Anomaly Detection for Medical Images Using Teacher-Student Model with Skip Connections and Multi-scale Anomaly Consistency

Mingxuan Liu, *Student Member, IEEE*, Yunrui Jiao, Jingqiao Lu, and Hong Chen, *Senior Member, IEEE*

***Abstract*—Anomaly detection (AD) in medical images aims to recognize test-time abnormal inputs according to normal samples in the training set. Knowledge distillation based on the teacher-student (T-S) model is a simple and effective method to identify anomalies, yet its efficacy is constrained by the similarity between teacher and student network architectures. To address this problem, in this paper, we propose a T-S model with skip connections (Skip-TS) which is trained by direct reverse knowledge distillation (DRKD) for AD in medical images. First, to overcome the low sensitivity to anomalies caused by structural similarity, we design an encoder-decoder architecture where the teacher network (T-Net) is a pre-trained encoder and the student network (S-Net) is a randomly initialized decoder. During training, the S-Net learns to reconstruct the shallow representations of images from the output of the T-Net, which is called DRKD. Secondly, we introduce skip connections to the T-S model to prevent the S-Net from missing normal information of images at multi-scale. In addition, we design a multi-scale anomaly consistency (MAC) loss to improve the anomaly detection and localization performance. Thorough experiments conducted on twelve public medical datasets and two private medical datasets demonstrate that our approach surpasses the current state-of-the-art by 6.4% and 8.2% in terms of AUROC on public and private datasets, respectively. Code and organized benchmark datasets will be available at https://github.com/Arktis2022/Skip-TS.**

***Index Terms*—Anomaly detection, medical image analysis, deep learning, teacher-student model, knowledge distillation**

I. Introduction

In medical image recognition, collecting and labeling abnormal data is time-consuming and expensive, especially when a disease is very rare [1]. Even with labelled data, supervised learning still faces the challenge of data imbalance, as abnormal data is typically in lower supply compared to normal data [2]. Anomaly detection (AD) aims to recognize test-time abnormal inputs based on normal samples observed during training [3]. Because training AD models relies solely on normal data to characterize the normal distribution, and normal medical images without lesion annotation are much easier to collect [4], researchers can conveniently establish AD models. As a result, the ability of AD to accurately detect anomalies in medical images using only normal data has made it a crucial technique in artificial intelligence-assisted diagnosis.

Most state-of-the-art works realize AD by building self-supervised tasks on the training dataset, which mainly include sample reconstruction [4]–[22], pseudo-outlier augmentation [23]–[26], and knowledge distillation (KD) [27]–[33]. Previous KD-based frameworks usually use a sufficiently pre-trained teacher network (T-Net) and a student network (S-Net) with a similar or identical structure to form a teacher-student (T-S) model. During training, knowledge is transferred from the T-Net to the S-Net, which allows the S-Net to learn the normal data manifold. Consequently, when fed with abnormal data, the features extracted by the S-Net and T-Net are anticipated to be inconsistent. However, the structural consistency of the T-Net and S-Net in the T-S model makes it a non-distinguishing filter [29]. Specifically, both T-Net and S-Net are encoders with the same network structure. S-Net may hence tend to overfit the output of T-Net, resulting in similar internal parameters. Therefore, traditional KD-based methods cannot guarantee the large gap between the features extracted by T-Net and S-Net for abnormal images, such that it is difficult to detect subtle anomalies in medical images.

Some studies address above problems by designing a smaller S-Net [28], [29], but their anomaly localization is limited by the weak feature extraction capability of shallow networks. Hanqiu *et al*. [31] proposed RDAD, a reverse distillation paradigm for the industrial AD, where a trainable one-class bottleneck embedding is utilized to connect the T-Net and S-Net, thus rendering the T-S model a framework with heterogeneity. However, this indirect connection necessitates the S-Net to recover shallow features of images from a low-dimensional representation that is dense with information. For medical tasks, normal images usually exhibit variable patterns, which will pose a challenge for the S-Net to accurately recover anomaly-free image representations and lead to AD failure consequently.

To mitigate the network structure consistency problem, in this paper, we propose a novel DRKD paradigm to train the Skip-TS model. Figure. 1 illustrates the T-S model and data flow of the proposed DRKD and the traditional KD paradigms [27]–[30]. First, to address the issue of non-distinguishing fil-

This work is supported by the National Science and Technology Major Project from Minister of Science and Technology, China (Grant No. 2018AAA0103100), Guangzhou Foshan Science and Technology Innovation Project (No. 2020001005585), and National Natural Science Foundation of China (No. 92164110 and U19B2041), partly supported by Beijing Engineering Research Center (No. BG0149). (Corresponding author: Hong Chen.)

Mingxuan Liu is with the Department of Biomedical Engineering, Tsinghua University, Beijing 100086, China (e-mail: liumx19@mails.tsinghua.edu.cn).

Yunrui Jiao is with the School of Integrated Circuits, Tsinghua University, Beijing 100086, China (e-mail: jyr19@mails.tsinghua.edu.cn).

Jingqiao Lu is with the Health Testing Technique and Equipment Research Center, Huangpu Joint Innovation Institute of Chinese Medicine, Guangzhou 510300, China (e-mail: lujingqiao@jiicm.org.cn)

Hong Chen is with the School of Integrated Circuits, Tsinghua University, Beijing 100086, China (e-mail: hongchen@tsinghua.edu.cn).

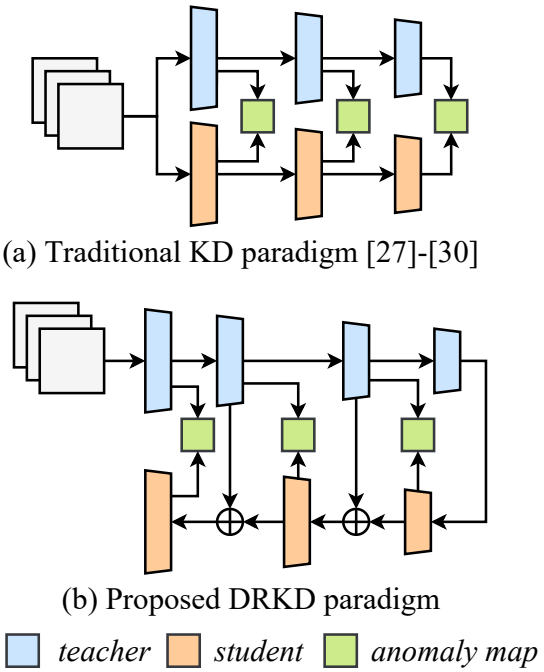


Fig. 1. T-S models and data flow in (a) Traditional KD paradigm [27]–[30], (b) proposed DRKD paradigm. Traditional KD uses teacher and student networks with similar architectures, both of which receive images as input. The proposed DRKD designs the student as a decoder, and directly receives multi-scale knowledge through skip connections.

ters encountered in traditional KD-based methods, we propose an encoder-decoder architecture for the T-S model, where deep features extracted by the T-Net are directly passed to the S-Net without intermediaries. During training, the S-Net undergoes a process known as DRKD, whereby it learns to reconstruct the multi-scale representation of the input image from its high-level semantic feature.

Secondly, to aid the S-Net in reconstructing an anomaly-free representation of the image, we propose a T-S model with skip connections between the T-Net and S-Net, called Skip-TS, with which the S-Net is able to receive multi-scale features of the image and the reconstruction difficulty is lessened. When normal medical images are used as input, the S-Net is able to produce features that are consistent with the output of T-Net at all scale to avoid false positives. Finally, the KD-based method needs to measure the difference between each layer's output to obtain the location information of anomalies. This anomaly localization method is based on the hypothesis that feature maps at each scale can reflect anomalous features. However, this hypothesis is not always correct in practice because sometimes the feature map at a certain scale exhibits high anomaly scores in normal regions [34]. We enforce the spatial consistency of the anomaly maps across layers by adding a multi-scale anomaly consistency (MAC) loss, which improves the model's ability of anomaly detection and localization.

Besides, previous studies only verified the performance of models on a small number of datasets, which is due to the lack of standardized benchmarks for medical AD. For the first time, we collect and organize fourteen medical datasets, twelve of them are existing public datasets [35]–[45], and two of them are private datasets collected in our work [46]. We verify the effectiveness of our method by comparing it with current state-of-the-art methods on these datasets. In addition, we qualitatively investigate the anomaly localization performance of the proposed method.

A preliminary version of this work was previously accepted [46]. Specifically, in the conference version of our paper:

1) We propose a novel distillation paradigm (DRKD). With this approach, the S-Net is designed as a decoder, which is directly fed with the high-level semantic features of images from the T-Net during the KD process. This effectively addresses the issue of non-distinguishing filters encountered in traditional KD-based methods.

2) To prevent false positives for normal medical images, a Skip-TS model is proposed to help the S-Net recover the multi-scale representation of normal images by introducing skip connections between the T-Net and S-Net.

3) We conduct extensive experiments on five medical datasets, and the results show that our proposed AD method achieves the best performance.

