder. These features not only serve as informative teacher representations for anomaly detection but also provide distilled supervision signals for the Student network.

The Student network is composed of the ResNet-based Encoder and two complementary Decoders. Specifically, the encoder produces multi-scale feature maps $\tilde{S}_1, \tilde{S}_2, \tilde{S}_3$. These multi-scale features are subsequently fused via convolutional operations and reconstructed using the Conv Upsample and Mamba UpNet decoders. Finally, the Multi-Feature Model aggregates the reconstructed outputs to effectively address diverse anomaly types: the Conv Upsample decoder is tailored for detecting local structural anomalies, whereas the Mamba

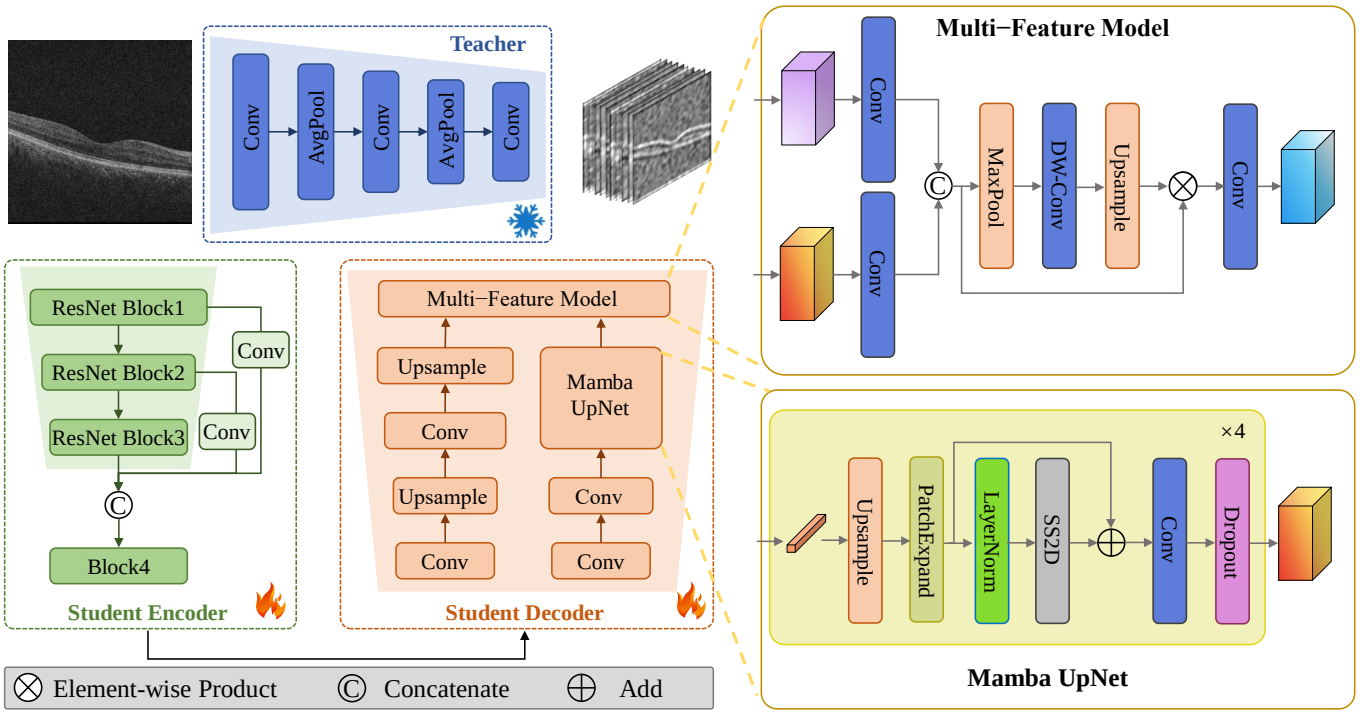

Fig. 2. Overview architecture of proposed Hybrid Dual-heterogeneous Knowledge Distillation Network.A spark icon indicates modules involved in training, while a snowflake icon denotes networks that are frozen.

UpNet decoder is designed to capture global anomalies with long-range dependencies. Ultimately, the discrepancy between the outputs of the Teacher and Student networks drives the anomaly detection process.

B. Teacher Network: Encoder-Only

The architecture of our teacher network F_T comprises convolutional layers interleaved with average pooling layers. Drawing inspiration from the Efficient Anomaly Detection [18] framework, our network is designed so that each neuron in the output layer covers a receptive field of 33×33 pixels. This configuration minimizes the likelihood that anomalies in one region will impact distant, unrelated areas, thereby enhancing the precision of anomaly localization. Notably, F_T is distilled from a pre-trained WideResNet-101 model [27], leveraging the extensive ImageNet dataset.

During the distillation process, the Mean Squared Error (MSE) is adopted as the loss function to guide the learning of F_T. To further improve our model, we employ a batch size of 32 during training and adopt the PatchCore [12] feature post-processing strategy, which fuses features from two layers into a 384-dimensional representation. Its encoder is capable of extracting high-dimensional features that embody richer spatial structure and semantic information, a property that is critical for detecting various complex pathological abnormalities in retinal OCT images. We obtain the features $f_t \in \mathbb{R}^{h \times w \times c}$ by feeding the training images x into F_T, formally defined as:

$$f_t = F_T(x). \tag{1}$$

Subsequently, the feature map f_t is employed to supervise the Student network.

C. Student Network: Encoder & Dual-Decoders

To overcome the identity mapping issue inherent in homogeneous networks used in traditional knowledge distillation, we introduce a heterogeneous architecture in the Student network. Specifically, the encoder is designed differently from that of the teacher network by employing a ResNet backbone, while the decoders are distinct from the student encoder.

