BMAD: Benchmarks for Medical Anomaly Detection

Supplementary Material

This is the supplementary document of our paper, entitled "BMAD: Benchmarks for Medical Anomaly Detection". It includes 4 sections, providing detailed information on datasets, supporting AD algorithms, experimental reproducibility, and evaluation metrics.

6. Datasets in BMAD

Our BMAD benchmark consists of six datasets sourced from five distinct medical domains, including brain MRI, retinal OCT, liver CT, chest X-ray, and digital histopathology. Due to the absence of specific anomaly detection datasets in the field of medical imaging, we construct these benchmark datasets by reorganizing and remixing existing medical image sets proposed for other purposes such as image classification and segmentation. Moreover, our codebase includes functionality for data reorganization, enabling users to generate new datasets tailored to their needs. In the this sections, we mainly focus on an overview of the original datasets and our data reorganization procedure.

6.1. Brain MRI Anomaly Detection and Localization Benchmark

The brain MRI anomaly detection benchmark is reorganized from the BraTS2021 dataset [3, 4, 40].

6.1.1 BraTS2021 Dataset

The original BraTS2021 dataset is proposed for a multimodel brain tumor segmentation challenge. It provides 1,251 cases in the training set, 219 cases in validation set, 530 cases in testing set (nonpublic), all stored in NIFTI (.nii.gz) format. Each sample includes 3D volumes in four modalities: native (T1) and post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR), accompanied by a 3D brain tumor segmentation annotation. The data size for each modality is 240 *240 *155.

Access and License: The BraTS2021 dataset can be accessed at http://braintumorsegmentation.org/. Registration for the challenge is required. As stated on the challenge webpage, "Challenge data may be used for all purposes, provided that the challenge is appropriately referenced using the citations given at the bottom of this page."

6.1.2 Construction of Brain MRI AD benchmark

After analyzing the BraTS2021 dataset, we built the brain MRI AD benchmark from the 3D FLAIR volumes. All data in our Brain MRI AD benchmark is derived from the 1,251 cases in the original training set. To account for variations in brain images at different depths, we specifically selected slices within the depth range of 60 to 100. Each extracted 2D slice was saved in PNG format and has an image size of 240 * 240 pixels. According to the tumer segmentation mask, we selected 7,500 normal samples to compose the AD training set, 3,715 samples containing both normal and anomaly samples (with a ratio of 1:1) for the test set, and a validation set with 83 samples that do not overlap with the test set. Fig. 4 illustrates the specific procedure we followed for data preparation, and Fig. 5 provides examples of our brain MRI AD benchmark.

6.2. Liver CT Anomaly Detection and Localization Benchmark

We structure this benchmark from two distinct datasets, BTCV [30] and LiTS [10]. The anomaly-free BTCV set is taken to constitute the normal train set in this benchmark and CT scans in LiTs is exploited to form the evaluation and test data.

6.2.1 BTCV Dataset

BTCV [30] is introduced for multi-organ segmentation. It consists of 50 abdominal computed tomography (CT) scans taken from patients diagnosed with colorectal cancer and a retrospective ventral hernia. The original scans were acquired during the portal venous contrast phase and had variable volume sizes ranging from 512*512*85 to 512*512*198 and stored in nii.gz format.

Access and License: The original BTCV dataset can be accessed from 'RawData.zip' at https://www.synapse.org/#!Synapse:syn3193805/wiki/217753. Dataset posted on Synapse is subject to the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

6.2.2 LiTS Dataset

LiTS [10] is proposed for liver tumor segmentation. It originally comprises 131 abdominal CT scans, accompanied by a ground truth label for the liver and liver tumors. The original LiTS is stored in the nii.gz format with a volume size of 512*512*432.

Access and License: LiTS can be downloaded from its Kaggle webpage at https://www.kaggle.com/datasets/andrewmvd/liver-tumor-segmentation. The use of the LiTS dataset is under Creative Commons Attribution-NonCommercial-ShareAlike(CC BY-NC-SA) [11].

