Skip-ST: Anomaly Detection for Medical Images Using Student-Teacher Network with Skip Connections

1st Mingxuan Liu
Department of biomedical engineering
Tsinghua University
Beijing, China
liumx19@mails.tsinghua.edu.cn

2nd Yunrui Jiao
School of Integrated Circuits
Tsinghua University
Beijing, China
jyr19@mails.tsinghua.edu.cn

3rd Hong Chen
School of Integrated Circuits
Tsinghua University
Beijing, China
hongchen@tsinghua.edu.cn

Abstract—Anomaly detection (AD) aims to recognize abnormal inputs in testing data when only normal data are available during training. Most AD models perform well on specific datasets but are difficult to generalize to other tasks, especially on medical datasets with high heterogeneity. In this paper, we propose a student-teacher network with skip connections (Skip-ST) which is trained by a novel knowledge distillation paradigm called direct reverse knowledge distillation (DRKD) to realize AD. Skip-ST consists of a pretrained teacher encoder and a randomly initialized student decoder. The output of the teacher encoder's last layer is the input of the student decoder, which aims to recover the multi-scale representation extracted by the teacher encoder. We introduce skip connections between the teacher encoder and student decoder to prevent the student decoder from missing normal information of images at multi-scale. Experimental results show that Skip-ST achieves a 7.95% Area Under the Receiver Operating Characteristic (AUROC) improvement averagely on five challenging medical datasets, outperforming the state-of-the-art AD models. Our code is available at https://github.com/Arktis2022/Skip-TS.

Keywords—anomaly detection, medical image analysis, deep learning, student-teacher network, 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]. Given that the training of an AD model necessitates only normal data to characterize the normal distribution, and normal medical images without lesion annotation are much easier to collect [4], researchers can use normal data to establish AD models for detecting anomalies in medical images. Consequently, AD has emerged as a critical component in AI-aided diagnosis.

Many efforts have been undertaken to achieve AD by building self-supervised tasks on visible normal data, which primarily encompass sample reconstruction, pseudo-outlier augmentation, and knowledge distillation (KD) [8]. In
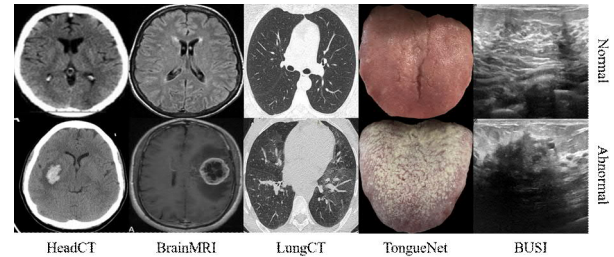
Fig. 1. Examples of normal and abnormal images for considered datasets. HeadCT: head CT slices from normal vs hemorrhage [3], BrainMRI: brain MRI images from normal vs tumors [3], LungCT: lung CT slices from normal vs covid-19 [5], TongueNet: tongue photos from normal vs rotten coating [6], BUSI: ultrasound images from normal vs breast cancer [7].

sample reconstruction-based AD methods, generative models are usually used, including generative adversarial nets (GAN) [9] and autoencoders (AE) [10]. Such AD models are claimed to be suitable for medical images. For example, f-AnoGAN [11] is dedicated to dealing with retinal OCT images. Nina Shvetsova *et al.* introduced a deep perceptual autoencoder and tested it on the NIH dataset and the Camelyon16 dataset [12]. To enforce the model to learn translation-consistent and representative features from the normal data, structure similarity loss and center constraint loss have been introduced to the reconstruction-based approaches [4]. Secondly, pseudo-outlier augmentation is widely used on industrial datasets. For example, CutPaste [13] generates pseudo outliers by cutting image patches and pasting them at random locations. AnoSeg [14] uses anomaly data generation based on hard augmentation, self-supervised learning with pixel-wise and adversarial losses, and coordinate channel concatenation. Lastly, KD based on the student-teacher network is proposed in 2020 [15]. For the student network, only normal data is visible during distillation, resulting in extracted features that may not align with those of the teacher network when abnormal data is inputted. In order to obtain multi-scale abnormal features, US [15] use an ensemble learning method, while STPM [16] choose to calculate multi-scale feature differences. In RDAD [18], an indirect reverse knowledge distillation (RKD) paradigm is realized by a trainable one-class bottleneck embedding (OCBE) module.