The student encoder is built upon a ResNet backbone and includes four sequential blocks, each comprising a convolution, normalization, and activation sequence. The outputs from ResNet Block1, ResNet Block2, and ResNet Block3 are individually processed by a convolution layer and then merged via the concatenation operator (denoted by C). The fused features are subsequently passed through Block4 to generate the final features representation f_{se}. The specific formula is expressed as:

$$\begin{aligned} \tilde{S}_i &= \mathrm{Conv}\big(\mathrm{ResNetBlock}_i(X)\big), \quad i = 1, 2, 3, \\ S_f &= \mathrm{ConCat}\big(\tilde{S}_1, \tilde{S}_2, \tilde{S}_3\big), \\ f_{se} &= \mathrm{Block4}(S_f), \end{aligned} \tag{2}$$

where $\mathrm{ResNetBlock}_i(\cdot)$ denotes the i-th residual block, $\mathrm{Conv}(\cdot)$ represents the convolution operation, and $\mathrm{C}(\cdot)$ stands for concatenation.

In the Student Decoder structure, we employ two decoding branches for the input feature f_{se} and integrate them through a feature fusion module to obtain the final output feature Y. As illustrated in the Fig. 2, the Student Decoder consists of the Conv Upsample Branch, Mamba UpNet Branch, and Multi-Feature Model.

Conv Upsample Branch: This branch progressively re-

stores spatial resolution and focuses on local anomaly detection. Specifically, the input feature X undergoes multiple upsampling operations, each followed by a convolutional layer to refine spatial features. The cascaded "upsample-convolution" structure enhances feature resolution while preserving high-level semantics, facilitating the capture of fine-grained details. The output of this branch, denoted as X_{local}, is computed as follows:

$$\begin{aligned} f_s{}' &= \text{Conv}(\text{Upsample}(f_{se})), \\ f_{\text{local}} &= \text{Conv}(\text{Upsample}(f_s{}')). \end{aligned} \tag{3}$$

Algorithm 1 Student Decoder — Conv Upsample Branch and Mamba UpNet Branch

Require: Student encoder feature f_{se}, number of Mamba modules N_{mamba}
Ensure: Local feature map f_{local}, Global feature map f_{global}

procedure CONVUPSAMPLEBRANCH(f_{se})
  $U_1 \leftarrow \text{Conv}(\text{Conv}(\text{Upsample}(f_{se})))$
  return U_1
end procedure
procedure MAMBAUPNETBRANCH(U_1, N_{mamba})
  for $i = 1$ **to** N_{mamba} **do**
    $f_{\text{global}} \leftarrow$ MAMBAMODULE(U_1)
  end for
end procedure
procedure MAMBAMODULE(U_1)
  $U_{\text{up}} \leftarrow \text{Upsample}(U_1)$
  $U_{\text{pe}} \leftarrow$ PATCHEXPAND(U_{up})
  $U_{\text{ln}} \leftarrow \text{LayerNorm}(U_{\text{pe}})$
  $U_{\text{ss}} \leftarrow \text{SS2D}(U_{\text{ln}})$
  $f_{\text{global}} \leftarrow \text{Dropout}(\text{Conv}(U_{\text{ss}} + U_{\text{pe}}))$ ▷ skip connection add ($\oplus$)
  return f_{global}
end procedure
procedure PATCHEXPAND(U_{up})
  $T' \leftarrow \text{LinearProj}(U_{\text{up}})$
  $U_{\text{pe}} \leftarrow \text{RearrangeToGrid}(T')$
  return U_{pe}
end procedure
procedure SS2D(U_{ln})
  $D_{\text{up}} \leftarrow$ SCANDIRECTION(U_{ln}, up)
  $D_{\text{down}} \leftarrow$ SCANDIRECTION(U_{ln}, down)
  $D_{\text{left}} \leftarrow$ SCANDIRECTION(U_{ln}, left)
  $D_{\text{right}} \leftarrow$ SCANDIRECTION(U_{ln}, right)
  $U_{\text{ss}} \leftarrow \text{MergeDirections}(D_{\text{up}}, D_{\text{down}}, D_{\text{left}}, D_{\text{right}})$
  return U_{ss}
end procedure

Mamba UpNet Branch:

To increase the sensitivity to global anomalies, the branch uses a stack of four Mamba UpNet structures. Mamba UpNet consists of multiple stacked modules, each comprising dilated convolution, PatchExpand, 2D-Selective-Scan, multi-scale pooling, and feature rearrangement operations to effectively capture global context. The 2D-Selective-Scan (SS2D) module [28] first decomposes the input image into independent sequences in the four directions (up, down, left and right). This ensures broad spatial coverage of the information and enables multi-directional feature capture. Next, feature extraction is performed by applying the formulas of the state space model. This process allows information from different directions to be integrated, while preserving important contextual information and filtering out irrelevant content. This achieves a global sense field while maintaining linear complexity. Finally, scanning and merging operations reconfigure these sequences to produce an output image matching the size of the original. PatchExpand linearly projects the features of each token into a higher-dimensional channel space, thereby doubling the channel capacity. The expanded channels are then rearranged into a finer spatial grid, which restores resolution while preserving feature coherence. The global feature map f_{global} is obtained by the steps outlined in Algorithm 1.