In this article, we extend our earlier work [46]:

1) We introduce a new method to accumulation the multiple anomaly maps generated by Skip-TS to realize anomaly localization.

2) A multi-scale anomaly consistency (MAC) loss is put forward to ensure the spatial consistency of the multi-scale anomaly maps obtained by Skip-TS, which improves the anomaly detection and localization performance.

3) We for the first time organize fourteen medical datasets (including the five datasets used in [46]) into benchmarks for medical AD. Extensive experiments on the benchmarks show that the proposed model Skip-TS has the best performance.

4) We analyse the interpretability of our model by conducting t-SNE analysis on the intermediate feature maps outputted by the T-Net and S-Net of Skip-TS, to assess which image features or regions are the most important for detecting anomalies.

The rest of our paper is organized as follows. In Section II, we introduce the related work on anomaly detection. In Section III, we describe the proposed KD-based method in details. In Section IV, extensive experiments on fourteen medical datasets are conducted to verify our method. Section V concludes our work.

II. Related Work

Anomaly detection, also known as out-of-distribution detection, can be solved by constructing the manifold of normal data. In addition, some studies are also devoted to pixel-level anomaly detection, *i.e.*, locating anomalies in images. This section will review previous efforts on these tasks.

A. Reconstruction-based Methods

Recently, reconstruction-based methods have dominated medical image anomaly detection. These methods learn a mapping function to reconstruct normal samples by using generative models, such as Autoencoder (AE) [47] and Generative Adversarial Nets (GAN) [48].

In particular, Akçay *et al.* [5] propose a Ganomaly model to jointly learn the generation of high-dimensional image space and the inference of latent space. Authors introduce skip connections in the model to thoroughly capture the multi-scale distribution of the normal data in image space [6]. AnoGAN [7] and f-AnoGAN [8] are also GAN-based anomaly detection models, the latter trains an encoder to replace the time-consuming iterative process in the testing phase.

However, due to the lack of focus on latent feature space, GAN-based methods fail to capture all the important features of normal data, leading to incorrect detections or misclassifications. In order to overcome the shortcoming, both image and feature spaces are considered through structure similarity loss and center constraint in SALAD [4], and experiments on optical coherence tomography (OCT) and chest X-ray datasets show its effectiveness. Many approaches also use AE and its variants. Salehi *et al*. [11] find that introducing self-supervised tasks such as puzzle-solving into AE-based methods can prevent overfitting and facilitate learning beyond pixel-level features.

Furthermore, in order to handle complex medical images, Shvetsova *et al*. [12] propose deep perceptual autoencoders that reconstructs high-resolution images through a progressive training process. Without constraints in the training process, traditional AE exhibit an overly generalized behavior and can sometimes recover anomalies. MAMA Net [9] utilizes a hash addressing memory module to solve this problem. Following the study of [9], Zhou *et al*. [10] propose a Proxy-bridged Image Reconstruction Network that bridges the input image and the reconstructed image with an intermediate proxy.

B. Pseudo-outlier Augmentation-base Methods

Pseudo-outlier augmentation converts AD to a supervised learning task by adding pseudo anomalies in normal images. CutPaste [23] generates pseudo outliers by cutting image patches and pasting them at random locations. To get synthetic anomaly images and reference masks for normal data, AnoSeg [24] uses hard augmentation to change the normal sample distribution. However, these methods are inefficient in medical image anomaly detection without medical expertise. AnatPaste [25] leverages a lung segmentation pretext task to generate anomalies in chest radiographs. But this technique is applied only in the anomaly detection of lung images.

C. Knowledge Distillation-based Methods

Some studies have found that networks pre-trained on ImageNet [49] or other large datasets can extract abnormal features, so it is possible to distinguish normal and abnormal images by directly modeling the deep features of images [50]–[54]. Since such methods need to save a lot of features of all training samples, which results in high computational overhead. CFA [55] reduces the complexity through a scalable memory bank at the expense of the model's capacity of capturing normal data distributions, leading to inadequate AD performance on intricate datasets.

In recent years, the KD-based method has emerged as a promising solution. By designing a S-Net to transfer the feature extraction capability of the pre-trained model, and the exclusive use of normal samples for training, the T-Net and S-Net in the T-S model are capable of producing distinct features for abnormal images. In order to obtain multi-scale abnormal features, US [18] uses an ensemble learning method, while STPM [28] and MKD [29] choose to calculate multi-scale feature differences. Kohei *et al*. [56] further improve the model in [28] using consistency between layer groups. In RDAD [31], a reverse distillation paradigm is realized by a trainable one-class bottleneck embedding module to introduce heterogeneity in the T-S model. However, above KD-based methods are limited in medical images because of the non-distinguishing filter problem [27]–[29] or the difficult reconstruction task faced by the S-Net [31]. Despite attempts made by Rui Xu *et al*. [32] to train an autoencoder as a teacher network and use the KD-based method for anomaly detection in CT images, there remains a need for a generalized T-S model for medical anomaly detection.

D. Summary

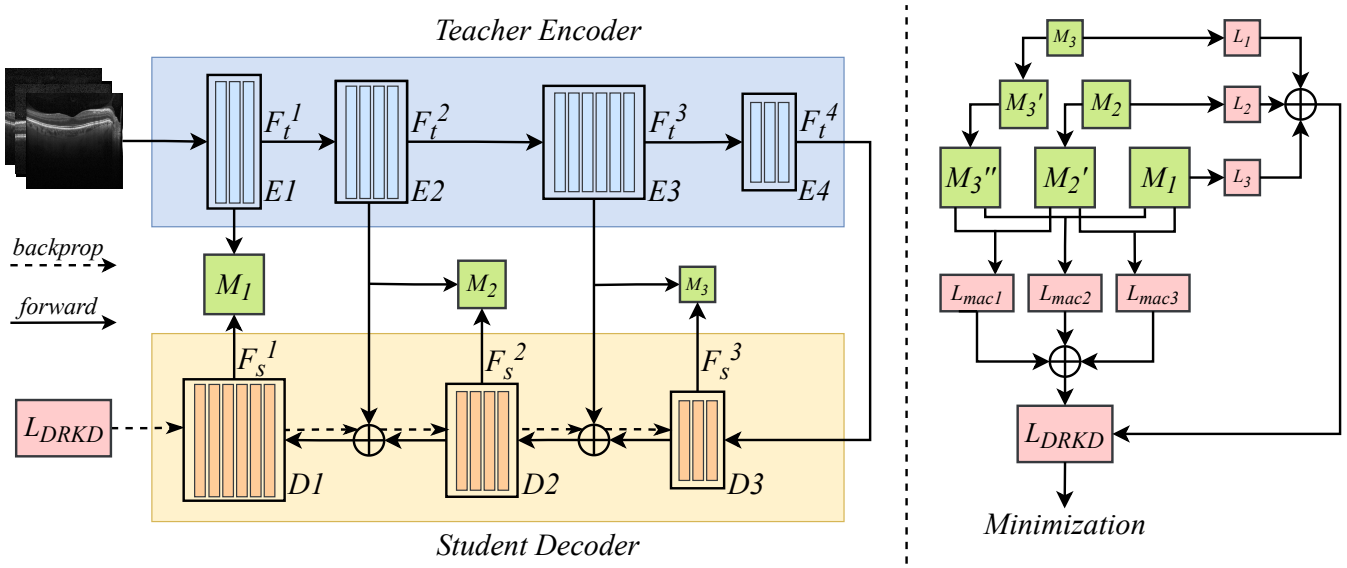
In Summary, most AD models for medical images use reconstruction-based schemes. However, these methods have the following problems: 1) the training process is complex and the performance is too sensitive to hyperparameters, which makes it hard to find the optimal settings for different datasets and tasks; 2) the generative model can reconstruct both normal and abnormal regions due to its high generalization capacity, and it is difficult for the model to reconstruct the high-frequency boundaries of normal images, resulting in false positives and negatives; 3) anomalies are found only by computing the pixel-level difference between the input and reconstructed images, resulting in more noise when anomaly localization; 4) the encoder-decoder architectures are trained from scratch, so that the powerful feature extraction capabilities of pre-trained models are not exploited, which results in less robust models and lower performance. Although previous studies have investigated the use of pseudo-outlier augmentation or KD-based methods for specific medical datasets [25], [32], these approaches suffer from limited generalization and only perform well on a few specific datasets. As such, this study aims to propose a simple, powerful, and general KD-based method that can be applied to various AD tasks in medical images.

III. METHOD

A. Pipeline

First of all, we define the anomaly detection task. Let $X_{train} = \{x_1, \cdots, x_n\}$ be a training dataset containing only normal images and $Y_{test} = \{y_1, \cdots, y_m\}$ be a testing dataset consisting of normal and abnormal data. Our goal is to train a model using X_{train} that can recognize and localize abnormal images in Y_{test}.