However, the above methods either perform well on industrial datasets MVTecAD [17] but not for complex medical data [13,14,15,16,18], or suitable for only a few medical datasets and the training process is complicated [4,11,12,19]. In this paper, we propose a simple and robust AD method, which is based on KD and verified on five medical datasets (shown in Fig. 1). Fig. 2 shows three

(a) Traditional KD paradigm [15,16]   (b) RKD paradigm [18]

(c) DRKD paradigm

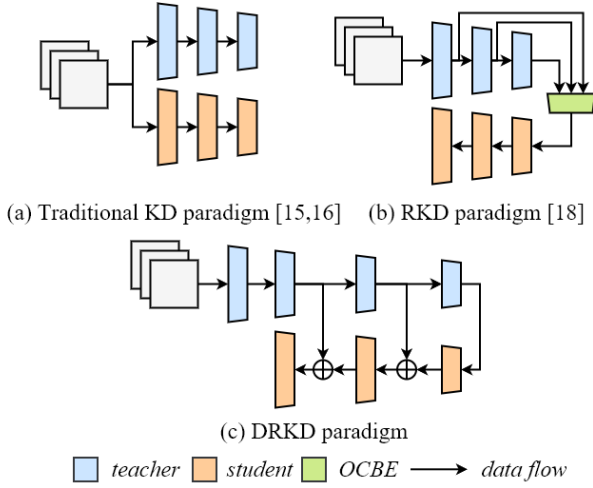□ *teacher* □ *student* □ *OCBE* ⟶ *data flow*

Fig. 2. Student-teacher network and data flow in (a) Traditional KD paradigm [15,16], (b) RKD paradigm in RDAD [18] and (c) proposed DRKD paradigm.

different student-teacher networks, with traditional KD, indirect RKD, and DRKD paradigms, respectively. With the traditional KD paradigm, the student network is an encoder with a similar structure to the teacher network [15,16]. The student network with the RKD paradigm is designed as a decoder and the output of the teacher encoder is indirectly entered into the student decoder through the OCBE module [18]. The proposed method has two important features that distinguish it from others: First, we implement a DRKD paradigm, where the student decoder directly takes the last layer's output of the teacher decoder as input, without any intermediaries. Furthermore, the student decoder is trained to reconstruct the multi-scale features extracted by the teacher encoder. Secondly, we put forward a student-teacher network called Skip-ST suitable for DRKD. This network has skip connections between the teacher encoder and the student decoder, aimed at facilitating the student's reconstruction of features. We conduct extensive experiments on five medical datasets, and the results show that our proposed AD method achieves the best performance.

II. PROPOSED METHODS

A. Problem Formulation and System Overview

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. The goal of anomaly detection is to train a model that can recognize abnormal data in Y_{test} using X_{train}.

We implement AD by training a student-teacher network through DRKD. The proposed Skip-ST consists of a pre-trained teacher encoder E and a randomly initialized student decoder D. First, we adopt a sufficiently pre-trained network to extract discriminative features of images [20], so the feature F_t extracted by the teacher encoder contains normal and abnormal information. Then, during training, the student decoder needs to reconstruct the output of the teacher encoder's each layer from the input. Since only normal data is included in X_{train}, the student decoder learns the patterns of normal images from F_t. Finally, in the testing process, because the student decoder only has knowledge of normal data, its output F_s only contains normal information and ignores anomalies for abnormal data, which leads to the

difference between F_t and F_s. Therefore, for 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, Y_{test}, can be divided into two subsets, normal and abnormal. As a result, anomaly detection is realized.

