Generalized Anomaly Detection in Medical Imaging with Vision-Language Models

by

Jinan Bao

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Signal and Image Processing

Department of Electrical and Computer Engineering
University of Alberta

Abstract

In the realm of image processing, machine learning models have achieved remarkable progress in tasks such as classification, recognition, and video analysis. However, their reliance on closed-set assumptions limits their performance in real-world scenarios where unseen anomalies frequently occur. This limitation is particularly critical in medical applications, where undetected anomalies can have severe consequences. Addressing this open-set problem through anomaly detection is imperative for developing robust systems capable of adapting to unpredictable real-world scenarios.

In this context, medical anomaly detection stands as a pivotal challenge due to the inherent unpredictability of pathological conditions. Despite recent advances, existing benchmarks for anomaly detection primarily focus on industrial and natural images, neglecting the specific requirements of medical domains. This has led to inconsistencies in data utilization and a lack of standardized evaluation protocols, impeding fair comparisons among methods. To bridge this gap, we introduce BMAD, a unified benchmark tailored specifically for assessing anomaly detection methods on medical images. BMAD comprises six reorganized datasets spanning five medical domains, alongside standardized evaluation metrics and a comprehensive codebase supporting 15 state-of-the-art algorithms.

The analysis on BMAD reveals that no single model achieves universal effectiveness across multiple medical domains, emphasizing the need for more generalized approaches. To this end, we propose a novel multimodal framework leveraging the Contrastive Language-Image Pre-training (CLIP) model to identify anomalies within medical data. By employing language models to describe and reconstruct image infor-

mation, our approach achieves state-of-the-art performance across multiple domains, demonstrating improved generalization capabilities.

Preface

This thesis was supported by the Natural Sciences and Engineering Research Council of Canada and Alberta Innovates.

Chapter 3 of this thesis has been successfully published under the title "BMAD: Benchmarks for Medical Anomaly Detection" by Jinan Bao, Hanshi Sun, Hanqiu Deng, Zhaoxiang Zhang, and Xingyu Li, in the prestigious CVPR Workshop series in 2024. This groundbreaking effort stems from the collaborative contributions of the research group, with Jinan Bao serving as the lead author.

Acknowledgements

I would like to express my deepest gratitude to all those who have contributed, in various ways, to the completion of this thesis. This work would not have been possible without the generous support and invaluable guidance of several individuals.

First and foremost, I am immensely grateful to my supervisor, Professor Xingyu Li, for her patience, unwavering encouragement, and insightful feedback throughout this research journey. Professor Li's expert guidance and rigorous academic approach have been instrumental in shaping this thesis and fostering my critical thinking and research skills. Her dedication to excellence has inspired me to strive for the best in my work.

I am also deeply indebted to the members of my committee, Dr.Mrinal Mandal, Dr.Merak Reformat, and Dr.Xingyu Li, for their valuable time, constructive criticisms, and thoughtful suggestions during the various stages of this thesis development. Their diverse perspectives and expertise have enriched my understanding of the subject matter and significantly improved the quality of this work.

I am sincerely thankful to the staff of the Electrical and Computer Engineering, for their administrative support and assistance, which facilitated a smooth research process. My heartfelt appreciation goes out to my fellow students and research colleagues, who have been a constant source of intellectual stimulation, encouragement, and camaraderie. The countless discussions, debates, and brainstorming sessions with you have been invaluable in refining my ideas and perspectives.

Furthermore, I am profoundly grateful to my family and friends for their unwavering love, support, and understanding. Your encouragement during the challenging

moments and celebration of the small victories have been the cornerstone of my journey. Your belief in me has given me the strength and motivation to pursue this endeavor.

In conclusion, this thesis stands as a testament to the collective efforts and contributions of many individuals. I am deeply grateful to each and every one of you for being a part of this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Thesis Motivation and Scope

With the proliferation of machine learning in image processing, numerous models have emerged, proficiently tackling diverse tasks like classification, recognition, and video analysis. While these methods excel in their respective niches, they often suffer from a key limitation: their performance is inherently tied to the specific datasets they've been trained on. This limitation, known as the closed-set problem, becomes particularly pronounced in the realm of detection task. In the canonical setting, classifiers are designed and trained to identify instances belonging to known categories, but in the real world, systems may be confronted with inputs that do not belong to any of the predefined classes they were trained on. These unknown or unseen classes pose a significant challenge, as misclassifying them as one of the known classes can lead to catastrophic consequences, especially in safety-critical applications such as autonomous driving, medical diagnosis, and surveillance systems. Hence, the quest for effective strategies to address the open-set problem is necessary, among which unsupervised anomaly detection relying on normal samples during model training has potential to advance the field.