Multi-Feature Model: After extracting features from both branches, the Multi-Feature Model integrates f_{local} and f_{global} at the feature level to fully exploit both local and global information. In this process, f_{local} and f_{global} are individually transformed via convolution and then concatenated ($\oplus$) to yield f_m. Subsequently, f_m undergoes max pooling (MaxPool), depthwise separable convolution (DWConv), and upsampling (Upsample). This output is then element-wise multiplied ($\otimes$) with f_{m} to enhance feature interactions, and the final fused feature Y is obtained through an additional convolution operation, which is computed as:

$$\begin{aligned} f_m &= \text{ConCat}(\text{Conv}(f_{\text{global}}), \text{Conv}(f_{\text{local}})), \\ f_{m'} &= \text{Upsample}(\text{DWConv}(\text{MaxPool}(f_m))), \\ f_{sd} &= \text{Conv}(f_{\text{m}} \otimes f_{m'}). \end{aligned} \tag{4}$$

During the training process, we guide the student network learning by calculating the L2 loss between f_t and f_{sd}, which is formulated as:

$$\mathcal{L} = ||f_t - f_{sd}||^2. \tag{5}$$

IV. EXPERIMENTS

A. Dataset

Retinal OCT 2017 Dataset [34] was acquired from the Spectralis OCT system (Heidelberg Engineering, Germany) and serves as a valuable benchmark for retinal image anomaly detection. It comprises four categories: choroidal neovascularization (CNV), diabetic macular edema (DME), Drusen, and normal. For balanced evaluation, the publisher partitions the data into a training set of 26,315 normal images and a test set of 1,000 images—250 normal and 750 abnormal (covering CNV, DME, and Drusen). Our model is trained solely on normal images to capture standard retinal features, and its performance is subsequently evaluated on the complete test set, ensuring a comprehensive assessment of its anomaly detection capabilities.

Retinal OCT 2022 Dataset [35] was obtained from Cirrus OCT (Carl Zeiss Meditec, Inc., Dublin, CA), comprising 31 OCT volumes from 18 healthy individuals and 100 OCT volumes from 50 patients diagnosed with retinal edema. Both oculus sinister and oculus dexter were included in the OCT volumes of both healthy and affected individuals, with no

TABLE I
QUANTITATIVE RESULTS ON RETINAL OCT 2017 DATASET FOR ANOMALY DETECTION.

Method	AUROC	F1-score	ACC	SEN	SPE
AE [29]	77.79	85.79	78.28	94.38	41.70
VAE [30]	80.04	85.55	77.63	95.32	37.45
Ganomaly [10]	83.53	88.65	81.60	95.86	38.80
f-AnoGAN [31]	83.35	84.73	77.50	89.89	49.36
PaDim [20]	90.67	89.84	85.10	87.87	76.80
SALAD [32]	96.42	93.42	90.64	95.69	79.15
STFPM [33]	96.86	95.80	93.70	95.60	88.00
MKD [15]	96.72	94.60	91.60	97.60	73.60
RD4AD [16]	97.64	96.40	94.60	96.40	89.20
Hetero-AE [17]	98.94	97.10	95.76	97.46	90.64
PatchCore [12]	97.52	96.96	96.61	97.65	91.25
EfficientAD [18]	98.95	97.01	96.68	98.51	91.40
Ours	99.72	99.07	98.60	99.07	97.20

exclusion criteria based on age, gender, or ethnicity. Each OCT volume has a voxel size of $512 \times 1024 \times 128$, covering an approximate region of $6\,\text{mm} \times 2\,\text{mm} \times 6\,\text{mm}$ in the central macula. After preprocessing, the training dataset consists of 3965 B-scan images from healthy individuals, while the test dataset contains 6059 normal and 6734 abnormal B-scan images, which were manually labeled and extracted from the OCT volumes of the patients.

Brain MRI (Br35H) Dataset [36] collection contains a total of 3,000 two-dimensional T1/T2 MRI slices, of which 1,500 have been designated as tumourous and the remaining 1,500 as non-tumorous. To conform to the unsupervised anomaly detection setting, we used 500 non-tumorous slices for training; the remaining 1,000 non-tumorous slices together with the 1,500 tumorous slices were retained for testing. This split emphasizes the model's ability to detect tumor-related deviations from a normal training distribution while preserving a realistic balance of normal and pathological examples at test time.

Chest CT Dataset comprises 10,810 axial slices collected from 110 patients presenting a spectrum of thoracic abnormalities (including, but not limited to, small-cell lung carcinoma, lung adenocarcinoma, fibrous hyperplasia, and granulomatous lesions). Imaging was acquired on commonly used clinical scanners (GE Discovery STe 16 and Siemens Biograph Vision 600). For the UAD protocol, only normal slices were included in the training set (4,456 images); the remaining 6,354 slices were reserved for evaluation.

PET Dataset includes a total of 10,000 PET slices, spanning diverse physiological and pathological conditions such as inflammatory responses, metabolic hyperactivity, and malignant uptake regions. Data acquisition was performed using Siemens Biograph Vision 600 scanners under standardized ^{18}F-FDG protocols. For the anomaly detection task, 4,000 normal PET slices were used for training, and 6,000 slices (3,000 normal and 3,000 abnormal) were used for evaluation. All PET slices were reconstructed with attenuation correction and rescaled to 256×256 for network input.

TABLE II
KEY EXPERIMENTAL SETTINGS FOR REPRODUCIBILITY.

