Anomaly Detection for Medical Images Using Heterogeneous Auto-Encoder

Shuai Lu, Weihang Zhang, He Zhao, Hanruo Liu, Ningli Wang, and Huiqi Li, *Senior Member, IEEE*

***Abstract*— Anomaly detection is an important task for medical image analysis, which can alleviate the reliance of supervised methods on large labelled datasets. Most existing methods use a pixel-wise self-reconstruction framework for anomaly detection. However, there are two challenges of these studies: 1) they tend to overfit learning an identity mapping between the input and output, which leads to failure in detecting abnormal samples; 2) the reconstruction considers the pixel-wise differences which may lead to an undesirable result. To mitigate the above problems, we propose a novel heterogeneous Auto-Encoder (Hetero-AE) for medical anomaly detection. Our model utilizes a convolutional neural network (CNN) as the encoder and a hybrid CNN-Transformer network as the decoder. The heterogeneous structure enables the model to learn the intrinsic information of normal data and enlarge the difference on abnormal samples. To fully exploit the effectiveness of Transformer in the hybrid network, a multi-scale sparse Transformer block is proposed to trade off modelling long-range feature dependencies and high computational costs. Moreover, the multi-stage feature comparison is introduced to reduce the noise of pixel-wise comparison. Extensive experiments on four public datasets (*i.e.*, retinal OCT, chest X-ray, brain MRI, and COVID-19) verify the effectiveness of our method on different imaging modalities for anomaly detection. Additionally, our method can accurately detect tumors in brain MRI and lesions in retinal OCT with interpretable heatmaps to locate lesion areas, assisting clinicians in diagnosing abnormalities efficiently.**

***Index Terms*— Anomaly detection, medical images, auto-encoder, heterogeneous network.**

I. Introduction

ANOMALY detection has attracted much attention in the medical image analysis community [1], [2], [3]. Although some supervised learning methods (*e.g.* ResNet [4]) have shown impressive performance in many computer vision tasks [5], [6], [7], they rely on large-scale natural datasets (*e.g.* ImageNet [8]). Unlike natural images, medical images with annotations by clinicians are difficult to obtain, especially for some rare diseases, which limits the performance of supervised learning in medical diagnosis [1], [9]. In addition, it is easy to collect normal medical data from healthy subjects compared with the expensive cost of annotating lesions. Therefore, some methods using only normal data are proposed for anomaly detection.

The existing anomaly detection methods can be divided into two groups: image reconstruction-based approaches [10], [11], [12], [13], [14], [15], [16] and non-reconstruction-based approaches [17], [18], [19], [20]. In this work, we focus on the reconstruction-based approach, in which the basic idea is to learn a mapping that can reconstruct normal data with a small error but a larger error for abnormal data. Usually, the pixel-wise reconstruction error of input and output is used as an indicator to measure the severity of anomaly in the test phase. However, this kind of method suffers from the negative effects of identity mapping [1] and poor anomaly localization, which limit the detection performance and the reliability of the model. Identity mapping means the model simply reproduces the inputs as outputs without learning the core features for reconstructing normal samples. When unseen samples are used for testing, the model simply outputs a replication of the input without understanding the semantic information. Therefore, the anomalous samples cannot be detected due to the small reconstruction error between the input and output caused by the identity mapping. As shown in Fig. 1(a), the auto-encoder (AE) fails to detect the anomalies because the model generates an identical output of abnormal images. Furthermore, pixel-wise differences between input and output can affect anomaly localization analysis. These methods commonly use the Mean Squared Error (MSE) as the reconstruction error to evaluate the RGB values of the corresponding pixel pairs at the input and output. The pixel value disturbance in the reconstructed image will increase the reconstruction error. This pixel-wise comparison makes the heatmap sensitive to noise, limiting the model's reliability. As shown in Fig. 1(a), although the SALAD method [9] achieves satisfactory detection results, the anomaly localization analysis is influenced by pixel-wise noise. Basic denoising operations such as mean and Gaussian filtering show ineffectiveness in addressing the issue, as these operations may reduce the errors associated with anomalous samples. It implies that pixel-level analysis may be too

Manuscript received 19 July 2023; revised 3 February 2024; accepted 10 March 2024. Date of publication 29 March 2024; date of current version 10 April 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 82072007, in part by Beijing Natural Science Foundation under Grant IS23112, and in part by Beijing Institute of Technology Research Fund Program for Young Scholars. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Stuart Weller. *(Corresponding authors: He Zhao; Huiqi Li.)*

Shuai Lu, Weihang Zhang, and Huiqi Li are with Beijing Institute of Technology, Beijing 100081, China (e-mail: huiqili@bit.edu.cn).

He Zhao is with Beijing Institute of Technology, Beijing 100081, China, and also with the Department of Eye and Vision Sciences, University of Liverpool, L7 8TX Liverpool, U.K. (e-mail: zhaohe@bit.edu.cn).

Hanruo Liu is with Beijing Institute of Technology, Beijing 100081, China, and also with Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China.

Ningli Wang is with Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing 100054, China, and also with Henan Province Academy of Medical Sciences, Zhengzhou 450046, China.

Digital Object Identifier 10.1109/TIP.2024.3381435

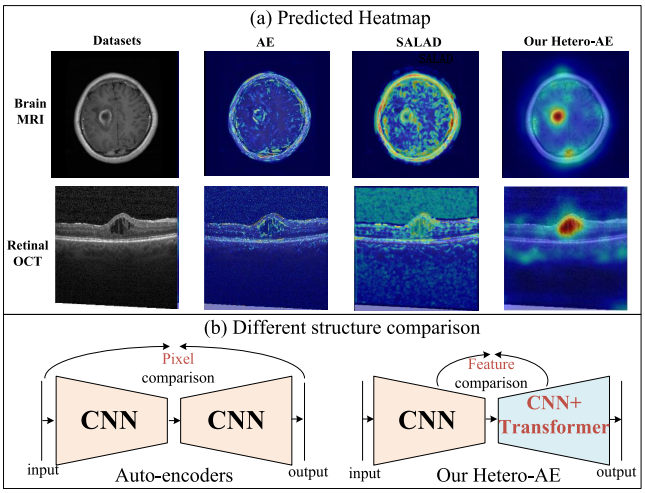

Fig. 1. (a) Heatmaps indicate anomalous areas on brain MRI and retinal OCT datasets. The red in the heatmap denotes regions where the predictions are more likely to be anomalies. (b) Different structure comparison.

sensitive to distinguish between noise and subtle anomalous patterns effectively, which motivates us to utilize a multi-level feature comparison to mitigate the negative impact of identity mapping and pixel-level noise.

The motivations for the two improvements are described in the following aspects. **1) The encoder and decoder with heterogeneous structures can easily generate large reconstruction errors for the unseen anomalous data.** The term "heterogeneous" emphasizes the structural differences between the encoder and decoder. Some existing reconstruction-based approaches [9], [21], [22] employ CNN or Transformer structures for both encoder and decoder, which makes anomalous samples well-reconstructed and hence difficult to detect. We hope that the structural differences can lead to greater feature differences between the encoder and decoder on anomalous samples compared with normal samples. CNN and Transformer are two different structures and thus have different inductive biases [23]. To make the reconstruction of anomalous samples difficult, a preferred choice is to utilize a CNN architecture for the encoder and a hybrid CNN and Transformer architecture for the decoder. We hope that the CNN part of decoder can effectively reconstruct and fit the features of encoder on normal samples, while the Transformer primarily serves to generate features with much larger differences on unseen abnormal samples. Additionally, Transformers can lead to high computational cost. To alleviate the quadratic computational complexity [24] of the Transformer, we design a multi-scale sparse Transformer block to trade off the capability of modelling features and the high computational cost. **2) Feature comparisons are more robust for anomaly localization than pixel-wise differences.** Reconstruction-based models generally compute pixel-wise differences for anomaly localization analysis, which are affected by pixel shifts or perturbations. Compared to the original image, the movement of pixels and changes in values in the reconstructed image can generate greater reconstruction errors. This increases the likelihood that normal areas might be misidentified as anomalous regions. The above problems can be alleviated by comparing mid-level features with a larger number of channels. Mid-level features are aggregations of local pixels, representing local information rather than single-pixel information. The comparison of mid-level features can reduce the impact of pixel-wise perturbations.