Figure 2 illustrates the overview of the proposed Skip-TS model and MAC loss for anomaly detection in medical images. As shown in Fig. 2, we train a T-S model called Skip-TS through DRKD (discussed in III.B). The Skip-TS consists of a pre-trained T-Net and a randomly initialized



(a) Data flow in teacher-student model with skip connections (b) Loss function

Fig. 2. Overview of the Skip-TS model and MAC loss for anomaly detection in medical images. (a) Skip-TS consists of a pre-trained T-Net and a randomly initialized S-Net. The T-Net can be divided into four sub-encoders *E1, E2, E3,* and *E4* and the S-Net can be divided into three sub-decoders *D1, D2*, and *D3*. During training, the S-Net needs to recover the multi-scale representation by minimizing the loss function L_{DRKD}. (b) The loss function L_{DRKD} is composed of two parts, one (*L1, L2, L3*) is directly obtained from 2D anomaly maps *M1, M2,* and *M3*, and the other ($L_{mac1}, L_{mac2}, L_{mac3}$) comes from the difference between *M1, M2,* and *M3*.

S-Net. First, we adopt a sufficiently pre-trained network to extract discriminative features of images [50]. Let F_t be the image feature extracted by T-Net, which contains normal and abnormal information. During training the S-Net reconstructs shallow features of each image from the input. Since only normal data is included in X_{train}, the S-Net learns the patterns of normal images from F_t. Let F_s be the image feature recovered by S-Net, in the testing process, because the S-Net only learned the manifold of normal data, F_s contains only normal feature and ignores anomalies, which brings the difference between F_t and F_s for abnormal data. That is, when all the input is normal image, F_t and F_s should be highly similar.In addition, the skip connections are helpful to maintain the consistent representation of normal features.

For each y_i in Y_{test}, we can obtain the anomaly score by measuring the similarity between F_s and F_t. The lower the similarity, the higher the anomaly score. By setting a threshold with the method proposed in [57], Y_{test} can be divided into two categories: the normal subset and the abnormal subset. As a result, anomaly detection is realized. Finally, by accumulating the anomaly maps M obtained from each F_s and F_t, the location information of the anomaly can be obtained.

It should be noted that the proposed DRKD paradigm also uses an encoder-decoder architecture. But unlike the generative model, we freeze the parameters of the encoder during training. Furthermore, we pay more attention to the gap of intermediate features extracted by the T-Net and S-Net rather than the pixel difference between the input image and the reconstructed image.

B. Direct Reverse Knowledge Distillation

As outlined in Section I and Section II.C, the traditional KD paradigm lacks sensitivity towards abnormal data and the RDAD model [31] has demonstrated a tendency to misclassify normal images as anomalies. These problems result in a low Area Under the Receiver Operating Characteristic (AUROC) [58] for KD-based anomaly detection methods when applied to medical images. In order to address the problems, we put forward a DRKD paradigm shown in Fig. 2, from which we can see that distillation is performed on Skip-TS with an encoder-decoder architecture, and features extracted by the T-Net are directly transferred to the S-Net without any intermediary.

With DRKD, T-Net is used to extract comprehensive features from input images including normal and abnormal information. For instance, when utilizing the pre-trained WideResNet50 [59] on ImageNet [49] and assign *E1, E2, E3,* and *E4* to its four layers. In this configuration, *E4* generates the input for S-Net, which represents the low-resolution feature set of the images. The task of other three sub-encoders *E1, E2,* and *E3* is to provide the multi-scale knowledge as a reference for the S-Net. During training, the parameters of T-Net are frozen to prevent the model from converging to a trivial solution [30].

The S-Net is structured symmetrically to that of T-Net to learn the intermediate representation of normal images from T-Net. This symmetry ensures that the output F_s of each layer in S-Net maintains consistent dimensions with its corresponding layer output F_t in T-Net. As a result, this design effectively prevents S-Net from processing anomaly data directly during the testing phase. In particular, when using WideResNet50 as T-Net, down-sampling is accomplished through convolutional layers with a kernel size of 3 and a stride of 2 [59]. Ac-

cordingly, the S-Net employs deconvolution layers [60] with a kernel size of 3 and a stride of 2 for up-sampling.

For the input $x_i \in R^{w \times h \times c}$ where h is the height, w is the width and c is the number of color channels, the k^{th} sub-encoder of the T-Net outputs $F_t^k(x_i) \in R^{w_k \times h_k \times d_k}$, where w_k, h_k and d_k represent the width, height and channel number of the feature map respectively. It is assumed that the number of sub-encoders is a, and $(a-1)$ for sub-decoders, the $(a-1)^{th}$ sub-decoder takes $F_t^a(x_i)$ as input and outputs $F_s^{a-1}(x_i) \in R^{w_{a-1} \times h_{a-1} \times d_{a-1}}$. For j^{th} $(j \in (0, a-1))$ sub-decoder, the input becomes $\left(F_t^{j+1}(x_i) + F_s^{j+1}(x_i)\right)$ because of the existence of skip connections (discussed in section III.C), and its output is $F_s^j(x_i) \in R^{w_j \times h_j \times d_j}$. It should be noted that when k is equal to j, that is, the dimension of $F_t^k(x_i)$ is equal to that of $F_s^j(x_i)$. We adopt the cosine similarity between F_t and F_s to compute the anomaly map. For F_t^l and F_s^l $(l \in (0, a))$, we can obtain a 2D anomaly map $M_l(x_i) \in R^{w_l \times h_l}$ by calculating the cosine similarity along the channel axis:

$$M_l(w,h) = 1 - \frac{\left(F_t^l(w,h)\right)^T \cdot F_s^l(w,h)}{||F_t^l(w,h)|| \cdot ||F_s^l(w,h)||} \tag{1}$$

1) Similarity Loss: As shown in Fig. 2(b), the output features of the T-Net and S-Net are required to be consistent when inputting a normal image to the model, that is, each pixel of the anomaly map should be close to zero. Therefore, we can constrain the output of S-Net by setting a similarity loss. The loss function L_l of each layer can be obtained by accumulating the loss map:

$$L_l = \frac{1}{w_l h_l} \sum_{w=1}^{w_l} \sum_{h=1}^{h_l} M_l(w,h) \tag{2}$$

then, the loss functions of each layer are added together to obtain the similarity loss L_{sim}:

$$L_{sim} = \sum_{l=1}^{a-1} L_l \tag{3}$$

2) Multi-scale Anomaly Consistency Loss: When KD-based methods are used for anomaly localization, it is commonly assumed that anomalous characteristics can be reflected across various scales by anomaly maps. However, this assumption may fail in certain scenarios. Inconsistencies in anomaly locations depicted by the anomaly map on a particular scale can undermine the final detection and localization of anomalies. In order to ensure that each scale's anomaly map accurately identifies anomalies, We use the MAC loss function to constrain the anomaly maps produced by the layers of the T-S model. First, we upsample each anomaly map to the same scale as the largest anomaly map via bilinear interpolation and get $M_l^{bi}(x_i) \in R^{w_{a-1} \times h_{a-1}}$. After that, we calculate the difference between all maps to obtain the MAC loss L_{mac}:

$$L_{mac} = \frac{1}{w_{a-1} h_{a-1}} \times \sum_{w=1}^{w_{a-1}} \sum_{h=1}^{h_{a-1}} \sum_{i,j \in (0,a)} \left(M_i^{bi}(w,h) - M_j^{bi}(w,h)\right)^2 \tag{4}$$

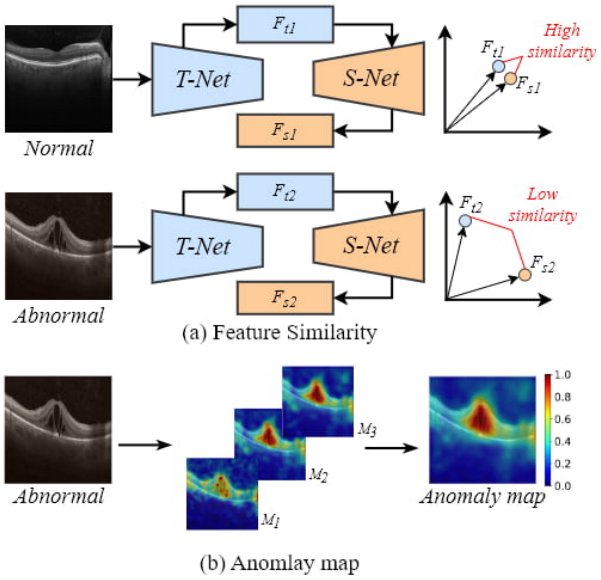


Fig. 3. Flowchart of anomaly detection and localization. (a) During testing, the T-Net outputs the real feature F_t and the S-Net outputs the anomaly-free one F_s. The anomaly score can be obtained by calculating the difference between F_t and F_s. (b) Anomaly localization can be achieved by merging anomaly maps on multi-scale. Red areas indicate higher anomaly scores.

Finally, with the loss function defined above, the overall objective of optimizing the student decoder can be expressed as:

$$L_{DRKD} = \alpha \cdot L_{sim} + \beta \cdot L_{mac} \tag{5}$$

where α, β are the loss weights. The α is set to 1.0 and β is set to 0.05.