Item	Setting
Python	3.8
Framework	PyTorch 2.4.1
Hardware	NVIDIA A100 GPU
Input resolution	256×256
Teacher backbone	WideResNet-101 (ImageNet pretrained, frozen)
Student backbone	ResNet-family encoder (scratch)
Decoders	Conv Upsample & Mamba UpNet
Loss	Feature-level L2
Optimizer	Adam
Learning rate	$1 \times 10^{-4} \rightarrow 1 \times 10^{-5}$ @ $0.95\times$ Training epochs (StepLR, γ=0.1)
Weight decay	1×10^{-5}
Batch size	1
Dropout	0.2 (decoder)
Training epochs	80,000
Evaluation metrics	AUROC, F1, ACC, SEN, SPE
Post-processing	Bilinear upsampling of anomaly map

B. Implementation Details

The proposed method has been implemented in PyTorch and executed on machines equipped with NVIDIA A100 GPUs. All input images were resized to 256×256 prior to training and evaluation. Following established practice for the considered OCT benchmarks, we adopted a one-model-per-category protocol: a separate student model was trained for each anomaly category to enable fair comparison with prior work. Dataset splits (training / validation / test) follow the predefined partitions provided with each public dataset.

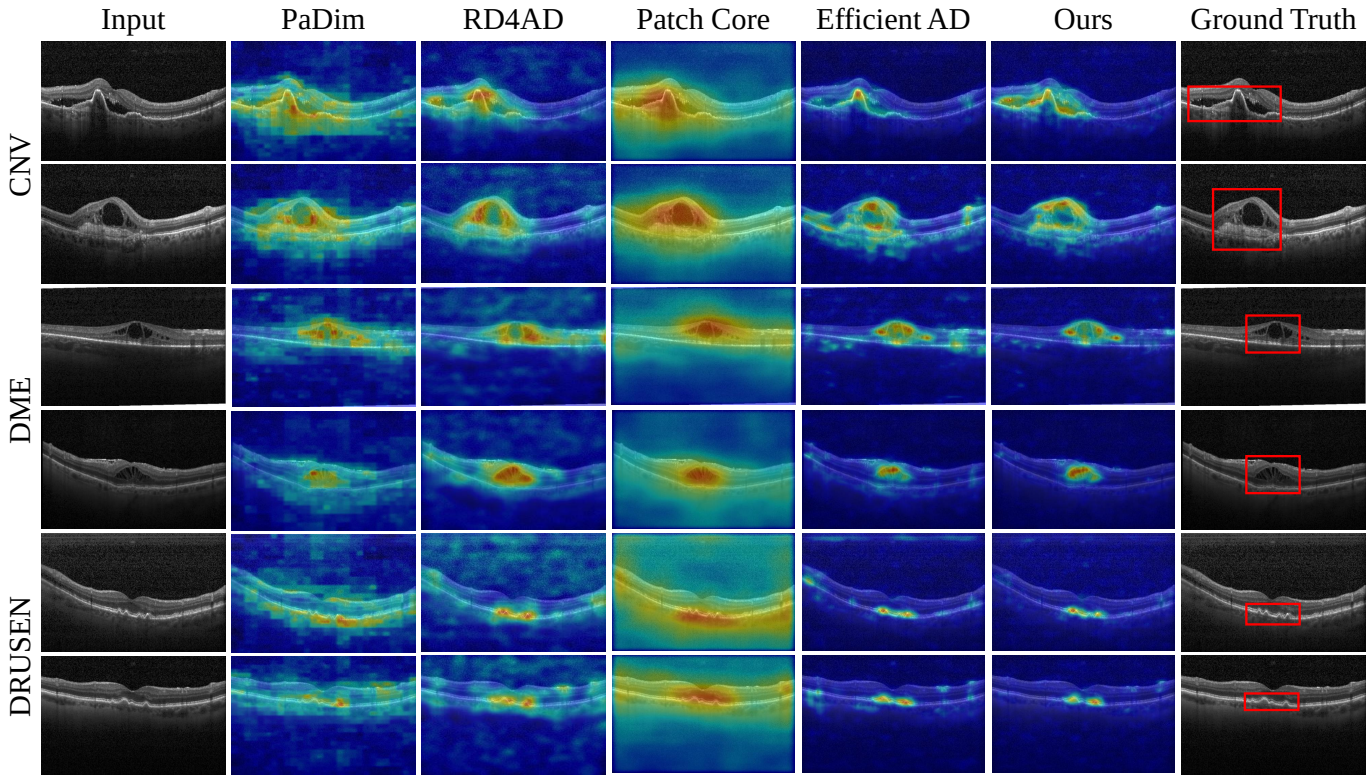

Fig. 3. Visualization of anomaly localization results on different categories of the Retinal OCT 2017 dataset.

The training configuration for each student model is as follows. Student models were trained from scratch with a ResNet-family encoder and two decoder branches (Conv Upsample and Mamba UpNet). The teacher encoder was initialized from a WideResNet-101 model pretrained on ImageNet and was kept frozen during student training; the teacher therefore provides fixed feature-level supervision. For specific details of a repetitive nature, refer to Table II. During inference, the anomaly score map $\mathcal{A}$ is bilinearly upsampled to the original image resolution for visualization and pixel-level localization. Complete training and evaluation scripts, environment specification (requirements/conda), and pretrained weights are provided in the accompanying code release to facilitate exact reproduction.

Evaluation metrics:

- **AUROC (Area Under the ROC Curve):**

$$AUROC = \int_0^1 TPR(FPR)\, dFPR,$$
$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{FP+TN}.$$

- **Accuracy (ACC):**

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}.$$

- **Sensitivity (SEN):**

$$SEN = \frac{TP}{TP+FN}.$$

- **Specificity (SPE):**

$$SPE = \frac{TN}{TN+FP}.$$

- **F1-score:**

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall},$$
$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN},$$

where TP (True Positives) denotes correctly identified positive samples, TN (True Negatives) denotes correctly identified negative samples, FP (False Positives) denotes negative samples incorrectly classified as positive, and FN (False Negatives) denotes positive samples incorrectly classified as negative.