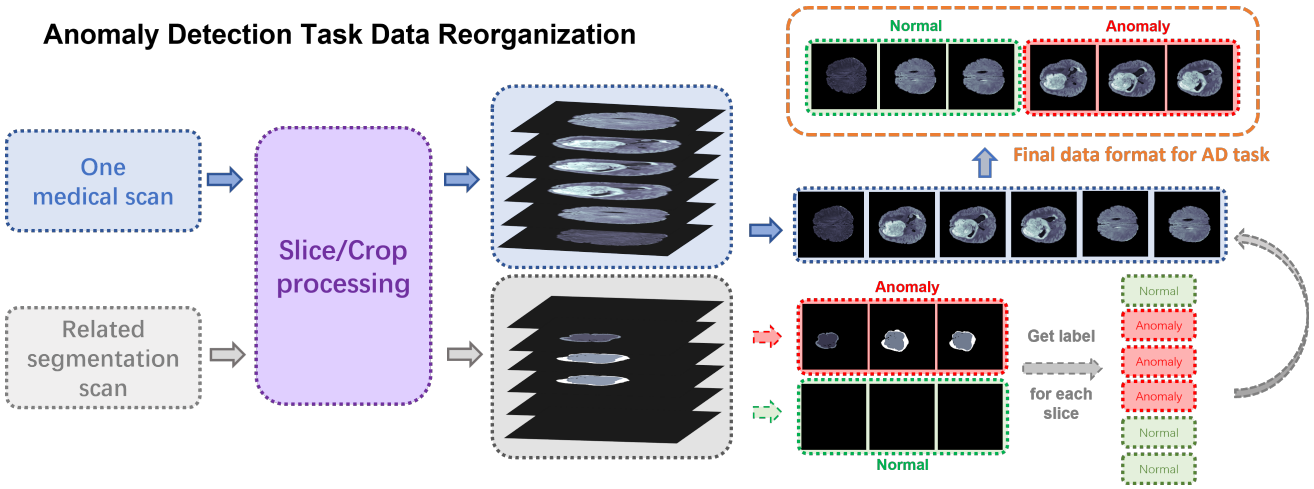

Figure 4. Diagram illustration of data preparation for the Brain MRI AD benchmark from 3D brain scans in BraTS2021.

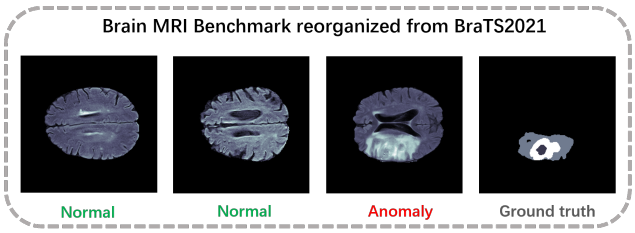

Figure 5. Visualization of our proposed Brain MRI benchmark.

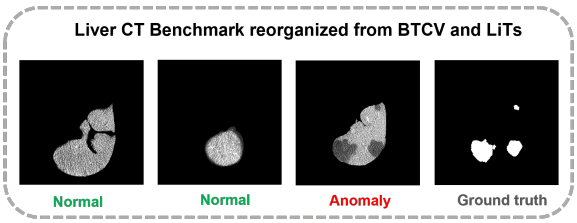

Figure 6. Visualization of our proposed Liver CT benchmark.

6.2.3 Construction of Liver CT AD Benchmark

In constructing the liver CT AD benchmark, we made a decision not to include lesion-free regions from the LiTS dataset as part of the training set. This choice was based on our observation that the presence of liver lesions in LiTS leads to morphological changes in non-lesion regions, which could impact the performance of anomaly detection. Instead, we opted to use the lesion-free liver portion from the BTCV dataset to form the training set. The LiTS dataset, on the other hand, is reserved for testing the effectiveness of anomaly detection and localization.

For both datasets, Hounsfield-Unit (HU) of the 3D scans are transformed into grayscale with an abdominal window. The scans are then cropped into 2D axial slices, and the liver's Region of Interest is extracted based on the provided organ annotations. We perform slide intensity normalization with histogram equalization. To be more specific, for the construction of the normal training set in the liver CT AD benchmark, we utilized the provided segmentation labels in BTCV to extract the liver region. From these scans, we extracted 2D slices of the liver with a size of 512 * 512, using the corresponding liver segmentation scans as a guide. The 2D slices were then converted to PNG format to serve as the final AD data. We selected 1542 slices to comprise the training set. To prepare the testing and validation sets, we sliced the data from LiTS and stored them in PNG format with dimensions of 512 * 512. Our testing and validation sets contain both healthy and abnormal samples. Fig. 6 demonstrates several samples in the Liver CT AD dataset. Fig. 6 provides visualization of the constructed Liver CT AD dataset.