C. Skip Connections

In order to help the S-Net reconstruct anomaly-free representations of images, we introduce skip connections to skip-TS inspired by image reconstruction [61]. For j^{th} $(j \in (0, a-1))$ sub-decoder, we let the $(F_t^{j+1}(x_i) + F_s^{j+1}(x_i))$ as its input. For abnormal images, as the S-Net is trained with normal data, $F_s^{j+1}(x_i)$ only contains normal patterns of images, and the sub-decoder learns how to decode the $F_s^{j+1}(x_i)$ into shallow features. However, because the pre-trained network T-Net can fully extract image features, $F_t^{j+1}(x_i)$ can be divided into $F_{t1}^{j+1}(x_i)$ and $F_{t2}^{j+1}(x_i)$:

$$F_t^{j+1}(x_i) = f(F_{t1}^{j+1}(x_i), F_{t2}^{j+1}(x_i)) \tag{6}$$

where $F_{t1}^{j+1}(x_i)$ and $F_{t2}^{j+1}(x_i)$ are normal and abnormal features respectively. The sub-decoder successfully decodes $F_{t1}^{j+1}(x_i)$ based on the knowledge gained during training but cannot decode $F_{t2}^{j+1}(x_i)$, which leads to a gap between $F_t^j(x_i)$ and $F_s^j(x_i)$. On the other hand, for normal image in testing, the information in $F_t^{j+1}(x_i)$ helps the sub-decoder reconstruct the shallow representation.

In summary, we argue that skip connections ensure the consistent features when using normal images as well as detects anomalies.

TABLE I
KEY STATISTICS FOR IMAGE DATASETS

Public Data	ID	Imaging Modalities	Associated Diseases	Training Data Normal	Testing Data Normal & Abnormal
HeadCT [35]	1	CT	Brain hemorrhage	60	40 & 100
BrainMRI [36]	2	MRI	Brain tumor	58	40 & 154
LungCT [37]	3	CT	Covid19	432	145 & 811
BUSI [38]	4	Ultrasound	Breast cancer	99	34 & 647
BreastMRI [39]	5	MRI	Breast cancer	525	175 & 700
Retinal OCT [40]	6	OCT	Diabetic macular edema	542	50 & 735
CP-CHILD-A [41]	7	Colonoscopy	Colonic polyps	800	800 & 800
CP-CHILD-B [41]	8	Colonoscopy	Colonic polyps	800	300 & 300
NAFLD [42]	9	Whole slide imaging	Non-alcoholic fatty liver	525	1645 & 2150
IQ-OTH/NCCD [43]	10	CT	Lung cancer	277	139 & 681
Covid-Xray [44]	11	Xray	Covid19	70	20 & 132
Malaria-Cell [45]	12	Microscopy imaging	Malaria	344	345 & 330
Private Data	**ID**	**Imaging Modalities**	**Associated Diseases**	**Training Data Normal**	**Testing Data Normal & Abnormal**
TongueNet-A [46]	13	Camera imaging	Rotten coating tongue	519	484 & 9
TongueNet-B [46]	14	Camera imaging	Peeled coating tongue	454	1363 & 30

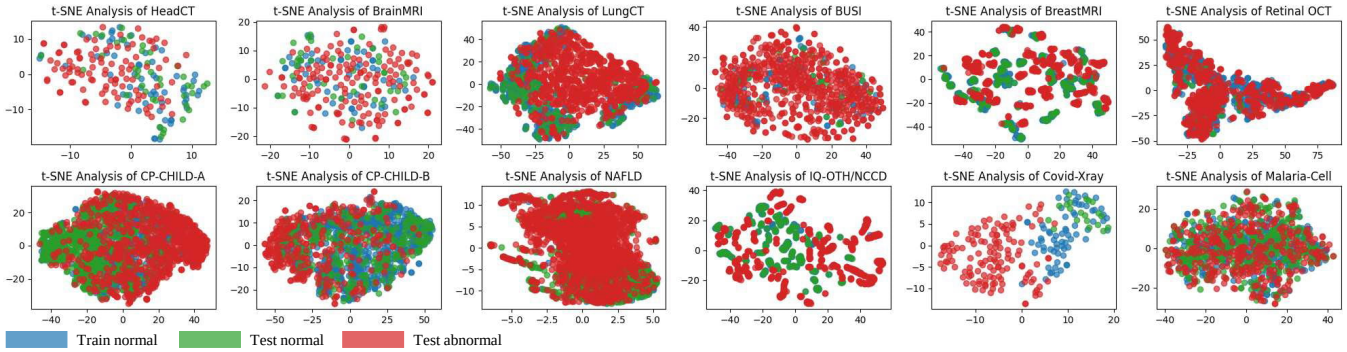


Fig. 4. t-SNE embeddings of train normal images (blue), test normal images (green) and test abnormal images (red) from the twelve public datasets.

D. Anomaly Detection and Localization

In the testing phase, we first implement anomaly localization. As shown in Fig. 3 (a), KD-based methods assume a low similarity between F_t and F_t when inputting abnormal images. Based on this assumption, for image $x_i \in R^{w\times h\times c}$, we calculate the anomaly maps of each scale $M_l(x_i) \in R^{w_l\times h_l}$ through (1),and upsample all anomaly maps to the same size as the input image through bilinear interpolation, finally the final anomaly map $M_{ano}(x_i) \in R^{w\times h}$ is the element-wise accumulation of them. Fig. 3 (b) shows the acquisition process of the final anomaly map:

$$M_{ano}(x_i) = \sum_{l=1}^{a-1} Upsample\left(M_l(x_i)\right) \tag{7}$$

For anomaly detection, we directly take the maximum value of $M_{ano}(x_i)$ as the anomaly score of x_i:

$$S_{AD}(x_i) = max\left(M_{ano}(x_i)\right) \tag{8}$$

IV. EXPERIMENTS

A. Datasets

Fourteen datasets, including two private datasets collected in our work are adopted to evaluate the proposed method as organized benchmarks. The key information and division methods of these datasets are shown in Table I. It should be noted that the training dataset only contains normal images, while the testing dataset consists of both normal and abnormal images. And the abnormal images are provided by patients with associated diseases shown in Table I. To illustrate the distribution of anomalies across the datasets, we generate t-SNE embeddings for both normal and abnormal images, as shown in Fig. 4. Additionally, to facilitate subsequent comparisons, we assign each dataset with an ID.

The difference between our organized AD benchmarks and the datasets used in previous works are as follows:

More Dataset: Most previous works verified their model on less than five datasets. For example, an OCT dataset and an Xray dataset is used in SALAD [4]. MAMA Net [9] and ProxyAno [10] were compared with other models on three datasets. The benchmark published by Cai et al. [67] only includes three CXR datasets, one brain MRI dataset, and one retina fundus image dataset. As we know, more datasets with different imaging modalities will help to fully test the generality of the model. Therefore, we propose the organized benchmarks with fourteen datasets encompassing various imaging modalities such as CT, MRI, ultrasound, colonoscopy, X-ray, and so on.

Small Sample Size: Due to the high cost of collecting medical data, the medical dataset usually includes less than one thousand medical samples. To adapt to the situation, we set the number of samples in the training dataset to less than

TABLE II
ANOMALY DETECTION AUROC (%) ON CONSIDERED DATASETS. THE BEST FOUR RESULTS ARE SHOWN IN **RED**, RED, BLUE AND **BLACK** FONT

Methods	Dataset ID															
	1	2	3	4	5	6	7	8	9	10	11	12	*AVG*	13	14	*AVG*
FCDD [62]	81.1	59.4	65.5	67.5	70.8	78.9	79.6	79.0	82.2	92.2	92.1	83.8	*77.7*	81.3	64.1	*72.7*
Ganomaly [5]	74.0	73.8	58.6	68.0	73.2	64.2	75.2	92.8	71.2	98.7	30.5	57.9	*69.8*	70.8	45.7	*58.3*
f-AnoGAN [8]	**82.6**	69.5	89.1	77.1	90.4	70.7	**97.0**	95.5	83.8	89.2	93.8	67.3	*83.8*	83.6	63.3	*73.5*
PuzzleAE [11]	83.8	71.1	75.2	81.8	**89.9**	75.1	92.7	96.4	79.9	**98.9**	69.7	63.5	*81.5*	81.9	55.9	*68.9*
CutPaste [23]	73.0	67.0	83.3	79.4	86.7	93.9	89.6	87.4	88.4	95.8	94.7	71.3	***84.2***	72.4	66.7	*69.6*
OrthoAD [63]	70.4	82.1	71.2	88.0	62.4	89.6	76.6	63.3	85.6	98.1	78.3	73.8	*78.3*	77.1	56.0	*66.6*
PaDiM [51]	77.1	85.3	77.0	**93.8**	80.4	**96.3**	85.0	80.9	79.5	98.1	88.2	87.4	*85.8*	73.6	66.1	*69.9*
CFA [55]	68.8	76.1	71.2	85.6	58.2	67.8	75.0	64.2	74.9	82.0	79.1	59.6	*71.9*	53.8	54.1	*54.0*
Patchcore [52]	72.4	65.5	80.5	81.5	72.0	94.5	97.8	**96.5**	91.1	84.3	78.5	84.8	*83.3*	80.8	70.3	***75.6***
STPM [28]	75.7	77.9	79.0	82.5	73.5	62.9	82.0	85.9	85.1	94.7	90.6	94.0	*82.0*	87.0	61.2	*74.1*
DFC [64]	65.6	70.2	74.2	91.7	**95.9**	88.5	92.8	92.0	87.9	93.8	83.3	**97.5**	*86.1*	72.9	58.1	*65.5*
RDAD [31]	74.3	**80.9**	**89.3**	**88.6**	85.9	88.5	89.1	80.5	89.4	95.6	88.2	83.3	*86.1*	92.4	**69.2**	*80.8*
DevNet [65]	84.2	63.3	**92.1**	68.4	55.9	84.8	97.6	82.1	**89.8**	90.0	**98.9**	**92.1**	*83.3*	80.1	63.3	*71.7*
DRA [66]	83.8	67.4	90.9	69.6	54.4	85.2	**98.1**	85.4	**92.2**	83.2	97.5	71.7	*81.6*	**85.6**	82.3	*84.0*
Ours	**85.7**	**88.2**	90.4	92.5	93.5	**91.3**	96.5	**95.3**	90.3	**97.3**	**94.1**	95.2	***92.5***	**94.8**	**83.1**	***89.0***