When delving into the applications of open-set concepts, unsupervised anomaly detection, which targets to identify the unseen abnormalities from the majority normal population, emerges as a particularly valuable tool in the medical domain. This

heightened practicality stems from the fact that abnormal pathological conditions are inherently unpredictable and difficult to simulate. Consequently, anomaly detection systems offer a means to identify these unforeseen conditions, leading to cost savings and enhanced healthcare. By flagging deviations from what is considered normal, these systems facilitate early interventions while minimizing the need for exhaustive diagnostic testing. As such, anomaly detection plays a pivotal role in medicine, ensuring more effective diagnoses and ultimately contributing to improved healthcare outcomes.

Given the paramount importance of anomaly detection, recent endeavors have led to the establishment of several benchmarks [1–4]. Nevertheless, these benchmarks predominantly concentrate on industrial and natural images, overlooking the critical need for medical-specific datasets. Due to the lack of dedicated medical anomaly detection benchmark, existing works on this topic usually utilize datasets originally intended for supervised classification [5–8] or segmentation tasks [9, 10] for constructing experimental datasets. In literature, we observed inconsistencies in data citation [11–13].Further more, since substantial data curation is required for this purpose, the lack of standardized protocols for reorganizing datasets suitable for anomaly detection and localization [14–16] impeds fair comparisons among methods.

To address these shortcomings, we introduce BMAD[1], a unified and exhaustive evaluation benchmark tailored specifically for assessing anomaly detection methods on medical images. BMAD comprises six meticulously reorganized datasets spanning five medical domains (brain MRI, liver CT, retinal OCT, chest X-ray, and digital histopathology) alongside three pivotal evaluation metrics. Furthermore, it incorporates fifteen state-of-the-art (SOTA) anomaly detection algorithms, providing a standardized and well-maintained platform for comprehensive comparisons. Through rigorous evaluations of these algorithms on BMAD, we offer insightful discussions on

[1]A CC BY-NC-SA license is granted to BMAD, ensuring compliance with all original dataset licenses.

the results, outlining promising research avenues for the future.

In our endeavor to address the scarcity of comprehensive benchmarks, we uncovered a salient observation: the absence of a single, universally effective, and generalizable model that can proficiently address anomaly detection tasks spanning multiple medical domains. To address this limitation, we introduce a novel approach which takes a multimodal framework, the Contrastive Language-Image Pre-training (CLIP) model, as a cornerstone for identifying anomalies within medical data. This approach employs the pretrained language models to describe image information, setting it apart from previous endeavors in medical anomaly detection that only replies on visual data in data analysis. Our experimental results suggest that our proposed language-image model-based method enables a more nuanced distinction between anomalous and normal medical images. This enhancement also facilitates a more precise localization of anomalies. Importantly, our approach exhibits remarkable effectiveness across a wider array of medical domains, underscoring its improved generalization capabilities on multiple domains, thereby contributing significantly to the academic discourse in medical anomaly detection.

Our contributions of this thesis can be summarized as follows:

- We have developed a comprehensive and standardized benchmark that includes six datasets from five common medical domains. To ensure consistency and comparability, we have made significant efforts to reorganize and adapt the datasets to the unsupervised anomaly detection setting in computational medical imaging.

- We have created a well-structured and user-friendly codebase that supports 15 state-of-the-art anomaly detection algorithms and their evaluations.

- We have conducted a thorough analysis of the strengths and weaknesses of the algorithms on the BMAD datasets. Our findings and discussions will inspire researchers to develop more advanced anomaly detection models for medical

data.

- We have devised, a novel, high-performance model that exhibits the capability to encompass multiple medical imaging domains for anomaly detection.