6.3. Retinal OCT Anomaly Detection and Localization Benchmark

The BMAD datasets includes two different OCT anomaly detection datasets. The first one is derived from the RESC dataset [26] and support anomaly localization evaluation. The second is constructed from OCT2017 [29], Which only support sample-level anomaly detection.

6.3.1 RESC dataset

RESC (Retinal Edema Segmentation Challenge) dataset [26] specifically focuses on the detection and segmentation of retinal edema anomalies. It provides pixel-level segmentation labels, which indicate the regions affected by retinal edema. The RESC is provided in PNG format with a size of 512*1024 pixels.

Access and License: The original RESC dataset

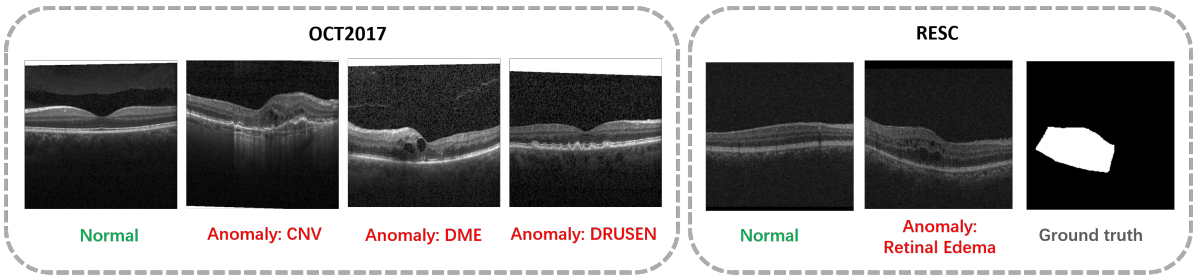

Figure 7. The Retinal OCT benchmarks consist of two separate datasets, each representing different anomaly types. These datasets are used to evaluate and benchmark various methods in the field of retinal OCT imaging. The datasets are designed to assess the performance of algorithms in detecting and localizing specific anomalies in retinal images.

can be downloaded from the P-Net github page at https://github.com/CharlesKangZhou/P_Net_Anomaly_Detection. As indicated on the webpage, the dataset can be only used for the research community.

6.3.2 OCT2017 dataset

OCT2017 [29] is a large-scale dataset initially designed for classification tasks. It consists of retinal OCT images categorized into three types of anomalies: Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen Deposits (DRUSEN). The images are continuous slices with a size of 512*496.

Access and License: OCT2017 can be downloaded at https://data.mendeley.com/datasets/rscbjbr9sj/2. Its usage is under a license of Creative Commons Attribution 4.0 International(CC BY 4.0).

6.3.3 Preparation of OCT AD benchmarks

To construct the OCT anomaly detection and localization dataset from RESC, we utilize the segmentation labels provided for each slice to get the label for AD setting. We select the normal samples from the original training dataset and adapt the original validation set into the AD setting for evaluation. The RESC is provided in PNG format with a size of 512*1024 pixels. On the other hand, on the OCT2017 dataset, we specifically select the disease-free samples from the original training set as our training data for the anomaly detection task. The test set is further divided into evaluation data and testing data for AD setting. Fig. 7 demonstrates several examples in the two OCT AD datasets.

6.4. Chest X-ray Anomaly Detection Benchmark

6.4.1 RSNA dataset

RSNA [63], short for RSNA Pneumonia Detection Challenge, is originally provided for a lung pneumonia detection task. The 26,684 lung images are associated with three labels: "Normal" indicates a normal lung condition, "Lung Opacity" indicates the presence of pneumonia, "No Lung Opacity/Not Normal" represents a third category where some images are determined to not have pneumonia, but there may still be some other type of abnormality present in the image. All images in RSNA are in DICOM format.

Access and License: RSNA can be accessed by https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/overview. Stated in the section of Competition data: A. Data Access and Usage, "... you may access and use the Competition Data for the purposes of the Competition, participation on Kaggle Website forums, academic research and education, and other non-commercial purposes."

6.4.2 Preparation of Chest X-ray AD Benchmark

We utilized the provided image labels for data re-partition. Specifically, "Lung Opacity" and "No Lung Opacity/Not Normal" were classified as abnormal data. The reorganized AD dataset including 8000 normal images as training data, 1490 images with 1:1 normal-versus-abnormal ratio in the validate set, and 17194 images in the test set. Examples of the chest X-ray dataset are provided in Fig. 8.