B. Direct Reverse Knowledge Distillation (DRKD)

As shown in Fig. 2(a), the student and teacher network in traditional KD paradigms such as US [15] and STPM [16] have similar structures; both receive images as input and extract features F_t and F_s, respectively. For abnormal inputs, F_t and F_s are expected to be different. However, due to the same data flow in the KD process, the dissimilarity between F_t and F_s may disappear for abnormal data, resulting in anomaly detection failure. RDAD attempts to improve the model's sensitivity to out-of-distribution data by creating an RKD paradigm (Fig.2(b)). Its student decoder is connected with the teacher encoder through a trainable OCBE module [18]. The student decoder first touches the high-level semantics of images and then gradually recovers their multi-scale representation. But due to the existence of OCBE, the features at each level extracted by the teacher encoder are densely compressed together. The student decoder may struggle to recover anomaly-free representations of images from highly abstract low-dimensional features, especially for complex medical images. So the RKD paradigm will result in a high anomaly score for normal images.

Therefore, the traditional KD paradigm is not sensitive enough to abnormal data and RKD may identify anomalies from normal images mistakenly. These problems result in a low AUROC for KD based methods. In order to address them, we put forward the DRKD shown in Fig. 2(c), from which we can see that distillation is performed on Skip-ST with an encoder-decoder architecture, and the knowledge of the teacher encoder is directly transferred to the student without any intermediary.

The proposed student-teacher network with skip connections (Skip-ST) for anomaly detection in medical images is illustrated in Fig.3. With our DRKD, teacher encoder E needs to extract comprehensive features from input images. We choose WideResNet50 [21] pre-trained on ImageNet [22] because it can output relevant features for anomaly detection [23]. The four layers of WideResNet50 are set as $E1$, $E2$, $E3$, and $E4$ respectively. $E4$ is used to generate the input of the student decoder, which can be regarded as the low-resolution feature of images. The task of the other three sub-encoders $E1 \sim E3$ is to provide the multi-scale knowledge as a reference for the student decoder. During training, the parameters of E are frozen to prevent the model from converging to a trivial solution [24]. The student decoder D has a symmetric structure with the teacher encoder E, and it aims to learn the intermediate representation of E from normal images. The symmetry allows each layer's output Fs to be consistent with Ft in dimension, and the reverse design prevents D from directly receiving abnormal data during testing. Specifically, for WideResNet50, down-sampling is realized by convolutional layers with a kernel size of 3 and a stride of 2 [21]. Correspondingly, D uses deconvolution layers [26] with a kernel size of 3 and a stride of 2 for up-sampling.