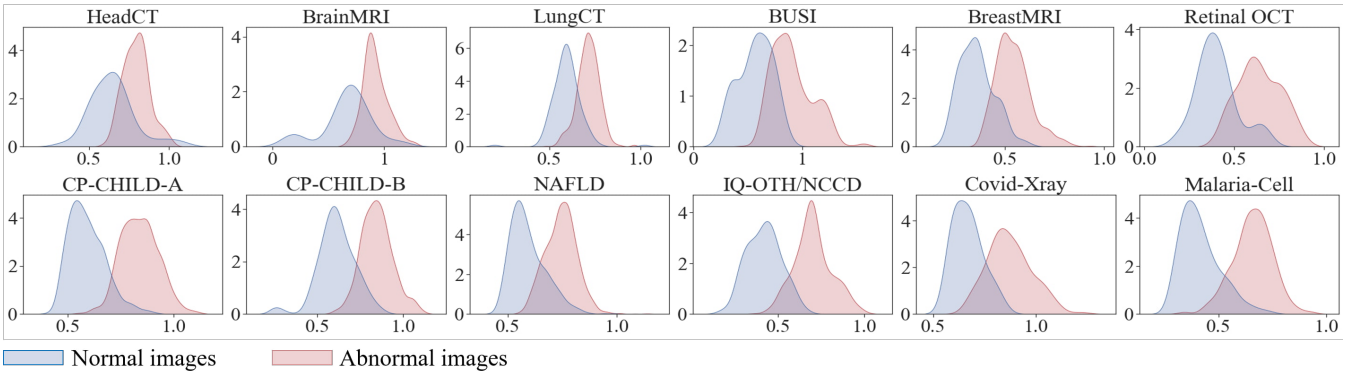


Fig. 5. Kernel density estimate (KDE) plots of anomaly scores for all considered public datasets by our method (x-axis: anomaly score S_{AD}, and y-axis: density).

a thousand. In addition, as can be seen from Table I, for some datasets (HeadCT, BrainMRI, BUSI, and Covid-Xray), no more than 100 images are used as training data in our work to obtain the AD performance of models with such very small samples.

Dedicated Datasets for Anomaly Detection: In real application, AD models are usually used to find abnormal images when there are too few abnormal images to train supervised models. Therefore, we collect two private datasets (TongueNet-A and TongueNet-B) to demonstrate such scenarios. The tongue images in the datasets were collected from several traditional Chinese medicine (TCM) hospitals and labelled by 30 senior doctors. According to the pathological characteristics of tongue coating, the abnormal tongue images in TongueNet-A has rotten coating while peeled coating for TongueNet-B. It can be seen from Table I that our private datasets contain only 9 and 30 abnormal images respectively, which makes it impossible to train the supervised model to achieve binary classification, and only AD models are able to detect abnormal images.

B. Implementation Details

All images in datasets are uniformly resized to 256×256. We use Adam optimizer [68] with the learning rate of 0.005 and $\beta = (0.5, 0.999)$. The training epoch is set 200, and batch size is 16. For the T-Net, we use the pre-trained WideResNet50, and the architecture of S-Net is symmetrical to that of the WideResNet50. Note that we only train the S-Net and freeze the T-Net. All experiments are implemented using PyTorch and conducted on the NVIDIA GeForce RTX 3090.

C. Evaluation Metrics

Considering that the choice of threshold will greatly affect indicators such as F1-score [10], we adopt Area Under the Receiver Operating Characteristic (AUROC) [58] to evaluate AD performance. The AUROC measures how well the model can separate normal samples from abnormal ones. In order

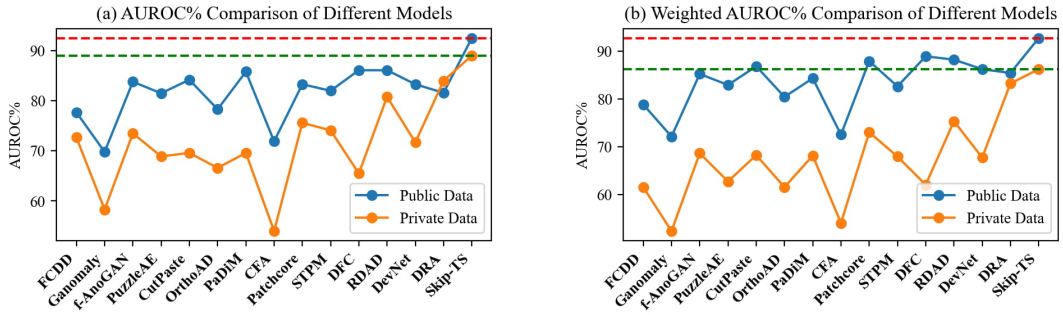


Fig. 6. Anomaly detection AUROC% and weighted AUROC% comparison of different models.

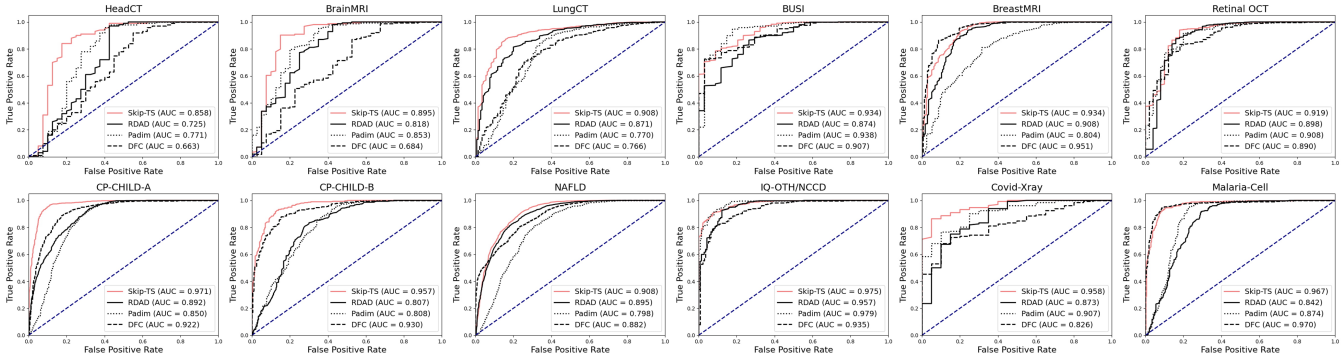


Fig. 7. AUROC curves of the top four performing models (**Skip-TS**, RDAD, DFC, Padim) on public datasets.

to verify the robustness of our models, we test the last epoch under five different random number seeds as the final result for all models. Note that this is a very harsh experimental setup and we do not allow any unprincipled early stops [11], [29].

D. Competing Methods

We compare our method with several state-of-the-art AD methods, including one-class classification model FCDD [62], reconstruction-based PuzzleAE [11], f-AnoGAN [8], Ganomaly [5], pseudo-outlier augmentation-based CutPaste [23], feature extraction-based CFA [55], PaDiM [51], OrthoAD [63], and Patchcore [52], KD-based RDAD [31], STPM [28], and DFC [64]. In addition, we also test two state-of-the-art open-set supervised models for anomaly detection, DevNet [65] and DRA [66], to observe the limitations of our proposed models. Such open-set supervised models need to use a few labeled abnormal images as examples. In the experiments, we use the code provided in the GitHub to implement above methods. In training DevNet and DRA, we use training dataset and one abnormal image.