6.5. Digital Histopathology Anomaly Detection Benchmark

6.5.1 Camelyon16 Dataset

The Camelyon16 dataset [6] was initially utilized in the Camelyon16 Grand Challenge to detect and classify metastatic breast cancer in lymph node tissue. It comprises 400 whole-slide images (WSIs) of lymph node sections stained with hematoxylin and eosin (H&E) from breast cancer patients. Among these WSIs, 159 of them exhibit tumor metastases, which have been annotated by pathologists. The WSIs are stored in standard TIFF files, which include

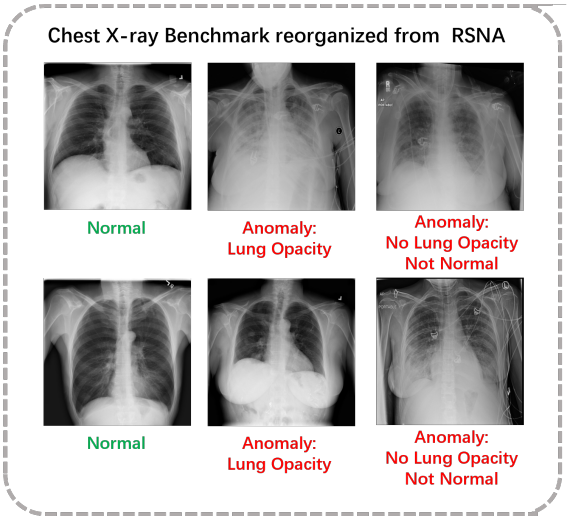

Figure 8. Our proposed chest X-ray benchmark consists two types of anomalies. These anomalies are clearly labeled in the images, and all of them are considered as anomaly samples.

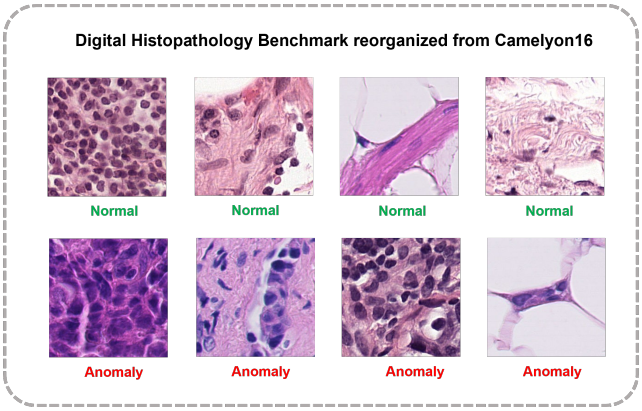

Figure 9. Examples of the digital histopathology AD benchmark. Unlike other medical image AD benchmarks, histopathology images shows higher diversities in tissue components.

multiple down-sampled versions of the original image. In Camelyon16, the highest resolution available is on level 0, corresponding to a magnification of 40X.

Access and Licence: The original Camelyon16 dataset can be found at https://camelyon17.grand-challenge.org/Data/. It is under a license of Creative Commons Zero 1.0 Universal Public Domain Dedication(CC0).

6.5.2 Preparation of histopathology AD Benchmark

To ensure a comprehensive evaluation of anomaly detection models for histopathology images, considering their unique characteristics such as large size, we opted to assess AD models at the patch level. To construct the benchmark dataset, we randomly extracted 5,088 normal patches from the original training set of Camelyon16, which consisted of 160 normal WSIs. These patches were utilized as training samples. For the validation set, we cropped 100 normal and 100 abnormal patches from the 13 testing WSIs. Likewise, for the testing set, we extracted 1,000 normal and 1,000 abnormal patches from the 115 testing WSIs in the original Camelyon16 dataset. Each cropped patch was saved as a PNG image with dimensions of 256 * 256 pixels. Fig. 9 presents several examples in the constructed histopathology AD benchmark.

7. Supported AD Models

Fig. 10 provides conceptual illustration of various AD architectures from the feature embedding-based methods and data reconstruction-based approaches. We conducted benchmarking using the Anomalib [2] for CFA, CFlow, DRAEM, GANomaly, PADIM, PatchCore, RD4AD, and STFPM. For the remaining algorithms, we provided a comprehensive codebase for training and inference with all proposed evaluation metrics functions. By utilizing these codebases and following the instructions provided, researchers can replicate and reproduce our experiments effectively. In addition to the codebase, we also provide pre-trained checkpoints for different benchmark on our webpage.