Now we describe the training process of DRKD. For the input $x_i \in \mathbb{R}^{w \times h \times c}$ where h is the height, w is the width and c is
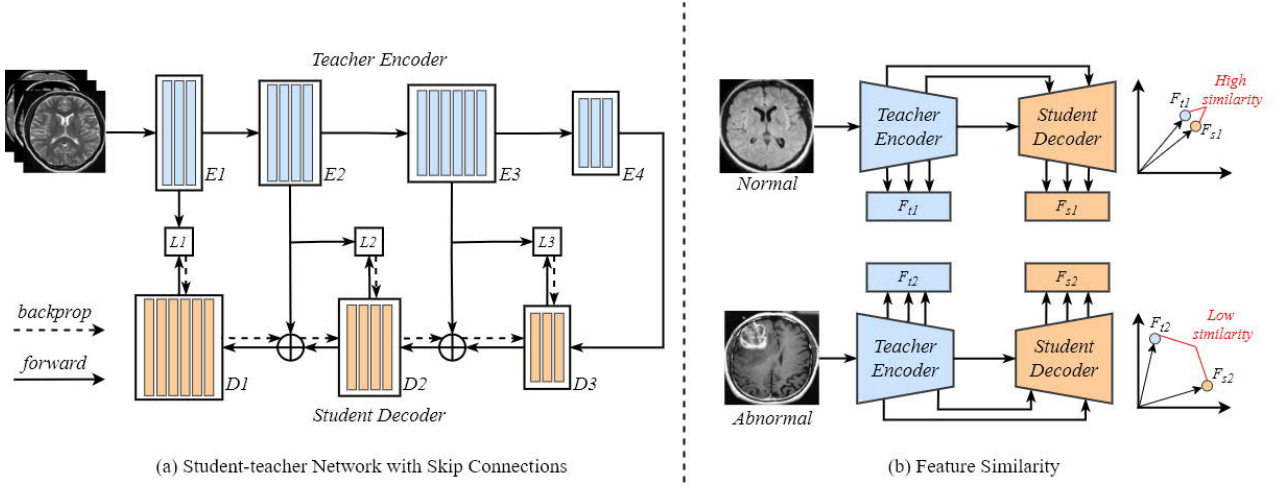
Fig. 3. Overview of student-teacher network with skip connections (Skip-ST) for anomaly detection in medical images. (a) Skip-ST consists of a pre-trained teacher encoder E and a randomly initialized student decoder D. The teacher can be divided into four sub-encoders $E1{\sim}E4$ and the student can be divided into three sub-decoders $D1{\sim}D3$. During training, the student D needs to recover the multi-scale representation extracted by the teacher E by minimizing the loss value $L1{\sim}L3$. (b) During testing, the teacher encoder outputs the real feature F_t and the student decoder outputs the anomaly-free one F_s. The anomaly score can be obtained by calculating the difference between F_t and F_s. For normal images, there should be a short distance between F_t and F_s.

the number of color channels, the kth sub-encoder of the teacher E outputs $F_t^k(x_i) \in \mathbb{R}^{w_k \times h_k \times d_k}$, 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 of E is a, and $(a\text{-}1)$ sub-decoders for D, the $(a\text{-}1)$th sub-decoder of D takes $F_t^a(x_i)$ as input and outputs $F_s^{a-1}(x_i) \in \mathbb{R}^{w_{a-1} \times h_{a-1} \times d_{a-1}}$. For jth ($j \in (0, a-1)$) sub-decoders, the input becomes $(F_t^{j+1}(x_i) + F_s^{j+1}(x_i))$ due to the existence of skip connections (Sec. II.C), and its output is $F_s^j(x_i) \in \mathbb{R}^{w_j \times h_j \times d_j}$. It should be noted that when k is equal to j, the dimension of $F_t^k(x_i)$ is equal to that of $F_s^j(x_i)$. We use the cosine similarity [25] between F_t and F_s to compute the loss value. For F_t^l and F_s^l ($l \in (0, a)$), first we obtain a 2D loss map $M^l(x_i) \in \mathbb{R}^{w_l \times h_l}$ by calculating the cosine similarity along the channel axis:

$$M^l(w,h) = 1 - \frac{(F_t^l(w,h))^T \cdot F_s^l(w,h)}{\| F_t^l(w,h) \| \cdot \| F_s^l(w,h) \|} \qquad (1)$$

and then accumulate the loss map to get the loss function L^l of this layer:

$$L^l = \frac{1}{w_l h_l} \left(\sum_{w=1}^{w_l} \sum_{h=1}^{h_l} M^l(w,h) \right) \qquad (2)$$

Finally, the loss functions of each layer are added to obtain the multi-scale loss L_{DRKD} of the DRKD phase:

$$L_{DRKD} = \sum_{l=1}^{a-1} L^l \qquad (3)$$

We train the student decoder D by minimizing L_{DRKD} and the parameters of the teacher encoder remain unchanged.