TABLE III
QUANTITATIVE COMPARISON WITH DIFFERENT ANOMALY DETECTION METHODS ON RETINAL OCT 2022 DATASET.

Method	AUROC	AUPRC	ACC
AE [29]	86.99	84.76	80.34
VAE [30]	87.63	84.69	81.56
AnoVAEGAN [37]	86.87	83.54	80.36
GANomaly [10]	88.96	90.88	81.19
Deep-SVDD [19]	80.7	81.44	74.78
f-AnoGAN [31]	89.75	91.03	80.99
P-Net [38]	93.87	95.15	86.64
KD_AD [15]	93.98	94.68	86.16
PaDim [20]	84.14	83.63	77.39
SCVAE-AC [35]	97.02	97.38	91.49
RD4AD [16]	95.54	95.26	89.48
PatchCore [12]	94.25	95.41	86.72
EfficientAD [18]	97.05	97.14	91.58
Ours	**97.42**	**97.56**	**92.36**

TABLE IV
COMPARISON OF ANOMALY DETECTION PERFORMANCE (AUROC AND F1-SCORE) ACROSS CT, MRI, AND PET DATASETS.

Method \ Dataset	CT		MRI		PET	
	AUROC	F1	AUROC	F1	AUROC	F1
AE [29]	62.75	65.06	85.82	87.82	66.91	70.34
GANomaly [10]	75.06	69.01	88.45	90.64	78.59	71.16
PaDiM [20]	54.84	67.86	91.69	92.38	62.37	72.95
PatchCore [12]	87.74	82.52	93.62	94.07	86.42	84.13
MKD [15]	87.24	79.73	94.65	95.23	84.64	83.46
RD4AD [16]	86.81	81.89	94.90	95.81	82.55	84.71
EfficientAD [18]	84.46	80.97	96.13	97.46	83.34	79.68
Ours	**87.94**	**82.95**	**96.38**	**97.67**	**86.96**	**85.52**

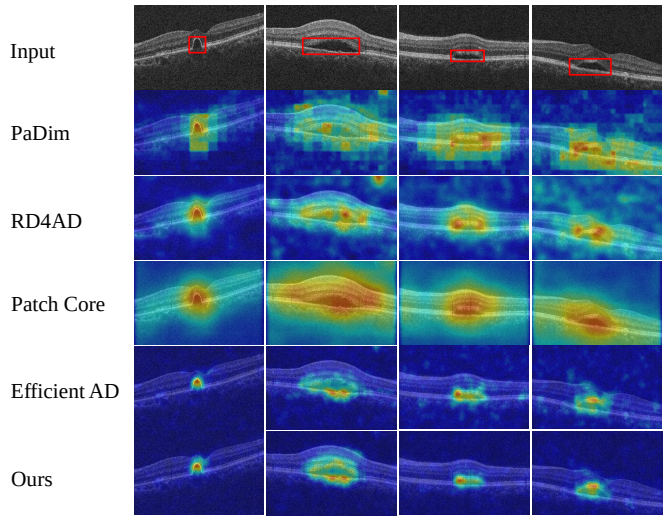

Fig. 4. Visualization of anomaly localization results on different categories of Retinal OCT 2022 dataset.

C. Anomaly detection and localization

Results on Retinal OCT 2017 Dataset

Table I presents a comparison of our approach (Ours) with state-of-the-art anomaly detection methods, including AE, VAE, Ganomaly, f-AnoGAN, PaDim, SALAD, STFPM, MKD, RD4AD, Hetero-AE, PatchCore, and EfficientAD. We report commonly used metrics in anomaly detection: Area Under the ROC Curve (AUROC), F1-score, Accuracy (ACC), Sensitivity (SEN), and Specificity (SPE). As shown in Table I, our proposed method achieves the highest AUROC (99.70%), outperforming all competing approaches by a notable margin. In particular, our model also attains the highest F1-score (99.07%), accuracy (98.60%), sensitivity (99.07%), and specificity (97.02%), indicating its robust ability to capture various abnormal patterns in retinal structures. Moreover, the overall performance is consistently strong across all metrics, highlighting the superiority of our hybrid distillation-based approach in balancing local structural cues and global contextual information.

To further demonstrate the effectiveness of our method, we present qualitative comparisons in Fig. 3, where we show the anomaly maps generated by PaDim, RD4AD, PatchCore, EfficientAD, and our method on representative examples of CNV, DME, and Drusen. The final column illustrates the ground truth annotations (highlighted in red boxes). From Fig. 3, it can be observed that our method generates clearer and more concentrated heatmaps for the abnormal areas, accurately localizing the lesions in CNV, DME, and Drusen images. Compared with other methods, which often produce over-smoothed or diffused activation regions, our proposed approach leverages dual-heterogeneous knowledge distillation to effectively learn discriminative features. This allows it to distinguish between normal retinal structures and pathological variations with higher fidelity, further confirming our method's superior anomaly detection performance.