The main contributions of the paper are summarized as follows:

1) To the best of our knowledge, the heterogeneous auto-encoder is firstly proposed for anomaly detection in medical images, where a vanilla ResNet network is used as the backbone of the encoder and a novel hybrid CNN-Transformer network is designed as the backbone of the decoder. Different inductive biases between CNN and Transformer help the auto-encoder to generate larger reconstruction errors on anomalous samples.
2) In the hybrid CNN-Transformer network, a Multi-scale Sparse Transformer Block (MSTB) is proposed to model the relationship between local and regional information, where regional data encompass multiple local data points. In this case, the MSTB trades off modelling long-range feature dependencies and high computational costs.
3) Unlike auto-encoder-based methods compute pixel-wise loss based on the final output of the network only, our method expands the pixel-wise loss computation. We calculate the differences not just at the final output but also across multiple intermediate layer features, enhancing the robustness of our model against pixel perturbations.

II. Related Work

A. Image Reconstruction-Based Approaches

The basic idea of image reconstruction-based approaches is with the assumption that the model can generate a well-reconstructed output for normal data while the abnormal ones will not be well reconstructed. The difference between the input and reconstructed images is used as an indicator to detect anomaly. Specifically, Chen et al. [25] utilized a convolutional AE to generate a low-dimensional representation and reconstruction error for each input data. Further methods based on variational auto-encoder (VAE) were used for image reconstruction. Lu et al. [26] utilized VAE for skin disease detection. Zimmerer et al. [11] proposed to combine a context encoder and a VAE for brain MRI image reconstruction. In addition, more methods [12], [13] tried to combine AE with generating adversarial network for anomaly detection in MRI images. Deecke et al. [14] used multiple points to reconstruct the test image, in which the latent variables and internal parameters of the generator were optimized in the iterative optimization process to improve the quality of image reconstruction. AnoGAN [2] generated synthetic samples similar to the test sample from the latent space and performed anomaly detection by calculating the difference between the test sample and the generated synthetic sample. To improve the inference speed of AnoGAN, a fast version called f-AnoGAN [3] was proposed, where an encoder was trained to map images to the latent space for fast inference.

GANormaly [27] used the encoder to map the original image and the generated reconstruction results into the latent space, and then compared the feature differences between the two latent spaces. Zhou et al. [1] proposed a proxy-bridged image reconstruction network for anomaly detection in medical images. Zhao et al. [9] proposed a method that combined a reconstruction framework and adversarial learning strategy, which learns the manifold of normal data through an encode-and-reconstruct translation between image and latent spaces. Wolleb et al. [28] proposed a weakly supervised anomaly detection method using denoising diffusion implicit models (DDIMs). The method combined iterative stochastic noising and deterministic denoising with classifier guidance to translate images from diseased to healthy subjects. SQUID [29] was proposed for detecting anomalies in chest X-ray, which is a GAN-based in-painting network with a memory bank. Dual-distribution Discrepancy for Anomaly Detection (DDAD) was proposed in [30] to improve anomaly detection in medical images. DDAD utilized both labeled normal images and unlabeled images containing anomalies during training. EDC [31] combined reconstruction and contrastive learning to optimize a pretrained encoder for the target domain without collapsing, and used a new global cosine distance loss to stabilize training. Li et al. [32] proposed an unsupervised anomaly detection framework called SSL-AnoVAE for retinal images. SSL-AnoVAE utilized a self-supervised learning module to obtain semantic prior information and concatenated representations from the SSL module and the encoder for improved image reconstruction and anomaly detection. UniAD [33] and related methods [34], [35] obtained multi-level features from the encoder as input to the decoder, and the decoder performed self-reconstruction on the obtained multi-level features. This direct self-reconstruction often faces the issue of identity mapping. To mitigate this problem, UniAD introduced masking and noise into the self-reconstruction process. Our approach takes the compressed bottleneck features of the encoder as input to reconstruct multi-scale hierarchical features, which is more challenging for feature reconstruction, and thus can suppress identity mapping.

B. Non-Reconstruction-Based Approaches

Non-reconstruction-based approaches can be roughly divided into two categories: one-class classification-based approaches and deep feature embedding-based approaches. The objective of one-class classification-based approaches [17], [18], [19] is to make the model generate a decision boundary for separating the negative samples during inference, but only the normal samples are provided during the training process. However, these one-class models may fail for normal classes with complex distributions.

Deep feature embedding-based approaches mainly generate abnormal feature maps by comparing the feature differences between target images and normal images. Knowledge distillation is an important method for deep feature embedding. The concept of knowledge distillation [36] was first proposed by Hinton. Uninformed students [37] was the first to use a knowledge distillation model for anomaly detection. It used multiple decoders to learn the output of an encoder network and detected abnormal images by comparing the differences between the outputs of multiple networks. Further, a multi-stage feature distillation strategy [20], [38] was introduced into the distillation framework. Deng and Li [39] introduced a reverse distillation paradigm into the teacher-student framework to enrich the representation of anomalies. Tian et al. [40] proposed a self-supervised pre-training method called PMSACL for unsupervised anomaly detection in medical images. PMSACL used a contrastive learning approach to discriminate between normal images and pseudo abnormal images synthesized via data augmentation. Liu et al. [41] proposed a student-teacher network with skip connections (Skip-ST) which was trained by a novel knowledge distillation paradigm called direct reverse knowledge distillation (DRKD) to realize anomaly detection.

Different from above methods, our method utilizes a heterogeneous structure for feature comparison of AE, which can increase the feature difference between the encoder and decoder on abnormal samples. Our proposed "heterogeneous" emphasizes the structural differences between the encoder and decoder. However, existing KD-based methods [20], [39] achieved feature differences between teacher and student networks by leveraging differences in parameter count or data flow direction between teacher and student networks.

III. Methodology

A. Overview

1) Problem Formulation: The training dataset with only n normal samples is denoted as $D_{train} = \{x_i \mid x_i \in \mathcal{X} = \mathbb{R}^{H\times W\times C}\}$, where x_i represents a medical image in the input space $\mathcal{X}$. The testing dataset with m normal and abnormal images is denoted as $D_{test} = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, where y_i represents the label of x_i in the output space $\mathcal{Y} = \{0, 1\}$, 0 denotes normal image and 1 denotes abnormal image. The goal is to use the training set only with normal samples $\{x_1, \ldots, x_i, \ldots, x_n\}$, to build a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ to detect abnormal samples in the testing set.

2) Framework Overview: To achieve the above goal, we propose an unsupervised heterogeneous Auto-Encoder (Hetero-AE) for anomaly detection in medical images, as shown in Fig. 2(a). Our key insight is to build a framework that enables the encoder and decoder to have similar features on normal medical images but large feature differences on anomalous images. Specifically, the proposed Hetero-AE consists of an encoder E and a decoder D, where F_E^i and F_D^i represent the features from their feature spaces, respectively. A medical image $x_i \in D_{train}$ is first fed into the encoder E, whose backbone is a classification network. Further, the final deep features F_E^4 extracted by the encoder are used as the input of decoder D. It is easy for the decoder to restore features of normal samples from the compressed deep feature F_E^4, while difficult to recover features of abnormal samples during inference. When there is a large difference between the features F_D^i from the decoder and the corresponding features F_E^i from the encoder, the test sample is more likely to be a medical image with anomalies. Thus, anomaly detection can be achieved by computing the feature distance between the encoder and decoder.