1.2 Thesis Structure

This thesis includes the following components. Chapter 1 provides a detailed introduction of the practical value of anomaly detection and highlights the absence of comprehensive benchmarks for medical anomaly detection. Additionally, we introduce our proposed benchmark effort and a more comprehensive model for medical anomaly detection. Chapter 2 gives a detailed background that have thoroughly introduced the background and current state-of-the-art advancements in anomaly detection methods. Chapter 3 exhaustively delineates the proposed medical anomaly detection benchmark, BMAD, which stands as a significant contribution in this thesis. The establishment of BMAD is grounded in meticulous medical annotations, thereby bridging the gap in the anomaly detection landscape by offering multi-domain medical data as a robust foundation. BMAD also provides open-source code for SOTA algorithm implementation and performance evaluation, fostering accessibility and reproducibility for researchers in the field. Chapter 4 meticulously outlines our novel multimodal medical anomaly detection framework that addresses the gap by offering a highly generalized detection model, thereby enhancing its applicability and efficacy across diverse medical domains. Finally, Chapter 5 discusses the limitations and shortcomings of our study. Potential directions for future work is also presented.

Chapter 2

Background

In this chapter, we present a comprehensive review on related works on anomaly detection that serves as the cornerstone for the subsequent discussions within this thesis. Specifically, we start from the traditional anomaly detection techniques proposed before deep learning. Then deep learning anomaly detection works are presented. In the last section of this chapter, we delve into the our target application scenario, medical anomaly detection. We summarize the prior arts and their pros and cons are discussed.

2.1 Traditional Anomaly Detection Works

Anomaly detection aims in identifying deviations from normal operational behavior, which can indicate potential damage or faults. These techniques analyze data to pinpoint unusual patterns that may signify the onset of damage, allowing for timely maintenance and intervention. The image anomaly detection based on traditional statistics and machine learning algorithms can often be categorized into the major groups presented in Table2.1.

2.1.1 Statistical Modeling

Gaussian model based methods assume data generated from a Gaussian distribution or a mixture of Gaussions. This category of methods estimates parameters like

mean and covariance using maximal likelihood estimation (MLE) from the training data. With the inferred distribution, they apply a threshold to the probabilities as anomaly scores to identify outliers.

Regression model based solutions are widely used for time-series anomaly detection [17, 18]. Fits a regression model to data, then computes anomaly scores based on residual errors for test instances.

Histogram based approaches follow a simple non-parametric approach that uses histograms to profile normal data. This category of methods is also known as frequency-based approaches and often favored in intrusion and fraud detection for capturing behavioral patterns efficiently.

Nearest neighbor strategy is the cornerstone of the methodology revolves around assessing the anomaly score of a specific data point by quantifying its distance or dissimilarity from its neighbor. If the computed distance separating a data point from its nearest neighbor surpasses those typically observed among other data points within the dataset by a substantial margin, it serves as a strong indication that the query data point exhibits anomalous or outlier behavior.

2.1.2 Bayesian Networks (BN)

Bayesian networks have been successfully employed for anomaly detection in multi-class scenarios. For univariate categorical datasets, a fundamental approach leveraging naïve Bayesian networks involves estimating the posterior probability of observing a specific class label c (drawn from a predefined set of normal class labels along with an anomaly class label), conditional on a given test data instance x. This can be formalized as estimating $P(c \mid x)$. The predicted class for this test instance is then determined as the one yielding the highest posterior probability, i.e., the class $c^* = \arg\max_c P(c \mid \mathbf{x})$. This approach allows for the detection of anomalies by comparing the posterior probability of the anomaly class label against those of the normal class labels, thereby enabling the identification of instances that are unlikely to belong

to any of the normal classes.

2.1.3 Support Vector Machines (SVM)

SVMs have been applied to one-class anomaly detection, using one-class learning techniques to define a region containing training data instances. Kernels, like radial basis function (RBF), can learn complex regions after projecting data into a high-dimensional space. The basic principle to classify test instances as normal is whether they fall within the learned region, and anomalous otherwise. Variants have been proposed for anomaly detection in audio signals, novelty detection in power plants, and system call intrusion detection. The technique has also been extended to detect anomalies in temporal sequences, with one variant finding the smallest hypersphere containing all training instances and classifying test instances outside it as anomalous. Robust Support Vector Machines (RSVMs), resilient to anomalies in training data, have been applied to system call intrusion detection. The fundamental formula for the Support Vector Machine is given by:

$$\min_{\mathbf{w},b,\xi} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right) \tag{2.1}$$

$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \tag{2.2}$$

2.2 Deep Learning Anomaly Detection Works

With the powerful capability to extract abstract numerical representations of data, deep learning is incorporated to better discriminating normal and abnormal samples for anomaly detection. Depending on how a deep neural network used for feature extraction, the existing algorithms for unsupervised anomaly detection, can be categorized into two paradigms: data reconstruction-based approaches and feature embedding-based (or projection-based) approaches. The former typically compares the differences between the reconstructed data and the original data in the data

space to identify potential anomalies, while the latter infers anomalies by analyzing the abstract representations in the embedding space.