The specific experimental settings for each of the supported methods are specified as follows.

PaDiM [17] leverages a pre-trained convolutional neural network (CNN) for its operations and does not require additional training. In our experiments, we separately evaluated all benchmarks using two backbone networks: ResNet-18 and WideResnet-50. For the dimension reduction step, we retained the default number of features as specified in the original setting. Specifically, we used 100 features for ResNet-18 and 550 features for WideResnet-50. These default values were chosen based on the original implementation and can serve as a starting point for further experimentation and fine-tuning if desired.

STFPM [67] utilized feature extraction from a Teacher-student structure. In our experiments, we evaluated all benchmarks separately using two backbone networks: ResNet-18 and WideResnet-50. We employed a SGD optimizer with a learning rate of 0.4. Additionally, we followed the original setting with a parameter with a momentum of of 0.9 and weight decay of 1e-4 for SGD. These settings were chosen based on the original implementation and can be adjusted for further experimentation if desired.

Patchcore [46] is a memory-based method that utilizes coreset sampling and neighbor selection. In our experiments, we evaluated Patchcore using two backbone networks: ResNet-18 and WideResnet-50. We followed the default hyperparameters of 0.1 for the coreset sampling ratio and 9 for the chosen neighbor number. These values were chosen based

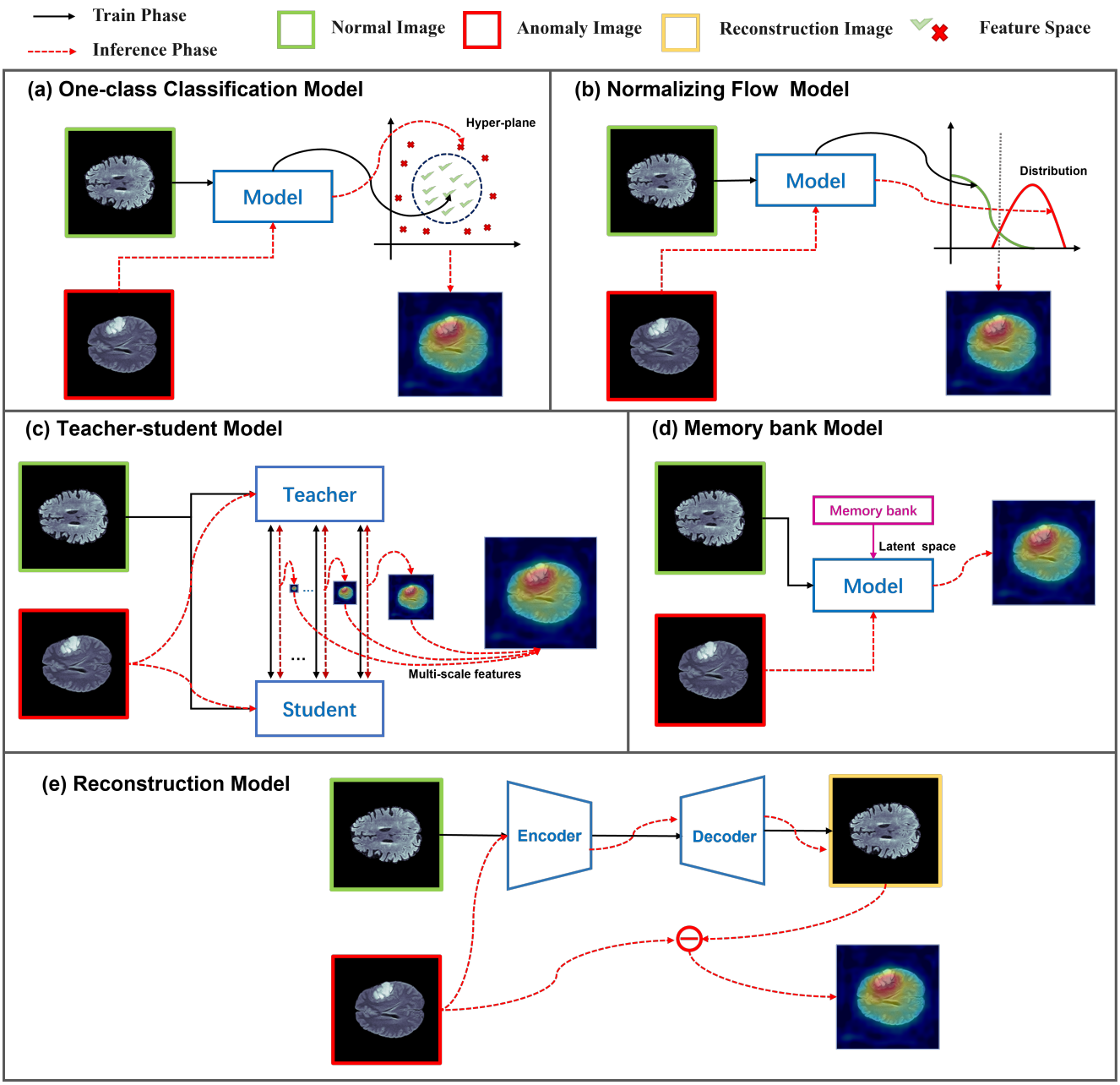

Figure 10. Conceptual illustration of various AD models. The one-class classification model, normalizing flow model, teaching-student model and memory bank model detects anomalies in the embedding space, and the reconstruction based method takes a generative model as its backbone for pixel-level anomaly comparison between the original query and reconstruction.

on the original implementation.

RD4AD [19] utilizes a wide ResNet-50 as the backbone network and applies the Adam optimizer with a learning rate of 0.005. In addition, we follow the defeat set of the beta1 and beta2 parameters to 0.5 and 0.99, respectively. For the anomaly score of each inference sample, the maximum value of the anomaly map is used. These settings were determined based on the original implementation of RD4AD and can be adjusted if needed.

DRAEM [73] is a anomaly augmentation reconstruction-based method utilized U-Net structure. The learning rate used for two sub network training is 1e-4, and the Adam optimizer is employed. For the remaining settings, we follow the default configurations specified in the original work.

CFLOW [22]is a normalizing flows-based method. We utilized WideResnet-50 as backbone and Adam optimizer

with a learning rate of 1e-4 for all benchmarks' experiments. And we follow the original parameter settings, including the selection of 128 for the number of condition vectors and 1.9 as clamp alpha value.