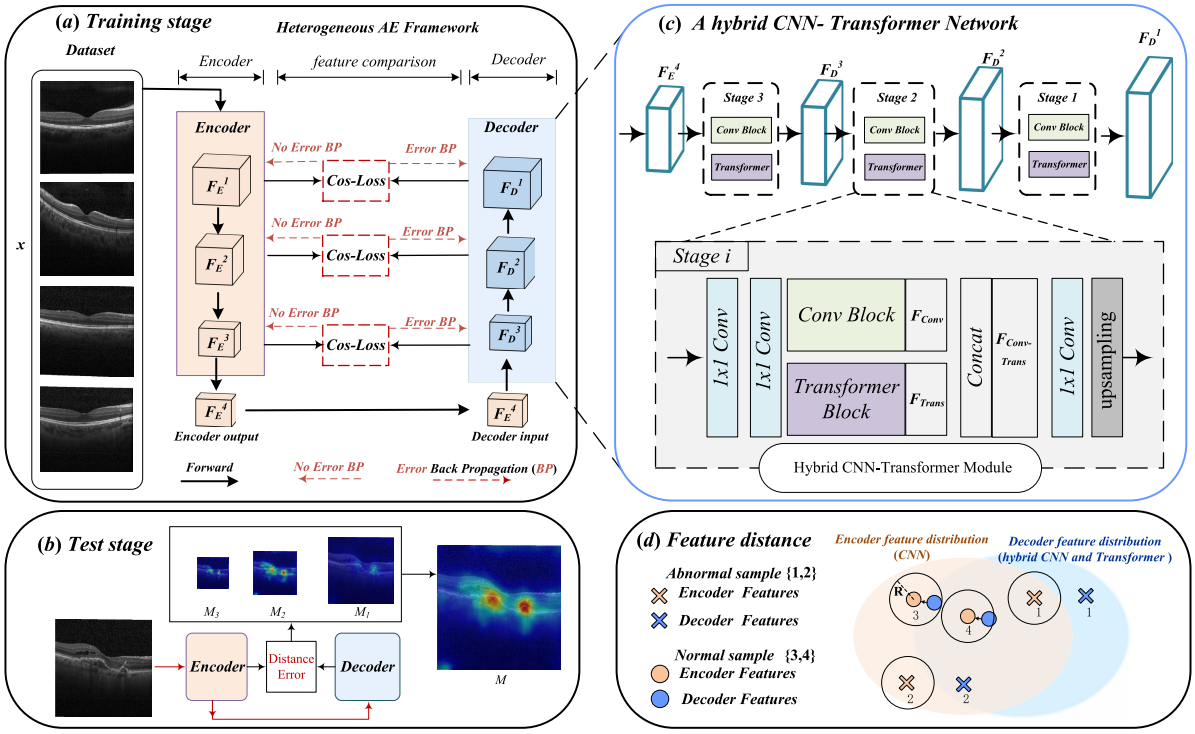

Fig. 2. Overall architecture of our Hetero-AE. First, the original image is sent to an encoder for feature compression. Then, the compressed feature with the smallest scale (F_E^4) is used as the input of a decoder. The decoder gradually recovers the compressed features and imitates the corresponding features of the encoder. (a) The training process of the model. (b) The testing process of the model. (c) The proposed decoder is based on a hybrid CNN-Transformer module. (d) Schematic representation of feature distances. Features from the encoder and decoder have large distance on abnormal samples and small distance on normal samples. Samples 1 and 2 represent abnormal samples while samples 3 and 4 represent normal samples, where each sample has two types of features (*i.e.*, encoder features and decoder features). The feature distance between the encoder and decoder is larger on abnormal samples than on normal samples.

Intuitively, if the decoder and encoder have larger feature differences on anomalous images, the anomaly detection performance will be further improved. In this case, our encoder and decoder are designed with different inductive biases. Since CNNs are extensively employed for image-related tasks, numerous pre-trained models on large datasets are available. Utilizing a pre-trained model can serve as a beneficial starting point for our application. Thus, we utilize CNN structures for our encoder, while the decoder uses a hybrid CNN-Transformer network (HCTN) as the backbone. Different inductive biases are able to make a great divergence in the feature distributions of encoder and decoder on abnormal medical images.

B. Heterogeneous Auto-Encoder Framework

1) Encoder: The main goal of an encoder is used to compress information, in which only the features related to normal samples are preserved through feature extraction. The encoder takes the classification network pre-trained on ImageNet as the backbone. Although the pre-trained encoder cannot extract specific medical features, it has the ability to extract features similar to natural images, such as the edges of objects. The structure of the encoder is described as follows. The simple ResNet [4] is chosen as the encoder backbone in our model. Different from the original ResNet, the encoder only uses the first four stages of ResNet, in which the pre-trained weights on ImageNet are used for initialization. In the training phase, only normal medical images are sent to the encoder, of which the four stages will sequentially generate hierarchical features $F_E = \{F_E^1, F_E^2, F_E^3, F_E^4\}$, as shown in Fig. 2(a). The features F_E will be used for feature reconstruction of the decoder.

2) Decoder (Hybrid CNN-Transformer Network): To alleviate the negative impact of identity mapping on anomaly detection performance of AEs, a hybrid decoder is proposed to make the model generate a large reconstruction error on anomalous samples. The input of the decoder is the final deep features $F_E^4 \in \mathbb{R}^{H_4 \times W_4 \times C_4}$ of the encoder. The decoder decodes F_E^4 sequentially to scales consistent with the encoder's hierarchical features. As shown in Fig. 2(a), the proposed decoder generates hierarchical features $F_D = \{F_D^3, F_D^2, F_D^1\}$. The core component in the decoder is the proposed Hybrid CNN-Transformer Module.

As shown in Fig. 2(c), the hybrid CNN-Transformer module contains two parallel sub-blocks (*i.e.*, the convolution block and the Transformer block) to extract features. Specifically, it first adjusts the dimension of the input feature F_E^4 to generate two connected groups through two 1×1 convolutions. The first group of features is fed into the convolution block to generate feature F_{Conv}. The second group of features is fed into the Transformer block to generate feature F_{Trans}. Then, F_{Conv} and F_{Trans} are concatenated together to generate the feature $F_{Conv-Tran}$, and the channel number of $F_{Conv-Tran}$ is adjusted by 1×1 convolution. An upsampling layer is finally used to double the scale of the features. The convolution block

in the hybrid CNN-Transformer module consists of two consecutive convolutions with a kernel size of 3×3 and a stride of 1×1. The Transformer block in the hybrid CNN-Transformer module is an improved version called Multi-scale Sparse Transformer Block (MSTB), which is described in detail as below.

C. Multi-Scale Sparse Transformer Block (MSTB)

The computational cost of the Transformer is positively correlated with the sequence lengths of Q (queries), K (keys), and V (values). To mitigate this, we decrease the sequence lengths of K and V to reduce the computational cost. By reducing the length of K and V from HW to $\frac{HW}{16}$ and $\frac{HW}{64}$ in the two branches of the regional information, we achieve a significant reduction in computational cost, decreasing from $\Omega\left(2\left(HW\right)^2 C + 4HWC^2\right)$ to $\Omega\left(0.15\left(HW\right)^2 C + 3.75HWC^2\right)$ with nearly 90% reduction. To achieve the above purpose, MSTB models the relationship between regional and local information as shown in Fig. 3, where regional data encompass multiple local data points. Following the design of vision Transformer (VIT) [23], MSTB mainly includes patch embedding, multi-head attention and multilayer perceptron. Different from VIT, the patch embedding of MSTB has multiple scales. According to the patch scale, the patch embedding of MSTB can be divided into two types: local information and regional information.

Local information $F_{local} \in \mathbb{R}^{(H_i \cdot W_i) \times C_i}$ is obtained by flattening input features. Specifically, the input feature $F \in \mathbb{R}^{H_i \times W_i \times C_i}$ is reshaped into $F^1 \in \mathbb{R}^{(H_i \cdot W_i) \times C_i}$, where H_i, W_i, and C_i represent the height, width, and number of channels of the i-th stage feature map, respectively. $F^1 = \left[f_1, \ldots, f_j \ldots, f_{N_i}\right]$ represents a sequence of flattened 2D patches after reshaping, where $f_j \in \mathbb{R}^{1 \times C_i}$ denotes a feature representation in F^1 and $N_i = H_i \cdot W_i$ denotes the length of the sequence. Further, F^1 is added with a learnable position embeddings $E^{local}_{pos} \in \mathbb{R}^{(H_i \cdot W_i) \times C_i}$ to obtain F_{local}.

Compared with local information, regional information reduces the resolution of spatial features through feature patches and feature projections as shown in Fig. 3. Firstly, the input feature $F \in \mathbb{R}^{H_i \times W_i \times C_i}$ is divided into non-overlapping feature patches with dimension ($\mathbb{R}^{\frac{H_i \cdot W_i}{p^2} \times p^2 \times C_i}$) by patches of resolution (p,p), and then the non-overlapping patches are reshaped into a sequence of flattened feature patches $F^p \in \mathbb{R}^{N_p \times (p^2 \cdot C_i)}$, where $N_p = \frac{H_i \cdot W_i}{p^2}$ denotes the number of patches. As shown in Fig. 3, the multi-scale spatial reduction (SR) in regional information is realized by adjusting different $p \in \{p_1, p_2\}$, where $\frac{p_2}{p_1} = 2$. For example, $p_1 = 4$ and $p_2 = 8$ are used in stage-1 of MSTB from the hybrid network. Following VIT, F^p is projected into a new patch sequence through the matrix $E \in \mathbb{R}^{\left(p^2 \cdot C_i\right) \times C_i}$, and further the new feature sequence is summed with a learnable position embedding ($E^p_{pos} \in \mathbb{R}^{N_p \times C_i}$) to generate regional information $F^p_{region} \in \mathbb{R}^{N_P \times C_i}$.