2.2.1 Reconstruction-based Methods

A reconstruction-based approach usually deploys a generative model for data reconstruction. It targets for small reconstruction residues for normal data, but large errors for anomalies. The distinction in reconstruction errors forms the basis for anomaly detection. **AutoEncoder (AE)** and **Variational AE** have been the first and most popular models for this purpose[27–36]. Later, **Generative Adversarial Networks (GANs)** are used to replace AE for its high-quality output[5, 30, 37–40]. Recently, there is a trend of exploiting **diffusion models** for normal sample generation[16, 41, 42]. In addition to convolutional neural networks, the transformer architecture is also explore in latest studies to build these generative models[13, 43–45]. To improve AD performance, regularization strategies are incorporated into normal sample reconstruction. Following the idea of denoising AE, Gaussian noise is added into normal samples for a better normal data restoration performance[33, 46, 47]. In the masking mechanism, a normal sample is randomly masked and then inpainted back[40, 48, 49]. Furthermore, many studies focus on synthesizing abnormalities on normal training samples and use the generative model to restore the original normal version[50–52]. Recently, the memory mechanism is exploited to further constrain model's capability on reconstructing abnormal samples[32, 34–36].

The following are some representative reconstruction-based algorithms for unsupervised anomaly detection.

f-AnoGAN[39] is to train a GAN model on a dataset of normal images only. The GAN consists of two main components: a generator and a discriminator. The generator's job is to create synthetic images that are as realistic as possible, while the discriminator's role is to distinguish between real images from the training set and fake images generated by the generator. As the training progresses, the two

components compete against each other in a zero-sum game, ultimately resulting in a generator that can produce highly realistic images that fool the discriminator.

GANomaly[38] is a semi-supervised anomaly detection approach that leverages the power of GANs and autoencoder principles to learn the distribution of normal samples and identify anomalies during the testing phase. This method addresses the challenge of scarcity or absence of labeled anomaly samples during the training phase, making it suitable for real-world scenarios where anomalies are rare or unknown.

UTRAD[13] is a framework for anomaly detection and localization that leverages the U-shaped Transformer architecture. It demonstrates significant performance advantages in industrial defect detection, medical disease diagnosis, and other domains.

DRAEM[50] consists of two main components: an encoder network and a decoder network. The encoder network maps the input surface image into a latent embedding space, while the decoder network aims to reconstruct the input image from this embedding. Critically, the training of Draem is discriminative in nature, meaning that it is optimized not only to minimize the reconstruction error for normal samples but also to maximize the reconstruction error for anomalous samples.

2.2.2 Projection-based Methods

A projection-based method employs either a task-specific model or simply a pre-trained network to map data into abstract representations in an embedding space, enhancing the distinguishability between normal samples and anomalies.

One-class classification based methods

One-class classification commonly employs normal support vectors or samples to delineate a tightly enclosed one-class distribution. Specifically, the one-class SVM approach, as introduced in [53], endeavors to identify a kernel function that projects the training data onto a hyperplane within a high-dimensional feature space. Any samples deviating from this hyperplane are subsequently designated as anomalous.

In a similar vein, methods such as Support Vector Data Description (SVDD) [54], DeepSVDD [55], and PatchSVDD [56] strive to encapsulate normal data within a hypersphere, leveraging either kernel-based methodologies or self-supervised learning techniques. For a comprehensive overview of these implementations, we present detailed information in Table 3.2.

The representative one-class classification based methods are described as follows.

DeepSVDD[55] employs a neural network as a nonlinear mapper that transforms the raw data into a low-dimensional representation. This allows the model to capture the complex structure of the data and adapt to various data types. Specifically, the neural network maps the input data into a feature space where a hypersphere is constructed to encompass the normal data points.