CFA [31] is also a memory bank-based algorithm. We employs a WideResnet-50 backbone and follows the parameter settings outlined in the original paper. The method utilizes 3 nearest neighbors and 3 hard negative features. A radius of 1e-5 is utilized for searching the soft boundary within the hypersphere. The model is trained using the Adam optimizer with a learning rate of 1e-3 and a weight decay of 5e-4. These specific parameter configurations play a crucial role in achieving the desired performance and effectiveness of the CFA approach, as determined by the original research paper or implementation.

MKD [53] utilizes the VGG16 backbone for feature extraction, and only the parameters of the cloner are trained. We follow the defeat setting with a batch size of 64. The learning rate is set to 1e-3 using the Adam optimizer. Additionally, the λ value is set to 1e-2, which represents the initial amount of error assigned to each term on the untrained network. These parameter settings are have been chosen based on the original research paper.

UTRAD [15] is based on Transformer backbone with a ReLu activation function. We trained the model with a defeat parameters setting: batch size of 8 and an Adam optimizer with a learning rate of 1e-4. The parameter settings are have been chosen based on the original research paper.

CutPaste [33] utilizes a Resnet-18 backbone. The backbone is frozen for the first 20 epochs of training. We trained the model using an SGD optimizer with a learning rate of 0.03. And the batch size for training is following to the defeat parameter, set to 64.

GANomaly [1] is trained using an Adam optimizer with a learning rate of 2e-4. The $\beta 1$ and $\beta 2$ parameters of the Adam optimizer are set to 0.5 and 0.999, respectively, following the original work. The weights assigned to different loss components are also set according to the original setting: a weight of 1 for the adversarial loss, a weight of 50 for the image regeneration loss, and a weight of 1 for the latent vector encoder loss. These parameter values have been chosen based on the original research paper and are crucial for the performance and effectiveness.

DeepSVDD [50] utilizes a LeNet as its backbone and is trained using an Adam optimizer with a learning rate of 1e-4. The model training follows the setting of weight decay as 0.5e-7 and a batch size of 200. These parameter values have been chosen based on the original research paper or implementation.

f-AnoGAN [55] is a generative network that requires two-stage training. During the training process, we use an Adam optimizer with a batch size of 32 and a learning rate of 2e-4. Additionally, the dimensionality of the latent space is set to 100. These parameter settings have been chosen based on the original research paper.