E. Anomaly Detection Results

We report the AUROC (%) of our method and competing methods on fourteen datasets in Table II, and highlight the best four results on each dataset. It can be observed that for the average outcome our method exceeds state-of-the-art by **6.4%** (86.1% $\rightarrow$ 92.5%) on public datasets, and **8.2%** (80.2% $\rightarrow$ 89.0%) on private datasets. However, the dataset has wide range of sample size hence simply averaging over AUROC percentages could lead to biased result. To solve the problem, we also calculate the weighted AUROC, where the size of each dataset is used as a weighting factor during the averaging process. The results, as shown in Fig. 6(b), demonstrate that Skip-TS is the best-performing model, achieving weighted AUROC of **92.7%** and **86.2%** on the public and private datasets, respectively. In addition, in order to visualize the misclassified and correctly classified classes, we also generated the AUROC curves for the top four ranked models on the public dataset, as depicted in Fig. 7. Further analysis of the data in Table II, we can get the following conclusions:

1) Generalization: Excluding the open-set supervised models DevNet and DRA, our Skip-TS emerges as the sole model that consistently ranks top four on all datasets. This demonstrates that the proposed method has robust generalization and works effectively on datasets with different imaging modalities, whereas other methods are only applicable on a subset of datasets.

2) Small-sample Learning: For HeadCT, BrainMRI, BUSI, and Covid-Xray with IDs of 1,2,4,11 respectively, less than 100 images are used for training. For HeadCT and BrainMRI, our method reaches new state-of-the-art AUROC of **85.7%** and **88.2%**, respectively. In addition, with our method the AUROC is only **1.3%** (93.8% $\rightarrow$ 92.5%) lower than that of PaDiM on BUSI and only **0.6%** (94.7% $\rightarrow$ 94.1%) lower than that of CutPaste on Covid-Xray. Therefore, the proposed method has the best performance overall on the small-sample dataset.

3) Pre-trained models: Upon analyzing the average results, we find that the top four performing methods use pre-trained models. Notably, for the public datasets, our proposed method

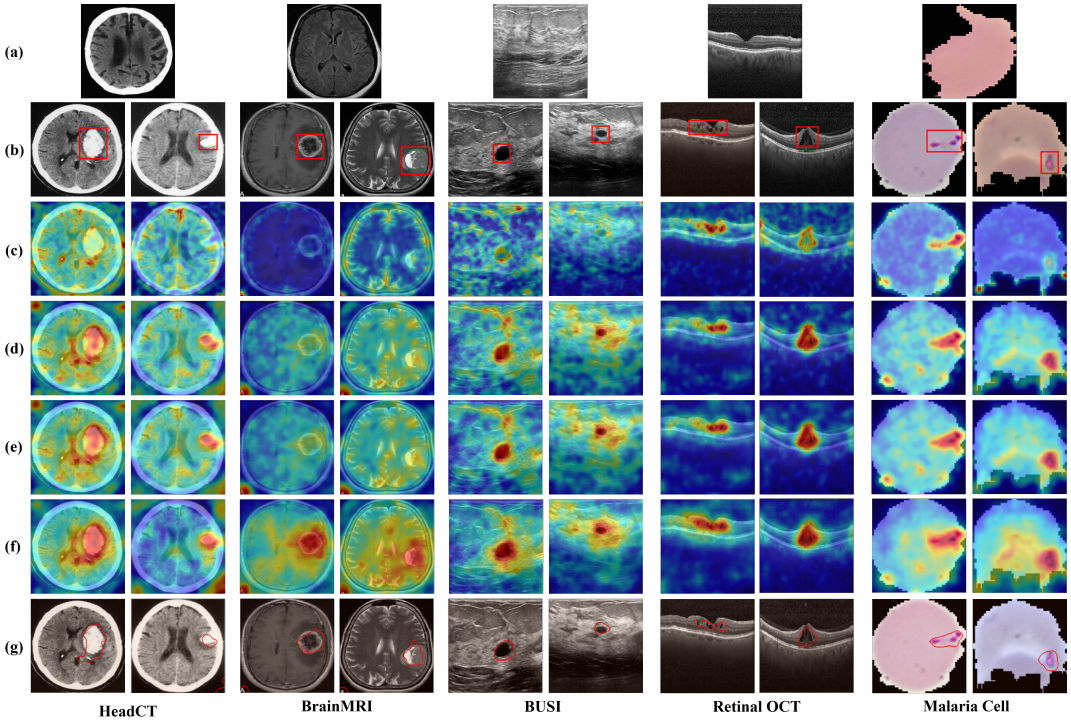


Fig. 8. The anomaly localization results of proposed method. Our model can precisely localize anomalies, whereas most reconstruction-based methods can only roughly detect them. (a) normal images; (b) abnormal images (the red rectangle denotes the abnormal area); (c) anomaly map $M1$; (d) anomaly map $M2$; (e) anomaly map $M3$; (f) final anomaly map M_{ano}; (g) segmentation results.

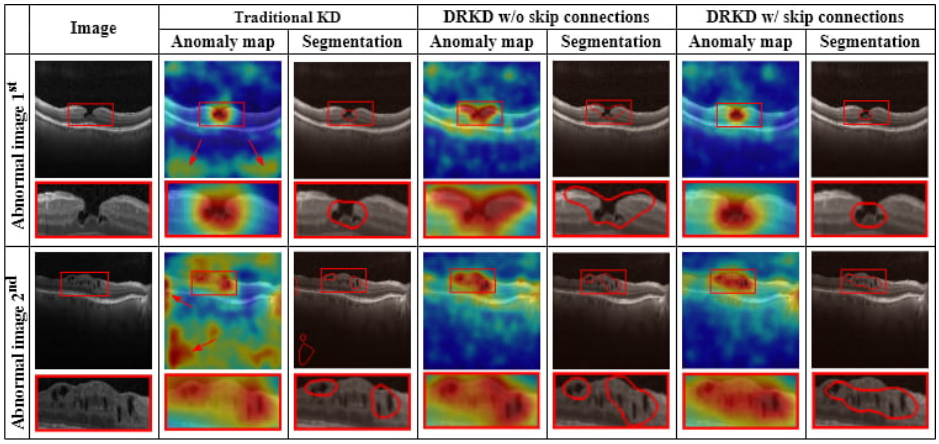


Fig. 9. The anomaly localization results with different distillation paradims.

ranks first, followed by RDAD, DFC, and PaDiM, respectively. Similarly, for the private datasets, the order of the top-performing methods are Ours, RDAD, Patchcore, and STPM. This outcome is largely attributed to the efficacy of pre-trained models in feature extraction, a trait which our method leverages to its fullest potential. It is worth noting, however, that most AD methods for medical images do not employ pre-trained models but instead prioritize training generative models from scratch [4], [7]–[10], [12].

4) Open-set Supervised Models: open-set supervised models utilize a limited number of abnormal images as a training set to enhance performance. Table II indicates that DevNet outperforms all unsupervised models on LungCT and Covid-Xray, while DRA achieves the state-of-the-art on CP-CHILD-A and NAFLD. Nevertheless, the average performance of DevNet and DRA on public datasets is inadequate, especially with regards to the AUROC on BreastMRI, which is remarkably low (55.9% and 54.4%, respectively). This is due to their susceptibility to overfitting to visible anomalies, resulting in low robustness, which makes them unsuitable for datasets with variable anomaly images. Therefore, unsupervised models remain the best option for medical AD.

5) Kernel density estimate: Finally, kernel density estimate (KDE) plots of anomaly scores S_{AD} of our method are shown in Fig. 5. We can intuitively see that the S_{AD} of normal (blue) and abnormal (red) images have non-overlapping distributions,

TABLE III
ABLATION STUDY OF THE PROPOSED METHOD (SC: SKIP CONNECTIONS)

	Public datasets	Private datasets
Distillation paradigm and skip connections		
Traditional KD	82.6	76.2
DRKD w/o SC	88.0	73.6
Loss functions		
$L_{sim}\,(MSE)$	92.4	86.9
$L_{sim}\,(MSE)$ + L_{mac}	92.2	86.7
$L_{sim}\,(Cosine)$	92.0	87.8
Multi-scale feature		
M1+M2	86.4	78.3
M1+M3	91.4	87.6
M2+M3	92.5	87.8
Our method		
DRKD & SC & $(L_{sim}\,(Cosine)$+$L_{mac})$ & M_{ano}	**92.5**	**89.0**

which indicates the proposed method has good anomaly detection ability.

F. Anomaly Localization Results

Noting that the anomalies in some datasets concentrate in local regions, we test the anomaly localization ability of proposed method on these datasets. As shown in Fig. 8, we select five datasets HeadCT, BrainMRI, BUSI, Retinal OCT, and Malaria Cell to show the segmentation results in each step. The original medical images is shown in Fig. 8(a) and the abnormal images marked with red rectangles is shown in Fig. 8(b). First, we obtain the anomaly maps $M1$, $M2$, and $M3$ by calculating the feature differences of each hierarchies (shown in Fig. 8(c)-(e) respectively). Secondly, we get the final anomaly map M_{ano} (shown in Fig. 8(f)) according to (7) and Gaussian filtering with $\sigma = 4$. Setting the threshold as 0.8, the segmentation results are shown in Fig. 8(g). Different from reconstruction-based methods that can only roughly locate anomalies by computing pixel-level differences [4], [8]–[10], our proposed method enables precise recognition and localization of anomalies on diverse images (shown in Fig. 8(g)), which are similar to those obtained by the semantic segmentation model [69] without using abnormal image and mask for training. Moreover, our analysis reveals that the segmentation outputs produced by our method fail to fully capture the abnormal regions in the HeadCT and BrainMRI datasets, implying that the performance of location may be hindered by the limited availability of training data. However, existing studies often ignore the impact of limited data.