Following the classic multi-head attention design, the proposed MSTB's attention module also contains multiple single attentions to improve feature representation. Different from classical attention, each single attention in our MSTB models the relationship between local information and regional information, which can reduce the computational cost of the classical attention. Specifically, the h group attention are realized by mapping queries (Q), keys (K), and values (V) h times through different learnable linear projections. We compute the j-th region attention inputs query Q_j, key K^p_j and, value V^p_j as an example. The local information F_{local} is linearly projected to $Q_j \in \mathbb{R}^{N_i \times d_{head}}$ with $W^Q_j \in \mathbb{R}^{C_i \times d_{head}}$, where $d_{head} = \frac{C_i}{h}$ represents the dimension of the single-head attention mechanism; meanwhile, the regional information F^p_{region} is projected linearly to $K^p_j \in \mathbb{R}^{N_p \times d_{head}}$ and $V^p_j \in \mathbb{R}^{N_p \times d_{head}}$ with $W^K_j \in \mathbb{R}^{C_i \times d_{head}}$ and $W^V_j \in \mathbb{R}^{C_i \times d_{head}}$, respectively. Query, key, and value are defined as follows:

$$\begin{cases} Q_j = F_{local} \boldsymbol{W}^Q_j, \\ K^p_j = F^p_{region} \boldsymbol{W}^K_j, \\ V^p_j = F^p_{region} \boldsymbol{W}^V_j. \end{cases} \tag{1}$$

Here the single attention of local information and regional information is calculated in Eq. (2).

$$head_j = softmax\left(\frac{\boldsymbol{Q}_j {K^p_j}^T}{\sqrt{d_{head}}}\right) \boldsymbol{V}^p_j \tag{2}$$

Here, $\frac{h}{2}$ single attentions are connected together to obtain multi-head region attention. The multi-headed attention is as follows:

$$\begin{aligned} &MHA\left(\boldsymbol{F}_{local}, \boldsymbol{F}^p_{region}\right) \\ &= concate\left(head_1, \ldots, head_j, \ldots, head_{\frac{h}{2}}\right). \end{aligned} \tag{3}$$

As shown in Fig. 3, the multi-head attention mechanisms of two scales are concatenated together. Specifically, the multi-head attention mechanism with patch p_1 ($MHA(\boldsymbol{F}_{local}, \boldsymbol{F}^{p_1}_{region})$) and p_2 ($MHA(\boldsymbol{F}_{local}, \boldsymbol{F}^{p_2}_{region})$) are concatenated together to get feature Z. Then Z is fed into layer norm (LN) and multilayer perceptrons (MLP) to enhance the representation of features.

D. Loss and Anomaly Score

1) Feature Comparison (FC) Loss: The loss of Hetero-AE contains three stages of feature differences between the encoder and decoder. $L^k(h, w)$ denotes the loss of the pixel at position (h,w) in the k-th stage, which is the combination of cosine similarity (cos) and Mean Squared Error (MSE), as described in Eq.(4).

$$\begin{aligned} L^k(h, w) = &-\alpha \cos\left(F^k_E(h, w), F^k_D(h, w)\right) \\ &+ (1 - \alpha)\, MSE\left(F^k_E(h, w), F^k_D(h, w)\right), \end{aligned} \tag{4}$$

where $F^k_E(h, w)$ and $F^k_D(h, w)$ represent one-dimensional feature vectors of the pixel at coordinates (h,w) in the k-th stage feature map of encoder and decoder, respectively. Our final feature comparison loss $\mathcal{L}_{\mathcal{FC}}$ can be described in Eq. (5).

$$\mathcal{L}_{\mathcal{FC}} = \sum_{k=1}^{3} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} L^k(h, w) \right\}, \tag{5}$$

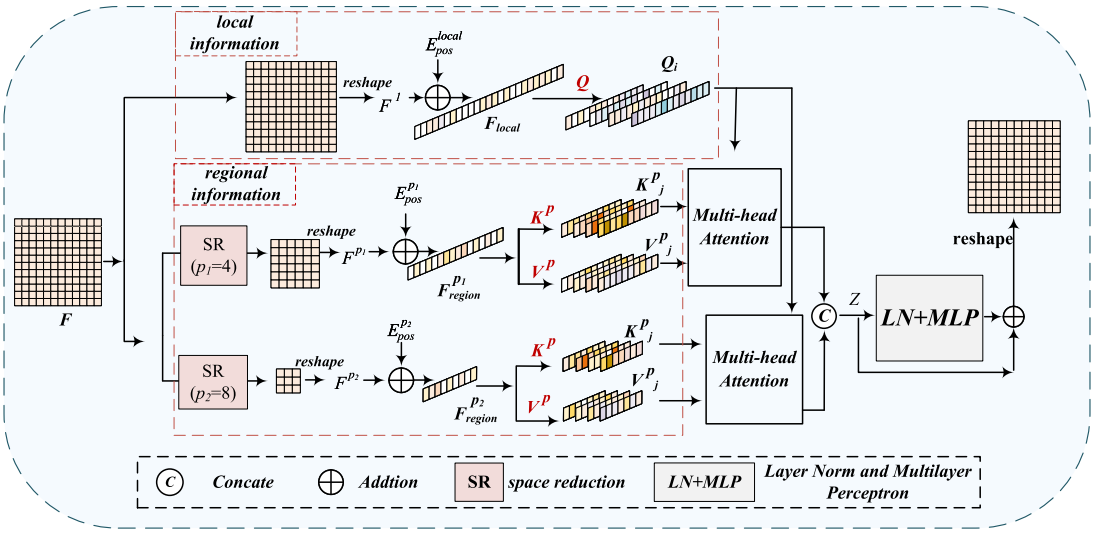

Fig. 3. The architecture of Multi-scale Sparse Transformer Block (MSTB). It consists of a local information block and a regional information block. The local information is utilized as a query to extract multi-scale regional features by the attention mechanism. The obtained features are then concatenated and input into MLP to derive the final feature representation.

where (H_k, W_k) represents the resolution of the feature output at the k-th stage.

2) Anomaly Score: Following Fig. 2, the anomaly score map is calculated based on the multi-stage feature cosine similarity. The features from encoder and decoder in different stages are resized to the resolution of input images in order to generate a final anomaly map M for anomaly localization. The equation of anomaly score map is defined as:

$$M_k(h, w) = 1 - \cos\left(F_D^k(h, w), F_E^k(h, w)\right), \tag{6}$$

$$M = \sum_k^3 Resize(M_k), \tag{7}$$

where F_D^k and F_E^k are the resized feature maps from different stages.

The anomaly score used for binary classification is formulated as follows:

$$score = \max_{h\in\{1,\ldots,H\}, w\in\{1,\ldots,W\}} M(h, w). \tag{8}$$

If this anomaly score is greater than the threshold σ, it is judged as an anomaly; otherwise, it is considered as normal. The optimal σ is the threshold value that achieves the maximal F1-score during the calculation of AUC curve. Consequently, all test images obtain a binary prediction result, classified as either anomalous or normal.

IV. Experiments

A. Datasets

1) Retinal OCT Dataset: The retinal OCT [44] dataset from Spectralis OCT (Heidelberg Engineering, German) is used in this section, and it contains four categories, including choroidal neovascularization (CNV), diabetic macular edema (DME), Rrusen, and normal. The training and test sets have been divided by the publisher for a fair comparison. The training set contains 26,315 normal images. The test set contains 250 normal images and 750 abnormal images from three diseases, including CNV, DME, and Drusen. We use the normal images in the original training set to train the model and use all the test images for performance evaluation.

2) Chest X-Ray Dataset: The Chest X-ray dataset [44] contains 5856 X-Ray images and two categories (Pneumonia/Normal). The training set has 5232 X-ray images, and the test set consists of 234 normal images and 390 pneumonia images. To train the proposed model, we only choose the normal class from the original training set, and the performance is evaluated using the full test set.

3) Brain MRI Dataset: The training datasets used in previous works of brain MRI [12], [13] are not released. A dataset called brain tumor MRI in Kaggle challenge was collected from multiple reports [45] for brain tumor classification. The dataset contains 7022 images of human brain MRI with four categories: glioma, meningioma, pituitary, and no tumor. The training set and test set have been divided for fair comparison. Only normal images in the training set are used to train the proposed Hetero-AE, and the performance is evaluated using the full test set.

4) COVID-19 Dataset: The COVID-19 image dataset is an open-source dataset published in Kaggle challenge from the report [46]. It contains three categories, including Covid-19, pneumonia, and normal image. The training set and test set have 251 and 66 images, respectively. We train the proposed Hetero-AE using only normal samples and then evaluate its performance with the full test set.