C. Skip Connections

During the DRKD process, the final layer output of the teacher encoder is directly connected to the student decoder. However, this simple way of connecting E and D creates a flaw in our model. It is well known that layers of neural networks correspond to features at various levels; the first few layers extract features such as colors, edges, and textures, while features from the latter layers contain more semantic information [3,16]. As a result, we face the same problem as

indirect RKD: D has difficulty reconstructing shallow features of an image from highly abstract features. In order to solve the problem and help the student decoder D reconstruct anomaly-free representations of images, we introduce skip connections between D and E by taking inspiration from some related works on image reconstruction [27,28,29,30]. Taking the second sub-encoder $D2$ as an example, we let the sum of the outputs from $D3$ and $E3$ be its input. During training, as there is only normal data in the dataset, F_t^3 only contains normal patterns of images, and $D2$ learns how to decode deep normal features in F_t^3 into shallow features. For abnormal images in testing, F_t^3 can be divided into two parts: normal features F_{t1}^3 and abnormal ones F_{t2}^3. $D2$ will successfully decode F_{t1}^3 based on the knowledge gained during training but cannot understand F_{t2}^3, which leads to a low similarity between F_s^2 and F_t^2. On the other hand, for normal input, the information in F_t^3 can help $D2$ reconstruct the shallow representation.

D. Anomaly Scoring

In the testing phase, we need to get the anomaly score of the input image. According to Eq. (1), we first calculate the cosine similarity between the output of each layer of the student encoder D and the teacher decoder E, and then obtain a 2D anomaly map M^l. Since the output scales are inconsistent, we need to up-sample M^l to the same size and obtain M_{re}^l by bilinear interpolation. The final anomaly map M is obtained by accumulating all the M_{re}^l. We designate the maximum value in M as the anomaly score. This is due to the inconsistent size of anomaly regions in medical images, making it inappropriate to utilize the average value as the anomaly score.

III. EXPERIMENTS AND DISCUSSIONS

A. Datasets and Competing Methods

We compare our model with four AD models to demonstrate its superiority: RDAD [18], STPM [16], CutPaste [13] and f-AnoGAN [11]. To accurately evaluate the performance of models with limited sample sizes, we selected five small medical datasets for experimentation. All
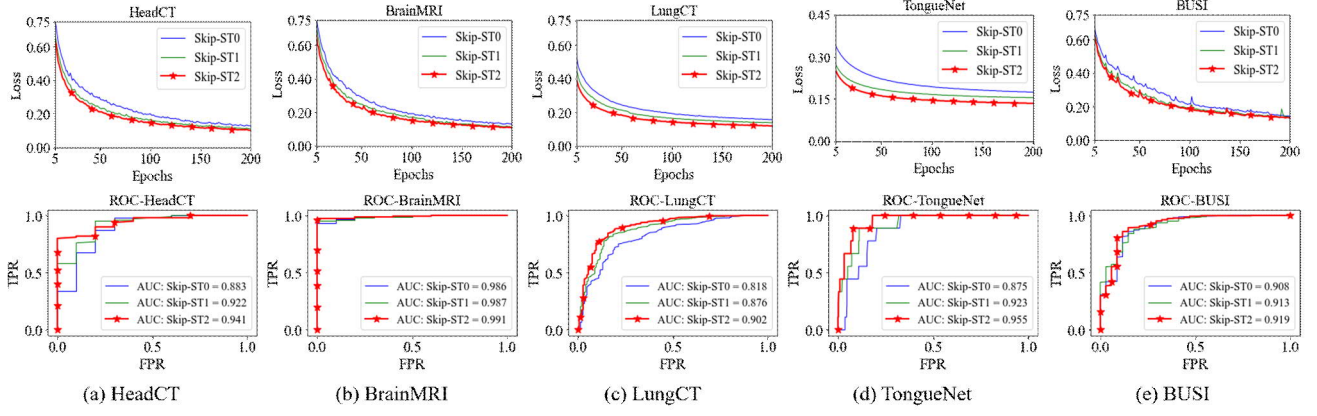
Fig. 4. Training performance curves (row 1) and ROC curves (row 2) of Skip-ST0, Skip-ST1 and Skip-ST2 on the HeadCT (a), BrainMRI (b), LungCT (c), TongueNet (d) and BUSI (e). All results gathered with seed 114514.