G. Computational Costs and Parameter Quantities

We compare the computational costs and parameter quantities of our proposed model, Skip-TS, with the state-of-the-art model, RDAD. The experimental results are presented in Table IV, where it can be observed that our model exhibits superior computational efficiency. Our model requires only **2.98E+10** FLOPs, which is 20% fewer than the FLOPs required by RDAD (3.72E+10 FLOPs). Additionally, our model has 31% fewer parameters, with a count of **1.33E+08**, compared to RDAD's parameter count of 1.93E+08. This difference arises from our use of simple skip connections to transfer multi-scale features extracted by the T-Net, instead of employing a trainable one-class bottleneck embedding module.

TABLE IV
THE COMPUTATIONAL COSTS OF SKIP-TS AND RDAD

Model	Computational Complex	
	FLOPs	Parameters
RDAD	3.72E+10	1.93E+08
Skip-TS	**2.98E+10**	**1.33E+08**

H. Interpretability Analysis

To analyze which image features or regions are the most important for detecting anomalies, we choose CP-CHILD-B [41] dataset as an example. Since three scale features extracted by T-Net and S-Net respectively are used for the computation of the three anomaly maps, $M1$, $M2$, and $M3$, we conduct t-SNE analysis on the multi-scale features F_t^1, F_t^2, and F_t^3 extracted by T-Net, as well as the features F_s^1, F_s^2, and F_s^3 extracted by S-Net. The analysis results are shown in Fig. 10, from which we can see that for T-Net, the extracted features F_t^1, F_t^2, and F_t^3 exhibit a significant overlap between the normal and anomalous data after t-SNE dimensionality reduction. This is because that although pre-trained models are good at extracting features from images, they are unable to specifically highlight the anomalous features of images, especially in medical images that differ significantly from natural images. Therefore, models directly learning the feature distribution extracted by pre-trained models, i.e., feature-based models (orthoAD [63], PaDiM [51], CFA [55], and Patchcore [52]), demonstrate suboptimal performance in medical image anomaly detection. However, for S-Net trained only on normal data, normal and abnormal images can be clearly distinguished as illustrated in the first row in Fig. 10.

We further evaluate the AUROC (%) of our model using anomaly maps with different scales. When using only $M1$ obtained from F_s^1 and F_t^1, the AUROC (%) is 63.8, whereas when using only $M2$ and $M3$, it is 85.7 and 95.1 respectively. When combining $M1$, $M2$, and $M3$, the AUROC (%) is 95.7. Therefore, the importance of AD across the three scales is ranked as follows: $M3 > M2 > M1$. This aligns with our ablation experimental results, the AUROC (%) ranking is $M2{+}M3 > M1{+}M3 > M1{+}M2$ when calculating anomaly scores using various anomaly map combinations (shown in Table III).

In summary, for Skip-TS, the depth features F_t^3 and F_s^3 play the most significant role for anomaly detection, and the optimal anomaly detection performance is achieved by fusing the three-scale features.

I. Ablation Study

After achieving state-of-the-art performance on the organized benchmark, we conduct an ablation study with the proposed method in terms of distillation paradigm, skip connections, loss functions and multi-scale feature. We delete one

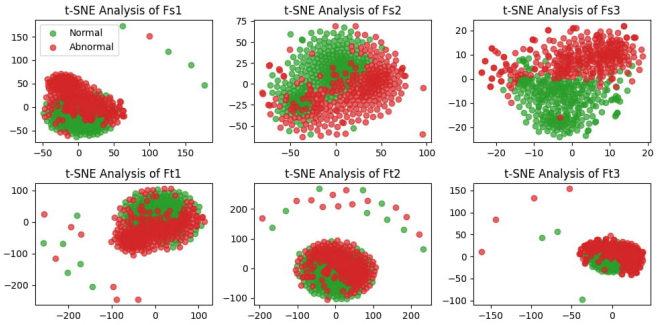


Fig. 10. t-SNE embeddings of multiscale features $\{F_t^1, F_t^2, F_t^3\}$ and $\{F_s^1, F_s^2, F_s^3\}$ extracted from the CP-CHILD-B dataset by T-Net and S-Net.

component each time and keep the other parts unchanged to test the function of each module. The results are shown in Table III.

1) DRKD and SC for Anomaly Detection: To begin, we verify the effectiveness of DRKD and skip connections in improving the performance of our model. Although our model performance comprehensively surpasses the STPM using the traditional KD paradigm, as shown in Table II. To ensure network structure consistency, we redesign a T-S model trained using traditional KD. Both T-Net and S-Net are set as encoders with WideResNet50 as backbone. Other experimental settings remained the same. As can be seen from Table III, compared with traditional KD, DRKD paradigm improves AUROC by **5.4%** (82.6% $\rightarrow$ 88.0%) on public datasets, but slightly decreases on the private dataset (76.2% $\rightarrow$ 73.6%). Skip connections further extend the advantages of DRKD, increasing the AUROC on public and private datasets to **92.5%** and **89.0%**, respectively, achieving state-of-the-art performance.

2) DRKD and SC for Anomaly localization: We qualitatively investigate effect of DRKD and skip connections for anomaly localization, and the results are shown in Fig. 9. As discussed in Section III.B and C, the traditional KD paradigm is not sensitive enough to anomalies and DRKD without skip connections may lead to false identification of normal regions as anomalies. We can see these phenomena in Fig. 9. For traditional KD, we first pay attention to the anomaly map, from which we find that there are a large number of high-heat areas (pointed by the arrow) outside the anomaly area. This is because the structural similarity of the T-S model in traditional KD paradigm makes it non-discriminative for anomalies. Although the segmentation result for the first anomaly image is good, it cannot accurately locate the second anomaly image and is accompanied by noise. For DRKD without skip connections, due to the difficulty of the S-Net to adequately reconstruct normal regions, the segmentation range is too large for the first abnormal image. For the second one, although the false positive area is reduced compared with the traditional KD, all abnormal areas are still not detected. However, after adding skip connections, the T-S model outputs accurate segmentation results for both abnormal images.

3) Loss Function: We compare four loss function selection strategies. For similarity loss L_{sim}, besides the cosine similarity loss function, the mean squared error (MSE) loss function is also commonly used. From Table III, we can see that the combination of $L_{sim}\,(Cosine)$ and multi-scale anomaly consistency loss L_{mac}, *i.e.* our method achieves the best performance.

4) Multi-scale feature: Since the feature maps from each layer of the pre-trained network can identify abnormalities of different scales [50], in Section III.D we advocate accumulating all anomaly maps $M1, M2, M3$ to calculate the abnormal score of each image. In this subsection, we explore the influence of multi-hierarchical features. From Table III we see that removing $M1$ brings the least impact on overall performance, reducing AUROC by only **1.2%** (89.0% $\rightarrow$ 87.8%) on private datasets. However, deleting $M3$ makes AUROC drop sharply, because $M3$ contains more dense abnormal information as a deep feature.

J. Limitations

The proposed Skip-TS model still has some limitations that will be addressed in our future work: (1) although AUROC has been improved by Skip-TS in anomaly detection, it is still difficult for Skip-TS to realize accurate anomaly localization on the datasets with very small samples like BrainMRI and HeadCT. The segmentation results for these datasets, as seen in Fig. 8(g), show that the abnormal part can be located, but not fully covered, which is due to the limited training data. We aim to enhance the abnormal localization performance with very small samples in the future. (2) Skip-TS currently can't detect anomalies in 3D medical images because it relies on 2D pre-trained networks. However, 3D images, such as 3D MRI [70] and 3D CT images [71], are common in clinical practice. We will improve our method for 3D medical images in the future. (3) the proposed AD model is exclusively trained on normal samples. When noisy anomalous samples are included in the training dataset, its performance will decline. Some supervised learning models attempt to address the problem of noisy labels. For example, in MST-TS [72], a meta-self-training method is introduced, which employs a self-training mechanism to train a teacher network and leverages the pseudo-labels generated by the teacher to train a student network. In the future, we will improve the robustness of our unsupervised AD model to noisy data.

V. Conclusion

In this paper, we propose a novel anomaly detection method for medical images based on knowledge distillation. To address the problem of insensitivity to anomalies caused by the structural similarity between the teacher and student networks, we propose a direct reverse knowledge distillation paradigm. In this approach, the S-Net serves as a decoder and receives input directly from the T-Net. Moreover, we introduce skip connections between T-Net and S-Net to help S-Net recover anomaly-free image representations. To further enhance the performance of our T-S model, we also design a multi-scale anomaly consistency loss. Experimental results demonstrate that our method surpasses state-of-the-art by 6.4% on public datasets and 8.2% on private datasets. Additionally, compared to reconstruction-based methods, our approach achieves

anomaly localization similar to that of semantic segmentation models, thus expanding the potential applications of medical anomaly detection. Future work will focus on improving the model's ability to locate anomalies and exploring additional applications in medicine.