Results on Retinal OCT 2022 Dataset

To validate the effectiveness of the proposed method, we performed experiments on the Cirrus retinal OCT 2022 dataset and compared our approach against various anomaly detection methods. For a fair comparison, all methods are evaluated under the same data splits and metrics. From Table III, we observe that traditional autoencoder-based methods (AE, VAE) achieve relatively modest performance on all three metrics, likely due to their limited capacity for representing complex structural and global anomaly patterns. GAN-based methods (GANomaly, f-AnoGAN) demonstrate some improvements in detection accuracy but still struggle to capture multi-scale anomaly features comprehensively. Deep-SVDD, which relies on a single feature center for anomaly measurement, demonstrates comparatively lower AUROC and AUPRC scores. KD_AD, both incorporating knowledge distillation or multi-branch architectures, exhibits better discriminative ability but still faces challenges in modeling long-range dependencies and detecting global logical anomalies. PatchCore and RD4AD, which perform well in industrial anomaly detection, show certain advantages but face difficulties in identifying the more intricate OCT anomalies. SCVAE-AC and EfficientAD perform relatively well across the metrics. Nevertheless, our proposed method achieves further improvement, obtaining

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT MODEL CONFIGURATIONS ON RETINAL OCT DATASETS

Module Configuration				Retinal OCT 2017		Retinal OCT 2022	
Conv Upsample	Mamba UpNet	Multi-Feature Model	SS2D	AUROC	ACC	AUROC	ACC
✓				96.93	93.65	95.26	88.87
✓	✓			98.05	96.18	96.29	90.96
✓	✓		✓	99.02	97.70	96.86	91.63
	✓	✓		97.84	96.02	95.78	90.88
	✓	✓	✓	98.36	96.66	96.57	91.33
✓	✓	✓		99.27	98.41	97.19	92.08
✓	✓	✓	✓	**99.72**	**98.60**	**97.42**	**92.36**

97.42%, 97.56%, and 92.36% on these three metrics, respectively, outperforming all compared methods. The dual-heterogeneous knowledge distillation framework effectively avoids the identity mapping problem by introducing architectural discrepancies between the teacher and student networks. In this teacher-student paradigm, the student network not only inherits the teacher's representational capabilities but also enhances them with specialized decoders, resulting in robust performance on both structural and global anomaly detection tasks. This confirms the efficacy and robustness of our approach in handling diverse types of anomalies in retinal OCT imaging scenarios. Fig. 4 presents our results of abnormal localization on retinal OCT 2022, from which it can be seen that our method is able to locate abnormalities on the retina more accurately compared to EfficientAD, yielding optimal results.

Cross-Modality Evaluation on CT, MRI, and PET Datasets.

To further evaluate the generalization capability of the proposed framework beyond retinal OCT images, we extended our experiments to three additional modalities: chest CT, brain MRI, and whole-body PET. All experiments followed the same unsupervised training protocol as in the OCT setup, where only normal samples were used for model training. For fairness and consistency, a one-model-per-category setting was maintained, and the same input resolution and preprocessing pipeline were applied across all datasets. Table IV summarizes the AUROC and F1-scores of representative anomaly detection baselines and our proposed model across the three modalities. The proposed method consistently achieves superior or comparable performance, reaching AUROC scores of **87.94%** on CT, **96.38%** on MRI, and **86.96%** on PET datasets. These results not only surpass strong baselines such as PatchCore and RD4AD but also demonstrate remarkable cross-modal robustness without any re-training or fine-tuning.

The AUROC for the CT dataset was relatively lower compared to OCT and MRI. This can be attributed to two main factors. First, CT images typically have lower soft-tissue contrast and higher noise levels than OCT and MRI, making subtle abnormalities harder to capture in an unsupervised anomaly detection setting. Second, the CT dataset contains multiple types of pathologies with substantially different appearance patterns and difficulty levels, which increases intra-dataset variability and further challenges anomaly detection performance. Notably, the performance gain is attributed to the proposed dual-heterogeneous distillation design, which effectively mitigates isomorphic mapping between teacher and student networks, thereby enhancing feature diversity and anomaly separability. Furthermore, the complementary local and global representation modeling through the Conv Upsample and Mamba UpNet decoders contributes to improved lesion localization and robustness under different anatomical contexts. Together, these findings highlight that the proposed method generalizes effectively across heterogeneous imaging modalities, underscoring its potential for broad clinical applicability and cross-domain deployment.

TABLE VI
PERFORMANCE COMPARISON OF PRE-TRAINED ENCODER BACKBONES ON RETINAL OCT DATASETS

Backbone \ Dataset	Retinal OCT 2017		Retinal OCT 2022	
	AUROC	ACC	AUROC	ACC
ResNet18	98.17	97.76	97.02	92.06
ResNet34	98.64	97.95	97.15	91.96
ResNet50	99.02	98.12	97.16	92.16
ResNeXt50	99.65	98.54	97.38	92.49
WideResNet50	99.72	98.60	97.42	92.36

V. DISCUSSION

A. Ablation Study

An extensive ablation study was conducted on the Retinal OCT 2017 and Retinal OCT 2022 datasets to evaluate the contribution of each major component in the proposed framework. The results of the experiment are shown in Table V. Four primary configurations were examined: (1) the baseline model employing only the Conv Upsample decoder; (2) Conv Upsample combined with the Mamba UpNet; (3) Mamba UpNet integrated with the Multi-Feature Model (MFM); and (4) the full model incorporating all three components. It is worth noting that SS2D, a key submodule within the Mamba UpNet, was also individually analyzed to quantify its contribution to long-range context modeling.

The baseline configuration achieves AUROC / ACC scores of 96.93% / 93.65% on Retinal OCT 2017 and 95.26% / 88.87% on Retinal OCT 2022. Introducing the Mamba UpNet alongside the Conv Upsample decoder leads to a

marked improvement, indicating that the selective state-space modeling capability of Mamba effectively enhances global anomaly recognition. When the Mamba UpNet is paired with the Multi-Feature Model (without the Conv Upsample branch), the model maintains competitive performance (AUROC / ACC = 98.36% / 96.66% on Retinal OCT 2017 and 96.57% / 91.33% on Retinal OCT 2022), demonstrating the benefit of multi-scale feature fusion for precise structural anomaly detection. The complete configuration—integrating Conv Upsample, Mamba UpNet, and MFM—achieves the highest performance across both datasets, with AUROC / ACC values of 99.72% / 98.60% on Retinal OCT 2017 and 97.42% / 92.36% on Retinal OCT 2022. This consistent improvement confirms that each component contributes complementary strengths: the Conv Upsample branch captures localized fine details, Mamba UpNet (empowered by SS2D) models long-range dependencies efficiently, and the Multi-Feature Model fuses hierarchical representations for balanced local–global learning. Their combined effect enables the student network to better approximate the teacher's representational manifold while maintaining heightened sensitivity to diverse types of anomalies. This synergy substantiates both the empirical performance gains and the theoretical rationale of the proposed dual-heterogeneous framework.