B. Experimental Setup

1) Implementation Details: We implemented our method in Pytorch 1.11 framework and performed training and evaluation with an AMD Ryzen Threadripper 3960X 24-Core CPU and an NVIDIA GeForce RTX 3090 GPU. The model was trained with batch size 200 and Adam optimizer, where the initial learning rate was set to 0.0001. The image was resized to

TABLE I
PERFORMANCE (AUC(%), F1-SCORE (%), ACC (%), SEN (%), AND SPE (%)) YIELDED BY DIFFERENT ALGORITHMS ON RETINAL OCT, CHESTX-RAY, BRAIN MRI, AND COVID-19 DATASETS

Datasets	Metrics	AE [42]	VAE [43]	Ganomaly [21]	f-AnoGAN [3]	SALAD [9]	STFPM [38]	MKD [20]	RD4AD [39]	Ours
Retinal OCT	AUC	77.79	80.04	83.53	83.35	96.42	96.86	96.72	97.64	**98.94 ± 0.16**
	F1-score	85.79	85.55	88.65	84.73	93.42	95.80	94.60	96.40	**97.10 ± 0.28**
	ACC	78.28	77.63	81.60	77.50	90.64	93.70	91.60	94.60	**95.76 ± 0.42**
	SEN	94.38	95.32	95.86	89.89	95.69	95.60	**97.60**	96.40	97.46 ± 0.52
	SPE	41.70	37.45	38.80	49.36	79.15	88.00	73.60	89.20	**90.64 ± 1.45**
Chest X-ray	AUC	59.87	61.81	67.93	75.46	82.65	84.80	85.84	77.70	**90.67 ± 0.80**
	F1-score	77.20	77.37	80.52	81.00	82.14	82.33	84.74	79.81	**88.64 ± 0.64**
	ACC	63.40	66.04	71.31	74.00	75.92	78.20	79.01	71.47	**85.35 ± 0.86**
	SEN	**98.97**	98.21	94.87	88.97	88.46	81.28	93.33	90.25	91.43 ± 0.83
	SPE	3.86	6.87	32.05	36.48	54.94	73.07	55.12	40.17	**75.21 ± 2.00**
Brain MRI	AUC	78.40	91.76	83.29	92.22	94.51	92.71	93.30	87.03	**99.21 ± 0.56**
	F1-score	85.77	93.85	90.42	92.25	96.05	92.74	92.84	89.56	**98.37 ± 0.29**
	ACC	78.94	91.30	86.65	89.16	94.43	89.85	89.93	84.59	**97.73 ± 0.40**
	SEN	92.49	96.13	91.16	93.37	97.90	93.81	94.59	95.69	**99.06 ± 0.40**
	SPE	48.14	80.49	76.54	79.75	86.66	80.98	79.50	59.75	**94.71 ± 0.28**
COVID-19	AUC	74.30	90.65	92.61	92.40	94.88	92.30	92.90	73.59	**97.35 ± 0.82**
	F1-score	80.00	88.42	92.47	90.52	89.74	88.88	90.32	77.97	**95.59 ± 0.06**
	ACC	72.72	83.33	89.39	86.36	86.67	84.84	86.36	75.00	**93.94 ± 0.01**
	SEN	78.26	91.30	93.47	93.47	87.50	86.95	91.30	71.88	**94.20 ± 1.26**
	SPE	60.00	65.00	80.00	70.00	85.00	80.00	75.00	80.00	**93.33 ± 2.89**

256×256. The maximal epoch for training was 200, and α was set to 0.7 in the loss function. Using brain MRI dataset as an example, the training time of one epoch is 19.46 minutes. The number of multi-head attention mechanisms (N_i) in the three stages are 2, 4 and 8, respectively. The number of feature channels (C_i) in the three stages are 64, 128 and 256, respectively. The sizes of p_2 in the three stages in the decoder are 8, 4 and 2, respectively.

2) Evaluation Metrics: For performance evaluation, we calculate the area under curve (AUC), F1-score, average classification accuracy (ACC), sensitivity (SEN), and specificity (SPE) as the evaluation metrics. These metrics can be divided into two groups. The first group (SEN and SPE) focuses on abnormal and normal samples respectively, while the second group (ACC, AUC and F1-score) evaluates the overall performance. SEN, also known as recall, refers to the ability to identify diseased samples. SPE refers to the model's ability to identify normal samples. ACC is the ratio of correctly predicted samples to the total input data. It provides a general indication of the model's performance but may not reflect the performance on imbalanced datasets. The AUC and F1-score are more reliable than ACC when dealing with imbalanced datasets. Here, F1-score, ACC, SEN, and SPE are defined as $F1\text{-}score = \frac{2TP}{2TP+FN+FP}$, $ACC = \frac{TP+TN}{TP+TN+FP+FN}$, $SEN = \frac{TP}{TP+FN}$, and $SPE = \frac{TN}{TN+FP}$, where TP, TN, FN, and FP are the true positives, true negatives, false negatives, and false positives, respectively. The thresholds used to calculate the evaluation metrics are chosen based on the best F1-score.

C. Comparison With State-of-the-Art Methods

In our experiments, apart from the state-of-the-art reconstruction-based approaches, including AE [42], VAE [43], f-AnoGAN [3], GANomaly [21], and SALAD [9], we also compare our method with non-reconstruction-based approaches including Student-Teacher Feature Pyramid Matching (STFPM) [38], Multiresolution Knowledge Distillation (MKD) [20], and RD4AD [39].

The experimental results of our Hetero-AE compared with the above eight methods are presented in Table I. Convolutional AEs with adversarial learning (*i.e.*, Ganomaly, f-AnoGAN, and SALAD) outperform convolutional AEs methods (*i.e.*, AE and VAE), and the main reason is that they benefit from the constraints of the adversarial strategy. Further, the knowledge distillation-based approaches (*i.e.*, STFPM, MKD achieve higher performance than the above two groups of methods. The better performance of the distillation methods may benefit from their multi-stage feature comparison between teacher and student. The RD4AD method is an improved version based on the distillation method. It outperforms the above two distillation methods on the retinal OCT dataset, but it shows performance degradation in Chest X-ray, Brain MRI and COVID-19 datasets. The decline in performance is attributed to the reverse distillation strategy of RD4AD, which resulted in poor discriminability for medical images with high similarity in both the untrainable teacher and trainable student networks of RD4AD. From Table I, it can be seen that the proposed Hetero-AE outperforms the other SOTA methods in terms of AUC and F1-score. The sensitivity of MKD is slightly higher than that of our method. To verify the significance of this difference, we performed a t-test. Both methods were trained five times with different random seeds to ensure fairness. The p-value (0.1968) of the test result is greater than 0.05, which indicates that the MKD's superiority in SEN over our method is not significant. Despite our approach yielding a lower SEN score than MKD and AE, it is worth noting that a much higher SPE performance is achieved by our approach. Balanced SPE and SEN demonstrate our model holds greater practical value. When evaluating the model comprehensively, metrics such as AUC, F1-score, and ACC