REFERENCES

[1] S. Pei, C. Wang, S. Cao, and Z. Lv, "Data augmentation for fmri-based functional connectivity and its application to cross-site adhd classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[2] G. Yue, P. Wei, Y. Liu, Y. Luo, J. Du, and T. Wang, "Automated endoscopic image classification via deep neural network with class imbalance loss," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[3] Y. Huang, G. Liu, Y. Luo, and G. Yang, "Adfa: Attention-augmented differentiable top-k feature adaptation for unsupervised medical anomaly detection," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 206–210.

[4] 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.

[5] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 622–637.

[6] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[7] 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 *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*. Springer, 2017, pp. 146–157.

[8] 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.

[9] Y. Chen, H. Zhang, Y. Wang, Y. Yang, X. Zhou, and Q. J. Wu, "Mama net: Multi-scale attention memory autoencoder network for anomaly detection," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 1032–1041, 2020.

[10] 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.

[11] M. Salehi, A. Eftekhar, N. Sadjadi, M. H. Rohban, and H. R. Rabiee, "Puzzle-ae: Novelty detection in images through solving puzzles," *arXiv preprint arXiv:2008.12959*, 2020.

[12] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," *IEEE Access*, vol. 9, pp. 118 571–118 583, 2021.

[13] J. Luo, J. Lin, Z. Yang, and H. Liu, "Smd anomaly detection: A self-supervised texture–structure anomaly detection framework," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[14] K. Song, H. Yang, and Z. Yin, "Anomaly composition and decomposition network for accurate visual inspection of texture defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[15] X. Li, J. Jing, J. Bao, P. Lu, Y. Xie, and Y. An, "Otb-aae: Semi-supervised anomaly detection on industrial images based on adversarial autoencoder with output-turn-back structure," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2023.

[16] K. Xiao, J. Cao, Z. Zeng, and W.-K. Ling, "Graph-based active learning with uncertainty and representativeness for industrial anomaly detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–14, 2023.

[17] H. Liang, L. Song, J. Du, X. Li, and L. Guo, "Consistent anomaly detection and localization of multivariate time series via cross-correlation graph-based encoder–decoder gan," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[18] R. Liu, W. Liu, H. Li, H. Wang, Q. Geng, and Y. Dai, "Metro anomaly detection based on light strip inductive key frame extraction and magan network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[19] M. Niu, Y. Wang, K. Song, Q. Wang, Y. Zhao, and Y. Yan, "An adaptive pyramid graph and variation residual-based anomaly detection network for rail surface defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[20] Y. Yan, D. Wang, G. Zhou, and Q. Chen, "Unsupervised anomaly segmentation via multilevel image reconstruction and adaptive attention-level transition," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[21] C. Zhang, Y. Wang, and W. Tan, "Mthm: Self-supervised multitask anomaly detection with hard example mining," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[22] H. Yang, Z. Zhu, C. Lin, W. Hui, S. Wang, and Y. Zhao, "Self-supervised surface defect localization via joint de-anomaly reconstruction and saliency-guided segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.

[23] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9664–9674.

[24] J. Song, K. Kong, Y.-I. Park, S.-G. Kim, and S.-J. Kang, "Anomaly segmentation network using self-supervised learning," in *AAAI 2022 Workshop on AI for Design and Manufacturing (ADAM)*, 2021.

[25] J. Sato, Y. Suzuki, T. Wataya, D. Nishigaki, K. Kita, K. Yamagata, N. Tomiyama, and S. Kido, "Anatomy-aware self-supervised learning for anomaly detection in chest radiographs," *arXiv preprint arXiv:2205.04282*, 2022.

[26] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.

[27] 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*, 2020, pp. 4183–4192.

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

[29] 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.

[30] Y. Cao, Q. Wan, W. Shen, and L. Gao, "Informative knowledge distillation for image anomaly segmentation," *Knowledge-Based Systems*, vol. 248, p. 108846, 2022.

[31] 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.

[32] R. Xu, Y. Wang, X. Ye, P. Wu, Y.-W. Chen, F. Xu, W. Zhu, C. Chen, Y. Zhou, H. Hu *et al.*, "Pixel-level and affinity-level knowledge distillation for unsupervised segmentation of covid-19 lesions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1376–1380.

[33] Q. Wan, L. Gao, and X. Li, "Logit inducing with abnormality capturing for semi-supervised image anomaly detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[34] S. Yamada, S. Kamiya, and K. Hotta, "Reconstructed student-teacher and discriminative networks for anomaly detection," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2725–2732.

[35] M. Toğaçar, Z. Cömert, B. Ergen, and Ü. Budak, "Brain hemorrhage detection based on heat maps, autoencoder and cnn architecture," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. IEEE, 2019, pp. 1–5.

[36] S. Rai, S. Chowdhury, S. Sarkar, K. Chowdhury, and K. P. Singh, "A hybrid approach to brain tumor detection from mri images using computer vision," *Journal of Innovation in Computer Science and Engineering*, vol. 8, no. 2, pp. 8–12, 2019.

[37] D. Gaurav, "Covid-19 detection," https://www.kaggle.com/datasets/gauravduttakiit/covid19-detection.

[38] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[39] K. Uzair, "Breast cancer patients mri's," https://www.kaggle.com/datasets/uzairkhan45/breast-cancer-patients-mris.

[40] 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.

[41] W. Wang and T. Jinge, "Cp-child.zip. figshare. dataset," https://doi.org/10.6084/m9.figshare.12554042.v1.

[42] I. Zingman, B. Stierstorfer, C. Lempp, and F. Heinemann, "Learning image representations for anomaly detection: application to discovery of histological alterations in drug development," *arXiv preprint arXiv:2210.07675*, 2022.

[43] h. alyasriy, "The iq-othnccd lung cancer dataset," https://data.mendeley.com/datasets/bhmdr45bh2/1.

[44] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[45] Arunava, "Malaria cell images dataset," https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria.

[46] M. Liu, Y. Jiao, and H. Chen, "Skip-st: Anomaly detection for medical images using student-teacher network with skip connections," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2023, pp. 1–5.

[47] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[50] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2020.

[51] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*. Springer, 2021, pp. 475–489.

[52] 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.

[53] M. Xu, X. Zhou, X. Gao, W. He, and S. Niu, "Discriminative feature learning framework with gradient preference for anomaly detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.

[54] Y. Liu, X. Gao, J. Z. Wen, and H. Luo, "Unsupervised image anomaly detection and localization in industry based on self-updated memory and center clustering," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.

[55] S. Lee, S. Lee, and B. C. Song, "Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," *IEEE Access*, vol. 10, pp. 78 446–78 454, 2022.

[56] K. Nakazawa, K. Hotta, J. Yu, and C. Zhang, "Student-teacher anomaly detection considering knowledge consistency between layer groups," in *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2022, pp. 381–382.

[57] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1038–1059, 2021.

[58] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[59] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[60] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2528–2535.

[61] H. Li and Y. Li, "Anomaly detection methods based on gan: a survey," *Applied Intelligence*, vol. 53, no. 7, pp. 8209–8231, 2023.

[62] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K. R. Muller, "Explainable deep one-class classification," in *International Conference on Learning Representations*, 2021.

[63] J.-H. Kim, D.-H. Kim, S. Yi, and T. Lee, "Semi-orthogonal embedding for efficient unsupervised anomaly segmentation," *arXiv preprint arXiv:2105.14737*, 2021.

[64] J. Yang, Y. Shi, and Z. Qi, "Learning deep feature correspondence for unsupervised anomaly detection and segmentation," *Pattern Recognition*, vol. 132, p. 108874, 2022.

[65] G. Pang, C. Ding, C. Shen, and A. v. d. Hengel, "Explainable deep few-shot anomaly detection with deviation networks," *arXiv preprint arXiv:2108.00462*, 2021.

[66] C. Ding, G. Pang, and C. Shen, "Catching both gray and black swans: Open-set supervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7388–7398.

[67] 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," *arXiv preprint arXiv:2210.04227*, 2022.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[69] M. Z. Khan, M. K. Gajendran, Y. Lee, and M. A. Khan, "Deep neural architectures for medical image semantic segmentation," *IEEE Access*, vol. 9, pp. 83 002–83 024, 2021.

[70] Q. Tian, B. Bilgic, Q. Fan, C. Liao, C. Ngamsombat, Y. Hu, T. Witzel, K. Setsompop, J. R. Polimeni, and S. Y. Huang, "Deepdti: High-fidelity six-direction diffusion tensor imaging using deep learning," *NeuroImage*, vol. 219, p. 117017, 2020.

[71] H. Chen, X. He, H. Yang, J. Feng, and Q. Teng, "A two-stage deep generative adversarial quality enhancement network for real-world 3d ct images," *Expert Systems with Applications*, vol. 193, p. 116440, 2022.

[72] X. Pu and C. Li, "Meta-self-training based on teacher–student network for industrial label-noise fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.