PatchSVDD[56] extends DeepSVDD to a patch-wise approach, enhancing its performance for anomaly detection and segmentation. Specifically, PatchSVDD divides the input image into multiple small patches and processes each patch independently. The core idea is to map spatially proximal patches to similar locations in the hypersphere, leveraging self-supervised learning to endow the encoder's features with positional discriminability.

Normalizing Flow based methods

Normalizing Flow models data distributions meticulously [57]. In AD, it transforms normal features into a invertible distribution, crucial for distinguishing patterns. During inference, normal samples naturally align with the model's distribution, while abnormal samples are projected separately. This separation, facilitated by Normalizing Flow, enables efficient anomaly identification. While during inference, the trained flow model demonstrates its efficacy by naturally funneling normal samples into a tightly defined distribution range, where they reside comfortably within the confines of the model's learned representation. Conversely, abnormal samples, which do not conform to the patterns inherent in the normal data, are deftly projected onto a

distinct, separate distribution range. This segregation process, facilitated by Normalizing Flow, enables the efficient identification of anomalies that deviate from the norm. Advancements in this field[58–61] have introduced enhancements such as improved computational efficiency in AD scenarios.

The latest flow-based models for anomaly detection includes the following.

CS-Flow[59] framework efficiently processes multiple feature maps of varying scales in a unified manner. By employing normalizing flows, it assigns meaningful likelihoods to input image samples, thereby facilitating efficient defect detection at the image level. Furthermore, the spatial arrangement within the latent space of the normalizing flow remains intact, rendering it interpretable. This interpretability enables the precise localization of defective regions within the image, enhancing the overall detection capabilities of the system.

CFLOW[61] model is grounded in a conditional normalizing flow architecture tailored for anomaly detection with the ability to pinpoint defective areas. Specifically, CFLOW incorporates a discriminatively pre-trained encoder, which is subsequently followed by multi-scale generative decoders. These decoders explicitly assess the likelihood of the encoded features, leading to a model that is both computationally and memory-efficient.

Memory Bank based methods

Memory Bank is a mechanism of remembering numerical prototypes of the training date[62–65]. Then various algorithms such as KNN or statistical modeling are used to determine the labels for queries. PaDim, Patchcore, and CFA are the major basic frameworks for memory bank based methods. We provide detailed information in 3.4 to summarize all implementation aspects.

PaDiM[63] iemploys a pre-trained convolutional neural network (CNN) for patch embedding, alongside multivariate Gaussian distributions, to derive a probabilistic portrayal of the normal class. Furthermore, it capitalizes on the intricate correlations

existing across various semantic levels within the CNN architecture, enhancing its ability to precisely localize anomalies. Notably, PaDiM surpasses the current state-of-the-art methodologies for both anomaly detection and localization, as evidenced by its exceptional performance on benchmarks such as MVTec AD and STC datasets.

PatchCore[64] method revolutionizes anomaly detection by leveraging unsupervised learning exclusively with normal samples. It employs pre-trained models, such as WideResNet50, to extract image features and constructs a memory bank containing normal block representations. During the testing phase, PatchCore assesses the presence of anomalies by comparing the similarity between test samples' features and those stored in the memory bank. Key to its success are techniques like local patch feature aggregation for broader context and robustness, and a greedy core set algorithm to efficiently reduce the memory bank size while preserving its representativeness.

CFA[65] is to mitigate the bias of pre-trained CNNs through feature adaptation. Specifically, CFA performs transfer learning on the target dataset, optimizing the parameters of the patch descriptor and the contents of the memory bank to form highly concentrated coupled hyperspheres centered around memory features in the feature space. These hyperspheres distinguish normal from abnormal features through contrastive supervision.

Teacher-student (T-S) based methods

Recently, the T-S (Teacher-Student) architecture for knowledge distillation has emerged as a prevalent approach in anomaly detection (AD), as evidenced by numerous studies [11, 66–70]. In this framework, the student network learns to represent normal samples by mimicking the teacher network's behavior. However, for abnormal cases, the student may struggle to accurately follow the teacher's guidance, leading to a representation discrepancy between the T-S pair. This discrepancy serves as the cornerstone for anomaly detection, enabling the identification of patterns that deviate from the learned normality. To provide a comprehensive overview of the implemen-

tation details of this T-S approach, we have summarized all relevant information in Table 3.6. This table encapsulates the key aspects of the various implementations, facilitating a deeper understanding of the nuances and variations within the T-S architecture for AD.