Benchmarks	BraTS2021	BTCV + LiTs	RESC
DRAEM [73]	19.31 ± 5.52	9.38 ± 0.78	33.51 ± 3.52
UTRAD [15]	7.27 ± 0.06	2.33 ± 0.06	22.81 ± 0.36
MKD [53]	28.89 ± 0.72	<u>14.92 ± 0.23</u>	43.53 ± 1.10
RD4AD [19]	28.28 ± 0.48	10.72 ± 2.50	33.51 ± 3.52
STFPM [67]	25.40 ± 0.82	8.87 ± 2.52	49.23 ± 0.23
PaDiM [17]	25.84 ± 1.20	4.50 ± 0.46	38.30 ± 0.89
PatchCore [46]	<u>32.82 ± 0.59</u>	10.49 ± 0.23	<u>57.04 ± 0.21</u>
CFA [31]	30.22 ± 0.32	<u>14.93 ± 0.08</u>	36.57 ± 0.18
CFLOW [22]	19.50 ± 2.73	7.58 ± 3.16	44.83 ± 1.78
SimpleNet [39]	28.96 ± 1.73	12.26 ± 2.41	30.28 ± 1.64

Table 3. Anomaly detection performance quantified by DICE over BMAD. The top method for each metric are underlined. Note that Dice is a threshold-dependent metric. The results in the table is obtained with threshold of 0.5. By adjusting the threshold for each result, it is possible to achieve higher performance.

CS-Flow [48] is trained using specific hyper-parameter settings. During the flow process, a clamping parameter of 3 is utilized to restrict the values. Gradients are clamped to a value of 1 during training. The network is trained with an initial learning rate of 2e-4 using the Adam optimizer, and a weight decay of 1e-5 is applied. These hyper-parameter settings have been determined through a process of optimization and are considered optimal for the CS-Flow method.

SimpleNet [39] was trained using the original hyper-parameters and includes two main modules. We retained the original parameters for the adapter and the Gaussian noise generation module. The results are based on the best performance achieved on the validation set during the top 40 training epochs, following the original settings.

8. Evaluation Metrics

8.1. AUROC

AUROC refers to the area under the ROC curve. It provides a quantitative value showing a trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across different decision thresholds.

$$AUROC = \int_0^1 (TPR)\mathrm{d}(FPR) \tag{1}$$

- To calculate the pixel-level AUROC, different thresholds are applied to the anomaly map. If a pixel has an anomaly score greater than the threshold, the pixel is anomalous. Over an entire image, the corresponding TPR and FPR pairs are recorded for a ROC curve and the area under the curve is calculated as the final metric.
- To calculate the image-level AUROC, each model independently calculates an anomaly score from the anomaly map as

a sample-level evaluation metric. Then different thresholds are applied to determine if the sample is normal or abnormal. Then the corresponding TPR and FPR pairs are recorded for estimating the ROC curve and sample-level AUROC value.

8.2. Per-Region Overlap (PRO)

We utilized PRO, a region-level metric,to assess the performance of fine-grained anomaly detection. To compute PRO, the ground truth is decomposed into individual unconnected components. Let A denote the set of pixels predicted to be anomalous. For connected components k, C_k represents the set of pixels identified as anomalous. PRO can then be calculated as follows,

$$PRO = \frac{1}{N} \sum_{k} \frac{|A \cap C_k|}{|C_k|}, \tag{2}$$

where N represents the total number of ground truth components in the test dataset.

8.3. DICE score

The Dice score is an important metric in medical image segmentation, evaluating the similarity between segmented results and reference standards. It measures the pixel-level overlap between predicted and reference regions, ranging from 0 (no agreement) to 1 (perfect agreement). Higher Dice scores indicate better segmentation consistency and accuracy, making it a commonly used metric in medical imaging for comparing segmentation algorithms. It should be noted that the Dice score is a threshold dependent metric. It requires different threshold values for different models and datasets to better suit the specific task. Therefore, we opted to not include the DICE comparison in the main experimentation. [Remark:] Due to the significance of DICE in medical segmentation, our codebase also includes a Dice function for its potential usage. For reference, Table 3 provides the Dice scores for the suppoeted AD methods with the threshold 0.5. By adjusting the threshold for each result, it is possible to achieve higher performance.