TABLE I. KEY STATISTICS OF IMAGE DATASETS

Dataset	Training dataset	Testing dataset	
	Normal	Normal	Abnormal
HeadCT	90	10	100
BrainMRI	88	10	154
LungCT	432	145	811
TongueNet	2597	484	9
BUSI	99	34	647

TABLE II. AUROC COMPRISON

Dataset	Method				
	f-AnoGAN	STPM	CutPaste	RDAD	Skip-ST
HeadCT	**0.907**	0.715	0.801	0.724	**0.929**
BrainMRI	0.708	0.859	0.893	**0.924**	**0.989**
LungCT	0.891	0.790	0.833	**0.893**	**0.902**
TongueNet	0.882	**0.914**	0.815	0.912	**0.944**
BUSI	0.771	0.825	0.794	**0.886**	**0.919**
Average	0.832	0.821	0.827	**0.868**	**0.937**

images in these datasets are uniformly resized to 256×256. We divide these datasets using the pattern shown in Table I.

B. Evaluation Metrics

We adopt the popular AD performance metric: Area Under the Receiver Operating Characteristic (AUROC) [31]. An AUROC value of 1 indicates optimal classification performance, while values close to 0.5 indicate random classification. In order to demonstrate the robustness of models, we test the performance of the last epoch under five different random number seeds as the final result. Note that this is a very harsh experimental setup and we do not allow any unprincipled early stops [3,32] to exist.

C. Experimental Results and Discussions.

The AUROC comparison of Skip-ST with other methods on five datasets is shown in Table II. The average outcome shows that our method exceeds SOTA (RDAD) by 7.95% and achieves the best performance. Compared with f-AnoGAN on HeadCT, STPM on TongueNet, RDAD on BrainMRI, LungCT, and BUSI, our method has an AUROC improvement of 2.43%, 3.28%, 7.03%, 1.01%, and 3.72%, respectively. It is noteworthy that the average performance of RDAD and Skip-ST exceeds the traditional KD method STPM by a significant margin, indicating the effectiveness of the RKD paradigm. Additionally, compared to RDAD, our DRKD paradigm demonstrates superiority as the student network is able to directly access knowledge extracted by the teacher network, reducing the potential for information loss.

D. Ablation Study

We validate the effectiveness of skip connections with an ablation study. Fig. 4 illustrates the training performance

curves and receiver operating characteristic (ROC) curves on five medical datasets. We conduct the experiments using Skip-ST with zero, one, and two skip connections, called Skip-ST0, Skip-ST1, and Skip-ST2, respectively. From the training performance curves (row 1 in Fig. 4), we can see that from Skip-ST0 to Skip-ST2, the loss value decreases faster, indicating that skip connections help the student decoder reconstruct normal features. For the ROC curves (row 2 in Fig. 4), the x-axis is false positive rate (FPR) and the y-axis is the true positive rate (TPR). The experimental results indicate that, on average, the AUROC of Skip-ST2 has been enhanced by 5.32% due to the implementation of skip connections compared to Skip-ST0.

IV. CONCLUSION

We proposed a student-teacher network with skip connections (Skip-ST) that is trained by direct reverse knowledge distillation (DRKD) for anomaly detection. Skip-ST has an encoder-decoder architecture, where the student encoder learns to reconstruct an anomaly-free representation of the input. Compared with other KD paradigms, our proposed method achieves higher AUROC by directly transferring the teacher's knowledge to the student and introducing skip connections. Experimental results on five medical datasets show that our method exceed RDAD by 7.95% on AUROC, outperforming the state-of-the-art AD models. In future work, we will implement anomaly localization in medical images and test our method on more medical datasets.

REFERENCES

[1] Sonal Kothari, John H Phan, Todd H Stokes, and May D Wang, "Pathology imaging informatics for quantitative analysis of whole-

slide images." *Journal of the American Medical Informatics Association* 20.6 (2013): 1099-1108.

[2] Der-Chiang Li, Chiao-Wen Liu, and Susan C. Hu, "A learning method for the class imbalance problem with medical data sets." *Computers in biology and medicine* 40.5 (2010): 509-518.