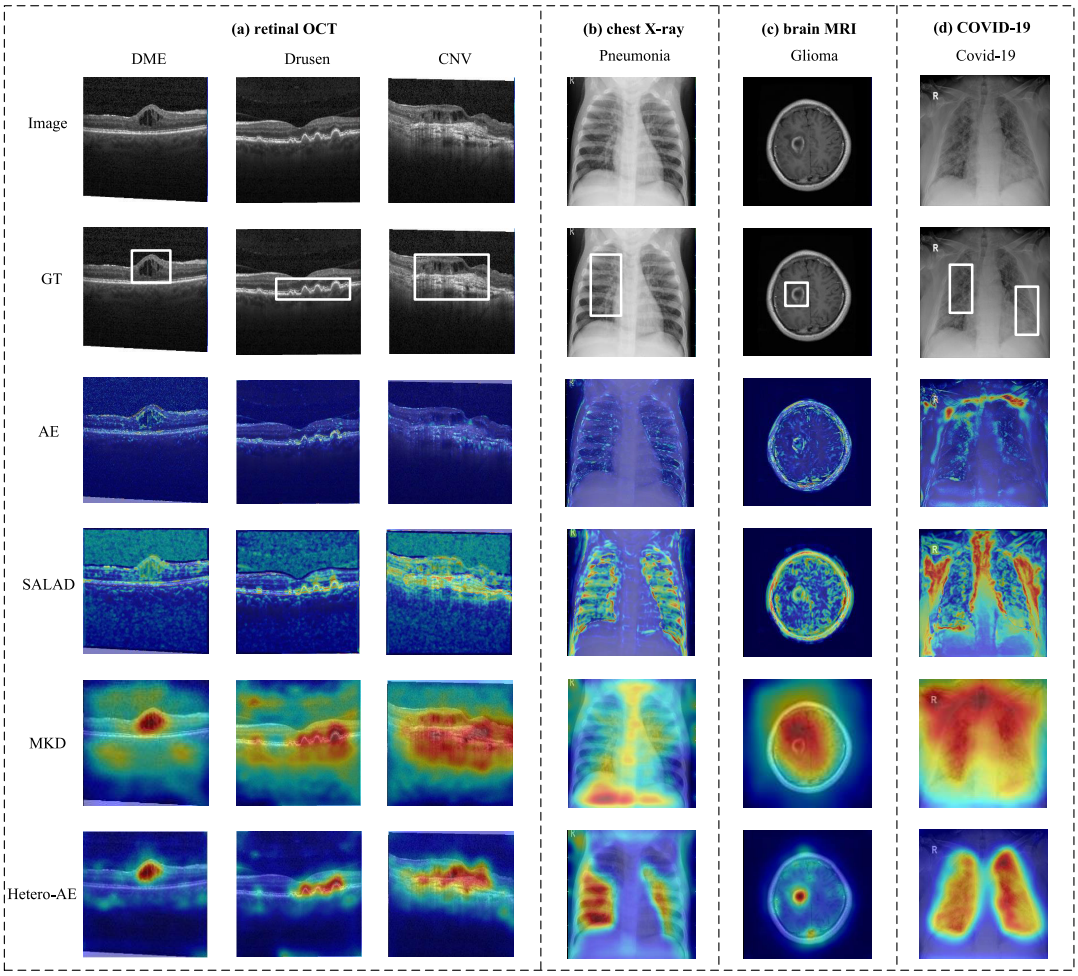

Fig. 4. The heatmaps locating the anomalous areas on retinal OCT, chest X-ray, brain MRI, and COVID-19. The red color in the heatmap represents areas that are more likely to be anomalies. The first row displays input images, the white boxes in the second row represent ground truth (GT) of anomalies, and the other rows show the heatmaps predicted by various methods, including AE, SALAD, MKD, and our Hetero-AE, respectively.

indicate that our method surpasses all comparison methods across four different datasets. The superior performance of Hetero-AE benefits from the heterogeneous structure and multi-stage feature comparison. The heterogeneous structure can enable the encoder and decoder to generate large feature differences on anomalous samples, which can improve the anomaly detection performance by avoiding identity mapping of AE. Meanwhile, multi-stage feature comparison can reduce the perturbation of pixel-wise reconstruction in complex textures.

It can be found that the performance of the listed models on chest X-ray consistently decreases by comparing the performance on the four datasets. The underlying reason for the performance degradation is that pneumonia lesions do not have sharp boundaries on chest X-ray dataset, which is difficult for unsupervised algorithms to detect anomalous samples.

It can be observed that the comparison methods demonstrate superior performance on retinal OCT compared with the other three datasets. This is because the Retinal OCT has the largest number of training samples, allowing the model to see abundant normal examples and learn the feature distribution of the normal OCT images. However, model performance depends not only on the dataset scale, but also on whether abnormalities in the dataset are salient. For example, the COVID-19 dataset has the fewest samples, but the differences between normal individuals and severe pneumonia patients are obvious. Therefore, comparison methods can achieve satisfactory anomaly detection on the COVID-19 dataset. In summary, both sample size and abnormality saliency jointly influence the model performance.

D. Visualization

1) Feature Visualization Using Heatmaps: We visualize the features of the proposed Hetero-AE on four public datasets using feature heatmaps as shown in Fig. 4. Different colors are used to indicate the feature distances between encoder and decoder. The red color indicates a large feature distance, which implies a more likely anomalous area. The blue color indicates that the feature distance is small, which implies it is more likely to be a normal region.

It can be observed from Fig. 4 those heatmaps generated by the AE method can hardly detect any anomalies. The underlying reason is that the AE method is influenced by identity mapping, which hinders its ability to effectively localize anomalous areas. Although the SALAD method (AE with

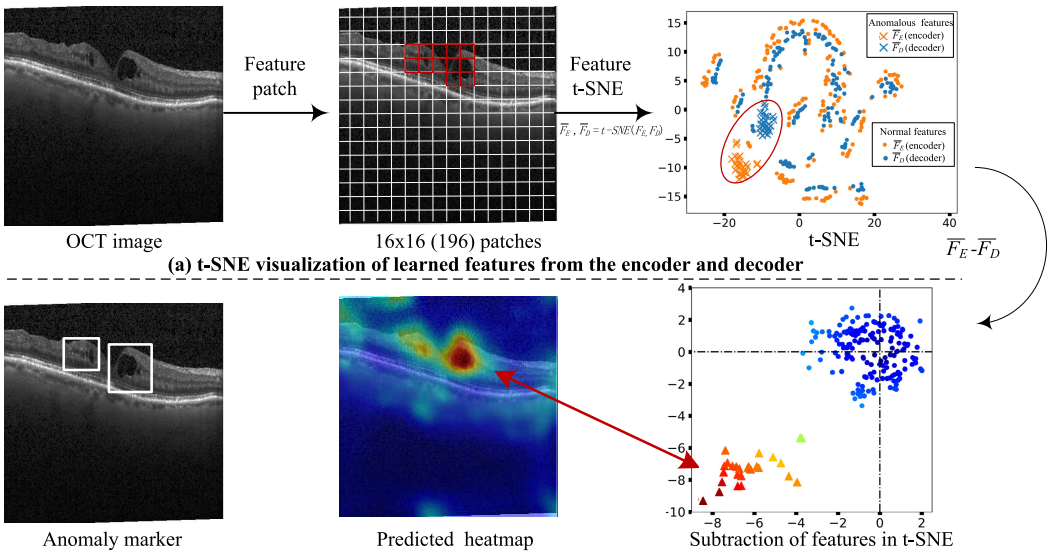

Fig. 5. Visualization of deep feature embeddings. (a) t-SNE visualization of learned features from the decoder and encoder. 196 pairs of features ($\overline{F}_E$ and $\overline{F}_D$) correspond to 196 image patches. (b) Relationship between predicted anomaly heatmaps and feature t-SNE. $\overline{F}_E - \overline{F}_D$ stands for feature subtraction between the encoder and the decoder after t-SNE dimensionality reduction.

adversarial learning) has achieved satisfactory classification results, its heatmaps suffer from noise caused by pixel-wise differences. The MKD method suppresses the noise caused by pixel-wise differences, but the accuracy of its heatmap in localizing the anomalous area still needs improvement. It can be observed that the proposed Hetero-AE can localize the anomalies more accurately. The superior results demonstrate that the proposed heterogeneous structure can effectively hinder identity mapping in AE methods. Furthermore, feature comparisons instead of pixel-wise comparisons can effectively suppress pixel-wise noise.

2) Feature Visualization Using t-SNE: To show the feature distribution of the encoder and decoder in the proposed method, t-Distributed Stochastic Neighbor Embedding (t-SNE) [47] is used to visualize the features. This section uses OCT images as an example. In Fig. 5(a), the OCT image is divided into 16 × 16 non-overlapping patches, and the encoder and decoder display patch features with t-SNE. Specifically, $\overline{F}_E$ and $\overline{F}_D$ represent the encoder features and decoder features after feature dimensionality reduction, respectively. As shown in Fig. 5(a), corresponding features from the encoder and decoder on normal regions are close in distance. This suggests that the multi-stage feature comparison loss can make the decoder features similar to the encoder features on normal samples. It is worth noting that the features (red circle) generated by the encoder are far away from those generated by the decoder on anomalous samples, which indicates the effectiveness of the proposed heterogeneous structure. To show the differences more clearly, we further let the encoder features subtract the decoder features and then show the distribution of features ($\overline{F}_E - \overline{F}_D$) in Fig. 5(b). It can be observed that the feature points of anomalous regions deviate from the coordinate origin which is the location of most normal patches. This observation supports the hypothesis that patches with large feature differences are anomalous regions.

E. Ablation Studies

Ablation studies are conducted to better understand the impact of each component in our framework. The contribution of the heterogeneous framework is discussed, followed by the analysis of multi-stage feature comparison. The ablation studies are carried on Chest X-ray and Brain MRI.