Further, the dedicated SS2D ablation within the Mamba UpNet verifies its critical role in efficient global dependency modeling—its removal leads to a measurable decline in AUROC, underscoring the necessity of this module within the overall decoder architecture. The synergistic integration of these components allows the student network to more accurately mimic the teacher network's representation while simultaneously addressing the diverse nature of anomalies, thus corroborating both the empirical improvements and the theoretical motivations for our proposed approach.

The feature residual distribution between the teacher and student networks inherently provides a statistical interpretation of anomaly likelihood. Given that normal features follow a compact distribution learned from normal-only data, deviations with higher residual magnitudes correspond to lower probability density regions, which implicitly quantify the uncertainty and rarity of the observed features. This perspective aligns with probabilistic formulations in density-based anomaly detection and provides an intuitive explanation for how our model discriminates between normal and abnormal patterns without requiring an explicit Bayesian formulation.

The choice of encoder architecture in the student network plays a pivotal role in hierarchical feature extraction and consequently affects anomaly sensitivity. In our framework, the encoder is responsible for capturing multi-scale representations of retinal images, which is crucial for distinguishing subtle pathological deviations from normal anatomy. We evaluate the performance of various pre-trained encoder backbones on two Retinal OCT datasets, as summarised in Table VI. In this study, the student network's encoder is initialized with five distinct architectures—ResNet18, ResNet34, ResNet50, ResNeXt50, and WideResNet50—to investigate the influence of hierarchical feature quality on anomaly detection performance. Each of these backbones provides a different capacity and style of feature learning: deeper ResNets offer more layers for abstraction, ResNeXt introduces parallel feature groups (cardinality) within residual blocks, and WideResNet expands the width (number of channels) per layer.

Our experimental results clearly indicate that increasing the complexity of the backbone architecture leads to incremental improvements in performance. For example, ResNet18 attains an AUROC of 98.17% and an ACC of 97.76% on the Retinal OCT 2017 dataset, while progressively more advanced models such as ResNet34 and ResNet50 show superior metrics. The WideResNet50, the most powerful among the evaluated backbones, achieves the highest performance, underscoring its enhanced capacity to capture subtle variations and complex patterns inherent in retinal images. The more sophisticated encoder like WideResNet50 provides richer, more discriminative representations that enable the decoders to operate more effectively, thereby mitigating the identity mapping issue common in conventional architectures. This synergy between a robust encoder and the subsequent decoders is fundamental to the improved anomaly detection performance observed across the datasets.

TABLE VII
COMPUTATIONAL EFFICIENCY ANALYSIS EXPERIMENTS ON THE RETINAL OCT 2017 DATASET

Method	AUROC	Number of Parameters ($\times 10^6$)	FLOPs ($\times 10^9$)
Baseline+FCN	97.45	46.6	594
Baseline+Transform	99.23	69.7	775
Baseline+Conv	98.36	26.2	265
Baseline+Mamba(Ours)	99.72	30.9	285

B. Complexity Analysis

To systematically evaluate the computational efficiency of the proposed Mamba UpNet within the dual-decoder framework, we conducted a controlled component replacement study on the Retinal OCT 2017 dataset. In this experiment, the Mamba UpNet was individually replaced with alternative decoder structures—namely a fully connected network (FCN), a Transformer block, and a standard convolutional (Conv) branch—while keeping the remainder of the network and training settings identical. The comparison was performed in terms of AUROC, the number of trainable parameters, and floating-point operations (FLOPs). All FLOPs were computed with an input resolution of 256×256 using the same profiling tool and conventions for fair comparison, excluding pre- and post-processing costs. As summarized in Table VII, the Baseline+Transformer configuration achieved a competitive result, yet at the cost of substantial computational demand, requiring 69.7×10^6 parameters and 775×10^9 FLOPs. The Baseline+FCN variant produced moderate accuracy (97.45%) but incurred the highest FLOPs, demonstrating that dense fully connected operations on high-dimensional features are computationally inefficient. The Baseline+Conv design significantly reduced both parameters and FLOPs but exhibited

limited accuracy (98.36%), indicating that purely convolutional decoders fail to capture global context effectively. By contrast, our Baseline+Mamba configuration achieved the best overall balance—yielding the highest AUROC (99.72%) with only 30.9×10^6 parameters and 285×10^9 FLOPs—thus achieving superior accuracy–efficiency trade-offs. The FCN variant flattens spatial features into a dense vector, resulting in extremely high-dimensional matrix multiplications and prohibitive FLOPs. In contrast, convolutional decoders maintain low local complexity but require multiple layers or enlarged kernels to approximate long-range dependencies, which limits representational richness. Mamba, leveraging selective state-space modeling, captures long-range dependencies through recurrent and convolutional operations with linear complexity in sequence length. This design allows Mamba to retain the representational benefits of global attention while maintaining computational efficiency closer to convolutional decoders.