[3] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee, "Multiresolution knowledge distillation for anomaly detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2021.

[4] He Zhao *et al.*, "Anomaly detection for medical images using self-supervised and translation-consistent features." *IEEE Transactions on Medical Imaging* 40.12 (2021): 3641-3651.

[5] Gaurav Dutta. COVID-19 Detection. https://www.kaggle.com/datasets/gauravduttakiit/covid19-detection, 2022.

[6] Mingxuan Liu, Yunrui Jiao, Hongyu Gu, Jingqiao Lu, and Hong Chen, "Data Augmentation Using Image-to-image Translation for Tongue Coating Thickness Classification with Imbalanced Data." *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022.

[7] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy, "Dataset of breast ultrasound images." *Data in brief* 28 (2020): 104863.

[8] Pang Guansong *et al.*, "Deep learning for anomaly detection: A review." *ACM Computing Surveys (CSUR)* 54.2 (2021): 1-38.

[9] Ian Goodfellow *et al.*, "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144.

[10] Diederik P. Kingma, and Max Welling, "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

[11] Thomas Schlegl, Philipp Seeböck, Sebastian M.Waldstein, Georg Langs, and Ursula Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks." *Medical image analysis* 54 (2019): 30-44.

[12] Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders." *IEEE Access* 9 (2021): 118571-118583.

[13] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[14] Jouwon Song, Kyeongbo Kong, Ye-In Park, Seong-Gyun Kim, and Suk-Ju Kang, "Anomaly Segmentation Network Using Self-Supervised Learning." *AAAI 2022 Workshop on AI for Design and Manufacturing (ADAM)*. 2021.

[15] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2020.

[16] Guodong Wang, Shumin Han, Errui Ding, and Di Huang, "Student-teacher feature pyramid matching for unsupervised anomaly detection." *The British Machine Vision Conference (BMVC)*. 2021.

[17] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, "MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2019.

[18] Hanqiu Deng and Xingyu Li, "Anomaly Detection via Reverse Distillation from One-Class Embedding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022.

[19] Kees M. van Hespen, Jaco J. M. Zwanenburg, Jan W. Dankbaar, Mirjam I. Geerlings, Jeroen Hendrikse, and Hugo J. Kuijf, "An anomaly detection approach to identify chronic brain infarcts on MRI." *Scientific Reports* 11.1 (2021): 1-10.

[20] Oliver Rippel, Patrick Mertens, and Dorit Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection." *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.

[21] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks." *The British Machine Vision Conference (BMVC)*. 2016.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks." *Communications of the ACM* 60.6 (2017): 84-90.

[23] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization." *International Conference on Pattern Recognition.* Springer, Cham, 2021.

[24] Yunkang Cao, Qian Wan, Weiming Shen, and Liang Gao, "Informative knowledge distillation for image anomaly segmentation." *Knowledge-Based Systems* 248 (2022): 108846.

[25] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation." *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2019.

[26] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus, "Deconvolutional networks." *2010 IEEE Computer Society Conference on computer vision and pattern recognition.* IEEE, 2010.

[27] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection." *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

[28] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections." *Advances in neural information processing systems* 29 (2016).

[29] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao, "Image super-resolution using dense skip connections." *Proceedings of the IEEE international conference on computer vision.* 2017.

[30] Yan Zou, Linfei Zhang, Chengqian Liu, Bowen Wang, Yan Hu, and Qian Chen, "Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections." *Optics and Lasers in Engineering* 146 (2021): 106717.

[31] Tom Fawcett, "An introduction to ROC analysis." *Pattern recognition letters* 27.8 (2006): 861-874.

[32] Mohammadreza Salehi, Ainaz Eftekhar, Niousha Sadjadi, Mohammad Hossein Rohban, and Hamid R. Rabiee, "Puzzle-ae: Novelty detection in images through solving puzzles." *arXiv preprint arXiv:2008.12959* (2020).