1) The Effectiveness of the Heterogeneous Framework: The purpose of the heterogeneous framework is to make the encoder and decoder have different inductive biases. Assuming that the encoder is only based on CNN module and the decoder is based on hybrid CNN-Transformer module, the AE framework can achieve better anomaly detection performance. To demonstrate the above statement, we conduct two groups of experiments to verify the effectiveness of the heterogeneous framework for anomaly detection. The first group is performed based on the homogeneous framework with the same structure. The popular ResNet and Efficientnet [48] are used as backbones for comparison, respectively. The second group is performed based on the heterogeneous framework, which refers to the different structures between encoder and decoder. In this group, ResNet (CNN) is selected as the encoder. We compare different structures for the decoder. For Transformer-based methods, we select Swin [24] and PVT [49]. For hybrid-structured methods, we employ CMT [50] and our proposed HCTN. The results of the two groups of experiments are presented in Table II. It can be observed that the performance of the heterogeneous framework is better than that of the homogeneous framework. To verify whether the improvement in model performance stems from the introduction of the Transformer. We evaluate the performance of the model with

TABLE II
PERFORMANCE COMPARISON BETWEEN HOMOGENEOUS AND HETEROGENEOUS FRAMEWORKSON CHEST X-RAY AND BRAIN MRI DATASETS

Type	Encoder	Decoder	Chest X-ray			Brain MRI			Param	FLOPs
			AUC	F1-score	ACC	AUC	F1-score	ACC	(10^6)	(10^9)
homogeneous framework	ResNet	ResNet(CNN)	85.72	84.77	80.76	95.50	94.02	91.53	**13**	39
	Efficientnet	Efficientnet(CNN)	84.09	82.59	78.04	94.50	93.33	90.61	20	**26**
	CNN+Trans	CNN+Trans	85.40	83.12	78.20	92.69	92.72	89.77	21	30
heterogeneous framework	ResNet	Swin (Transformer)	86.20	84.61	79.00	96.63	94.53	92.29	30	48
		PVT (Transformer)	85.94	84.27	78.04	96.38	94.39	92.06	23	45
		CMT(Hybrid)	86.29	85.35	81.09	97.53	97.61	96.64	24	39
		Our HCTN (Hybrid)	**90.13**	**88.22**	**84.93**	**99.20**	**98.68**	**98.16**	17	33

TABLE III
ABLATION STUDY ON HYBRID CNN-TRANSFORMER MODULE ON CHEST X-RAY AND BRAIN MRI DATASETS

Encoder	Decoder	Chest X-ray					Brain MRI					Param
		AUC	F1-score	ACC	SEN	SPE	AUC	F1-score	ACC	SEN	SPE	(10^6)
ResNet	CNN+CNN	85.54	82.51	78.20	82.30	71.30	93.92	92.99	90.38	92.38	85.92	17.90
	Trans+Trans	87.49	85.93	82.85	83.84	81.19	97.44	94.86	92.82	95.91	85.93	**16.74**
	CNN+Trans (HCTN)	**90.13**	**88.22**	**84.93**	**90.25**	**76.06**	**99.20**	**98.68**	**98.16**	**99.66**	**94.81**	17.32

TABLE IV
PERFORMANCE COMPARISON OF THE AUTO-ENCODER FRAMEWORK USING FEATURE COMPARISONON CHEST X-RAY AND BRAIN MRI DATASET

Methods	Feature comparison	Chest X-ray					Brain MRI				
		AUC	F1-score	ACC	SEN	SPE	AUC	F1-score	ACC	SEN	SPE
E-D (ResNet)	×	81.82	80.09	74.35	82.56	60.68	91.13	91.37	88.02	91.83	79.50
	✓	85.72	84.77	80.76	85.64	72.64	95.50	94.02	91.53	96.46	80.49
E-D (Transformer)	×	78.89	80.22	72.43	89.48	44.01	89.89	90.71	87.18	90.61	79.50
	✓	82.76	83.29	77.88	88.20	60.68	93.72	92.86	90.16	92.60	84.69
Our Hetero-AE	×	85.91	83.33	77.88	88.46	60.25	94.00	93.11	90.38	94.03	82.22
	✓	**90.13**	**88.22**	**84.93**	**90.25**	**76.06**	**99.20**	**98.68**	**98.16**	**99.66**	**94.81**

CNN+Trans architecture for both encoder and decoder. Compared with our final model (last row of Table II), the result does not exhibit superior performance. This indicates that the performance improvement is attributed to the heterogeneous framework rather than the introduced Transformer block. The heterogeneous structure makes anomalous samples have a larger reconstruction error, which is beneficial to distinguish them from normal samples. Furthermore, a comparison is made between two types of decoders in heterogeneous structures. As shown in Table II, it can be observed that the performance of the decoder using the hybrid network outperforms that of using the pure Transformer. This implies that the hybrid CNN-Transformer-based decoder can generate features that are closer to the encoder's features on normal samples while more different features on anomalous samples. In addition, it can be observed that the Transformer will increase computational cost of the model. The proposed hybrid network achieves a trade-off between detection performance and high computational cost.

To further validate the effectiveness of the cooperation of CNN and Transformer blocks, we conduct an ablation study on the hybrid CNN-Transformer module as shown in Table III. For all model structures, we use the same structure for the encoder while three different settings are used for the decoder. 1) CNN+CNN: only CNN blocks are used in decoder; 2) Trans+Trans: only Transformer blocks are used in decoder; 3) CNN+Trans: our final model using both CNN and Transformer (HCTN). As shown in Table III, the three different structures used for the decoder have similar parameter counts. The decoder implemented with Trans+Trans or CNN+Trans achieves better performance than the decoder using only CNN. This demonstrates that the performance improvement of our method stems from the structural differences between the encoder and decoder. Our proposed CNN+Transformer (HCTN) achieves superior performance than using Transformer alone. This demonstrates that when CNN is utilized in the encoder, employing a hybrid of CNN and Transformer in the decoder is more suitable for a heterogeneous framework than only using the Transformer.

To more intuitively observe the distribution of feature distances between the encoder and decoder, we use boxplots to compare the feature distances of different networks on normal and anomalous samples. Feature distances are compared when the encoder E only uses CNN, while the decoder D uses CNN, Transformer, and a hybrid of CNN+Transformer block respectively. We observe the impact of structural differences between E and D on feature distances in Fig. 6, where the three network structures are denoted as E(CNN)-D(CNN), E(CNN)-D(Transformer), and E(CNN)-D(CNN+Transformer). Specifically, the difference between these networks is that D(CNN) is obtained by replacing all Transformer modules of D(CNN+Transformer) with CNN,

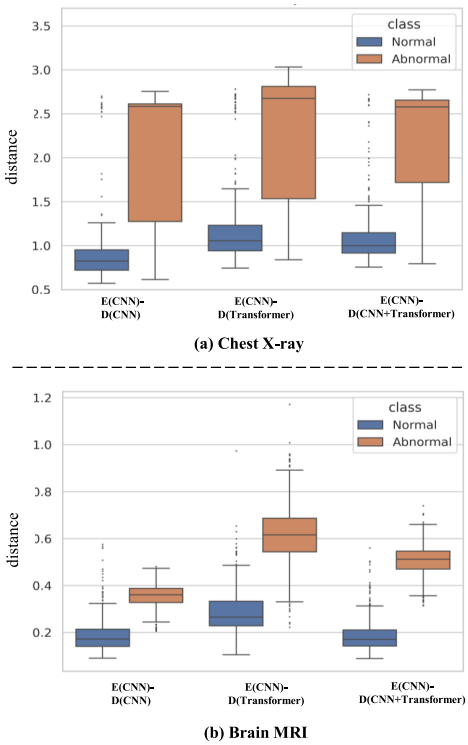

Fig. 6. Boxplots showing the distribution of feature distances. The x-axis represents different network structures. The y-axis represents the distance calculated from the multi-stage feature distance between the encoder and decoder. The blue boxplots and orange boxplots show the distribution of feature distances on normal samples and abnormal samples respectively. The lines in the box represent the 25th percentile, median, 75th percentile, and outliers are marked as dots. (a) Comparison of three networks on chest X-ray dataset. (b) Comparison of three networks on brain MRI dataset.

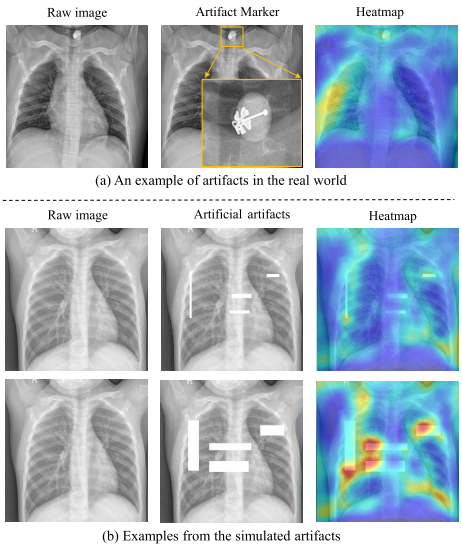

Fig. 7. The impact of artifacts on anomaly detection. (a) An example of artifacts in the real world. (b) Examples from the simulated artifacts.

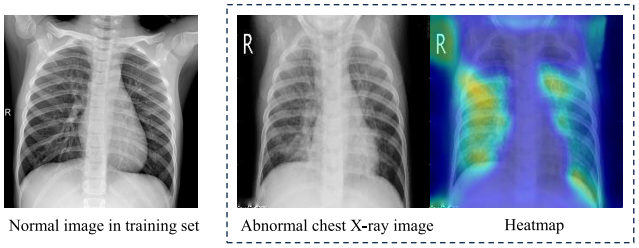

Fig. 8. Examples of chest X-ray images with complex textures. The left image shows a normal lung from the training set. The right image shows lungs infected with viral pneumonia and a heatmap predicted by our method.

and D(Transformer) is obtained by replacing all CNN modules of D(CNN+Transformer) with Transformer. It can be observed from Fig. 6 that our D(CNN+Transformer) results in a larger distribution difference in feature distances between normal and anomalous samples compared with D(CNN), which facilitates distinguishing between normal and abnormal.