Despite the efficiency advantages of the Mamba UpNet, several practical limitations remain. The reported FLOPs and parameter counts are theoretical indicators—they do not directly translate to real-world inference latency or energy consumption, which depend on GPU kernel efficiency, memory bandwidth, and batch-level parallelism. The present evaluation uses a one-model-per-category setting; while suitable for benchmarking, such design may increase storage and maintenance overhead when scaling to multi-class clinical systems.

In summary, the experimental and analytical results consistently demonstrate that Mamba provides a superior balance between modeling capacity and computational efficiency. It effectively captures global contextual dependencies at substantially lower complexity than Transformer-based alternatives and achieves higher accuracy than purely convolutional designs. These findings validate the suitability of Mamba as a decoder backbone for large-scale retinal OCT anomaly detection and motivate further research into optimizing memory footprint and real-time deployment strategies for clinical applications.

VI. Conclusion

In this paper, we propose a novel hybrid dual-heterogeneous knowledge distillation network for anomaly detection in retinal OCT images. The proposed method addresses key challenges in existing anomaly detection techniques, particularly the identity mapping problem and the difficulty in detecting diverse anomaly types in retinal OCT imaging scenarios. Furthermore, cross-modality evaluations on chest CT, brain MRI, and PET datasets confirmed the generalization capability of the proposed framework, highlighting its robustness and transferability across diverse imaging modalities. Looking forward, we aim to develop a unified and generalizable framework for medical anomaly detection capable of handling multiple imaging modalities within a single model. Such a universal paradigm is expected to enhance cross-domain adaptability and reduce dependence on large-scale annotated datasets, while maintaining high sensitivity and specificity. Integrating uncertainty estimation and probabilistic anomaly scoring mechanisms will further improve interpretability and clinical reliability. Ultimately, our contributions lay a solid foundation for advancing toward a universal, clinically deployable anomaly detection system that can support efficient, automated, and reliable analysis across diverse medical imaging scenarios.

VII. Acknowledgments

This work was supported in part by the Youth Project of Natural Science Foundation of Shandong Province, China (ZR2023QC262); the National Natural Science Foundation of China (62205181); the Natural Science Foundation of Shandong Province (ZR2022QF017); and the Shandong Province Outstanding Youth Science Fund Project (Overseas) (2023HWYQ-023); the Key R&D Program of Shandong Province, China (2024CXGC010106); the Taishan Scholar Foundation of Shandong Province (tsqn202211038).

VIII. References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[2] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, and C. Adak, "Unsupervised anomaly detection for surface defects with dual-siamese network," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7707–7717, 2022.

[3] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3694–3706, 2020.

[4] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, "Optical coherence tomography," *science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[5] Z. Chen, H. Wang, C. Ou, and X. Li, "Mutri: Multi-view tri-alignment for oct to octa 3d image translation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 885–20 894.

[6] B. E. Bouma, J. F. de Boer, D. Huang, I.-K. Jang, T. Yonetsu, C. L. Leggett, R. Leitgeb, D. D. Sampson, M. Suter, B. J. Vakoc *et al.*, "Optical coherence tomography," *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 79–79, 2022.

[7] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, 2019.

[8] K. Zhou, J. Li, W. Luo, Z. Li, J. Yang, H. Fu, J. Cheng, J. Liu, and S. Gao, "Proxy-bridged image reconstruction network for anomaly detection in medical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 582–594, 2021.

[9] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "Encoder-decoder contrast for unsupervised anomaly detection in medical images," *IEEE transactions on medical imaging*, vol. 43, no. 3, pp. 1102–1112, 2023.

[10] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2019, pp. 622–637.

[11] Y. Zhang, Z. Chen, and X. Yang, "Light-m: An efficient lightweight medical image segmentation framework for resource-constrained iomt," *Computers in Biology and Medicine*, vol. 170, pp. 108 088–108 088, 2024.

[12] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 318–14 328.

[13] T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen, "Panda: Adapting pretrained features for anomaly detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2806–2814.

[14] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4183–4192.

[15] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 902–14 912.

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

[17] S. Lu, W. Zhang, H. Zhao, H. Liu, N. Wang, and H. Li, "Anomaly detection for medical images using heterogeneous auto-encoder," *IEEE Transactions on Image Processing*, vol. 33, pp. 2770–2782, 2024.

[18] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 128–138.

[19] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 4393–4402.

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

[21] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.

[22] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157.

[23] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 2016, pp. 2360–2368.

[24] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations (ICLR)*, 2018.

[25] P. Samangouei, "Defense-gan: protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[26] H. Guo, L. Ren, J. Fu, Y. Wang, Z. Zhang, C. Lan, H. Wang, and X. Hou, "Template-guided hierarchical feature restoration for anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6447–6458.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[29] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

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

[31] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[32] H. Zhao, Y. Li, N. He, K. Ma, L. Fang, H. Li, and Y. Zheng, "Anomaly detection for medical images using self-supervised and translation-consistent features," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3641–3651, 2021.

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

[34] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[35] X. Zhou, S. Niu, X. Li, H. Zhao, X. Gao, T. Liu, and J. Dong, "Spatial–contextual variational autoencoder with attention correction for anomaly detection in retinal oct images," *Computers in biology and medicine*, vol. 152, pp. 106 328–106 328, 2023. [Online]. Available: https://pan.baidu.com/s/142fzaj_Q1DyKVh-SlMqlrQ

[36] A. Hamada, "Br35h :: Brain tumor detection 2020," 2025. [Online]. Available: https://dx.doi.org/10.21227/tbkk-q937

[37] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer, 2019, pp. 161–169.

[38] K. Zhou, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, Z. Gu, J. Liu, and S. Gao, "Encoding structure-texture relation with p-net for anomaly detection in retinal images," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 360–377.