2) The Effectiveness of Feature Comparison: We further compare the performance of auto-encoders with and without feature comparison in Table IV. Models without feature comparison means that pixel-wise comparisons are used as the optimization objective. E-D (ResNet/Transformer) represents that both the encoder and decoder are ResNet/Transformer networks. It can be observed from Table IV that the performance of homogeneous and heterogeneous frameworks using feature comparison can achieve better performance than pixel-wise difference comparison. This demonstrates that feature comparison can make the performance of auto-encoders no longer limited by small differences at pixel level. Feature-level comparisons are more robust than pixel-wise comparisons.

F. Discussion

Our method is trained solely on normal samples to detect anomalies. Images that are significantly different from normal samples are identified as anomalous. Examples of real-world artifacts and simulated artifacts are shown in Fig. 7. It can be observed that our method does not recognize small artifacts as anomalies, but for large artifacts, our method recognizes them as anomalies. Large artifacts similar to those in Fig. 7(b) are very unusual in actual clinical practice, and it is reasonable for our method to identify these areas as abnormalities.

A potential limitation is that the performance of the proposed method degrades for medical images with complex textures. The degradation in model performance on images with complex textures is a challenging problem in the field of medical anomaly detection [9]. Chest X-ray images show more complex tissue structures than retinal OCT and brain MRI images. A failure case is shown in Fig. 8. The abnormal chest X-ray image shows symptoms of coarsening lung textures due to viral infection in both lungs. The heatmap predicted by our algorithm indicates potential minor anomalies in the lung regions, but does not reach our predefined anomaly threshold. This is because the difference in texture between anomalous and normal images is too subtle, resulting in inaccurate predictions by our method. In future work, we will adapt the proposed method to medical images with complex textures.

The proposed method has practical significance for disease screening, where the population being screened consists mostly of healthy people and a small portion of diseased people, and the diseases are diverse. Screening a large number

of images is time-consuming, which can make clinicians overlook some abnormal images. The heatmaps of our method provide an intuitive way to highlight areas that may contain anomalies, allowing clinicians to prioritize these regions efficiently. The heatmaps serve as a reference to help reduce the risk of overlooking important details during image review. The proposed method can help shorten the time required for diagnosis, especially when dealing with large image sets.

V. Conclusion

A heterogeneous auto-encoder framework for unsupervised anomaly detection, namely Hetero-AE, is proposed in this paper, which reduces the feature difference on normal samples and increases that on anomalous samples. Specifically, a novel hybrid CNN-Transformer network is proposed to enable the decoder to learn long-range feature dependencies compared with the encoder. Further, a multi-scale sparse Transformer block is presented in the Transformer module, which improves the performance of the decoder and reduces the computational cost. In addition, the multi-stage feature comparison is introduced into the auto-encoder framework to mitigate pixel-wise noise. Experiments on four medical image datasets demonstrate the effectiveness of our approach.

References

[1] K. Zhou et al., "Proxy-bridged image reconstruction network for anomaly detection in medical images," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 582–594, Mar. 2022.

[2] 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 *Proc. Int. Conf. Inf. Process. Med. Imag.* Boone, NC, USA: Springer, Jun. 2017, pp. 146–157.

[3] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[7] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.

[8] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[9] H. Zhao et al., "Anomaly detection for medical images using self-supervised and translation-consistent features," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3641–3651, Dec. 2021.

[10] B. Zong et al., "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

[11] D. Zimmerer, S. A. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," 2018, *arXiv:1812.05941*.

[12] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 161–169.

[13] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders," 2018, *arXiv:1806.04972*.

[14] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2018, pp. 3–17.

[15] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[16] K. Zhou et al., "Memorizing structure-texture correspondence for image anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2335–2349, Jun. 2022.

[17] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[18] L. Ruff et al., "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4393–4402.

[19] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 375–390.

[20] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14902–14912.

[21] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semisupervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Springer, 2018, pp. 622–637.

[22] A. De Nardin, P. Mishra, G. L. Foresti, and C. Piciarelli, "Masked transformer for image anomaly localization," *Int. J. Neural Syst.*, vol. 32, no. 7, Jul. 2022, Art. no. 2250030.

[23] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[25] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Autoencoder-based network anomaly detection," in *Proc. Wireless Telecommun. Symp. (WTS)*, Apr. 2018, pp. 1–5.

[26] Y. Lu and P. Xu, "Anomaly detection for skin disease images using variational autoencoder," 2018, *arXiv:1807.01349*.

[27] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip connected and adversarially trained encoder–decoder anomaly detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[28] J. Wolleb, F. Bieder, R. Sandkuhler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2022, pp. 35–45.

[29] T. Xiang et al., "SQUID: Deep feature in-painting for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23890–23901.

[30] Y. Cai, H. Chen, X. Yang, Y. Zhou, and K.-T. Cheng, "Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102794.

[31] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "Encoder–decoder contrast for unsupervised anomaly detection in medical images," *IEEE Trans. Med. Imag.*, vol. 43, no. 3, pp. 1102–1112, Mar. 2024.

[32] Y. Li et al., "Self-supervised anomaly detection, staging and segmentation for retinal images," *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102805.

[33] Z. You et al., "A unified model for multi-class anomaly detection," in *Proc. NeurIPS*, vol. 35, 2022, pp. 4571–4584.

[34] Z. You, K. Yang, W. Luo, L. Cui, Y. Zheng, and X. Le, "ADTR: Anomaly detection transformer with feature reconstruction," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2022, pp. 298–310.

[35] X. Cai, R. Xiao, Z. Zeng, P. Gong, and Y. Ni, "ITran: A novel transformer-based approach for industrial anomaly detection and localization," *Eng. Appl. Artif. Intell.*, vol. 125, Oct. 2023, Art. no. 106677.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[37] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student–teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4183–4192.

[38] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for unsupervised anomaly detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2021, pp. 1–14.

[39] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9737–9746.

[40] Y. Tian et al., "Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images," *Med. Image Anal.*, vol. 90, Dec. 2023, Art. no. 102930.
[41] M. Liu, Y. Jiao, and H. Chen, "Skip-ST: Anomaly detection for medical images using student–teacher network with skip connections," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2023, pp. 1–5.
[42] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 372–380.
[43] T. Matsubara, K. Sato, K. Hama, R. Tachibana, and K. Uehara, "Deep generative model using unregularized score for anomaly detection with heterogeneous complexity," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5161–5173, Jun. 2022.
[44] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.
[45] J. Cheng, "Brain tumor dataset," Apr. 2017, doi: 10.6084/m9.figshare.1512427.v5.
[46] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "COVID-19 image data collection: Prospective predictions are the future," 2020, *arXiv:2006.11988*.
[47] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[48] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[49] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.
[50] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12175–12185.

Shuai Lu received the B.Sc. and M.Sc. degrees from Beijing University of Chemical Technology, Beijing, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with Beijing Institute of Technology. His current research interests include image processing, pattern recognition, and intelligent computing.

Weihang Zhang received the B.S. degree from the School of Instrumentation Science and Opto-Electronics Engineering, Beihang University, Beijing, China, in 2015, and the Ph.D. degree from the Department of Precision Instrument, Tsinghua University, in 2020. He was a Postdoctoral Researcher with the Department of Precision Instrument, Tsinghua University, from 2020 to 2022. He is currently an Assistant Professor with the School of Medical Technology, Beijing Institute of Technology. His research interests include image processing, pattern recognition, and intelligent computing.

He Zhao received the B.E. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 2014 and 2020, respectively. He was a Postdoctoral Researcher with the University of Oxford, working with multi-modality medical imaging analysis. He is currently a Lecturer with the University of Liverpool. His research interests include medical image analysis, computer vision, and deep learning.

Hanruo Liu received the Ph.D. degree from the University of East Anglia, U.K., in 2013. She is currently an Associate Professor of ophthalmology with Beijing Tongren Eye Center, Beijing Tongren Hospital. Her research interests include intelligent ophthalmology big data research.

Ningli Wang is currently the Director of Beijing Tongren Eye Center, Beijing Tongren Hospital; the Dean of the School of Ophthalmology, Capital Medical University; the Head of the National Committee for the Prevention of Blindness; an Advisory Board Member of the Chinese Academy of Medical Sciences; and the President of Asia–Pacific Academy of Ophthalmology. His research interests include pathogenesis, and diagnosis and treatment of glaucoma.

Huiqi Li (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2003. She is currently a Professor with Beijing Institute of Technology. Her research interests include medical image processing and computer-aided diagnosis.