The following are classic T-S models proposed for visual anomaly detection.

MKD[11] incorporates the concept of multiresolution, where the teacher network extracts features at multiple scales or resolutions. The student network is then trained to mimic this behavior, learning to reconstruct or interpret the multi-scale features in a way that captures the essence of normal patterns. By focusing on multiple resolutions, the method can capture both global and local information, making it more robust in detecting anomalies that manifest at different scales.

RD4AD[70] lies in utilizing a pre-trained teacher model as encoder to extract image features and distilling these features into a student decoder. The goal of the student decoder is to reconstruct the multi-scale features of the teacher model. However, since the student model only learns normal patterns during training, it is unable to reconstruct abnormal features, thereby enabling anomaly detection.

SimpleNet[67], an efficient network for image anomaly detection, integrates a pre-trained feature extractor, a feature adapter, an anomaly feature generator, and a binary discriminator. It leverages target-oriented features, synthetic anomalies in feature space, and a simple discriminator to achieve high performance while maintaining efficiency and practicality.

2.2.3 Vision-Language Model (VLM) Based Methods

In recent years, large-scale models have evolved rapidly, with Vision-Language models particularly standing out due to the abundance of pretrained data that enhances their capabilities for downstream tasks. Among these multimodal models, CLIP [71], whose diagram is illustrated in Fig 2.2, has emerged as a widely adopted multimodality approach for contrastive vision-language pretraining on vast datasets comprising

millions of image-text pairs. It leverages the unique one-to-one correspondence between paired data, generating natural positive and negative samples. Due to the rich knowledge gained from the contrastive learning on millions of data, the pre-trained multimodal model has imparted remarkable versatility, enabling them to excel even in zero-shot inference. Their prowess lies in their prompt-guided zero-shot capability, allowing them to accurately identify and categorize previously unseen images during the inference phase. Recent research has also expanded the zero-shot transferability of CLIP models to open-vocabulary semantic segmentation by extracting intrinsic dense features [63, 70, 72]. Efforts have been made to significantly improve CLIP's recognition performance, including prompt engineering [73] and adapter modules [71, 74].

Notably, CLIP's innate ability to detect out-of-distribution data without additional training inspired the us of CLIP for zero-shot anomaly detection. For instance, WinCLIP [75] introduced a multi-scale window moving approach for local patch processing, followed by CLIP-based classification of each window. However, it suffers from time-consuming computations due to the additional windowed image patch processing. Following WinCLIP, latest anomaly detection studies have begun incorporating CLIP models, replacing previously prevalent pretrained models like ImageNet [63, 70, 72]. for zero-shot AD tasks [76–78].

To further enhance the generalization ability of these large VLM models in specific anomaly detection domains, fine-tuning strategies have emerged. By tailoring the models to these domains based on their pre-existing large-scale structures, they become better suited for detecting anomalies in specific contexts, thereby improving their overall performance and robustness.

In this thesis, we proposed a network utilizes the CLIP model in a full-shot manner to enhance the anomaly detection performances across various medical domains. We incorporate the text encoder in the CLIP model to restore the normal image features. For specific details, please refer to Chapter 4 for a thorough reading.

2.3 Medical Anomaly Detection Works

Consistent with the categorization of general anomaly detection methods, approaches deployed in medical anomaly detection can be fundamentally divided into analogous groups. In the context of medical image anomaly detection, autoencoders have been extensively trained on datasets solely comprising healthy date [28, 79]. Deviations from the learned patterns subsequently result in an elevated anomaly score, highlighting potential abnormalities. This principle has been successfully applied to unsupervised anomaly detection within medical imagery [9, 80, 81], where the disparity between the reconstructed healthy image and the anomalous input serves to pinpoint abnormal pixels. Additionally, other research efforts have focused on harnessing Generative Adversarial Networks (GANs) [12, 82] for image-to-image translation tasks [83–85]. Transformer networks [14, 86] have also demonstrated remarkable success in brain anomaly detection. Moreover, in [87], a novel thresholding methodology is introduced, specifically tailored for segmenting brain anomalies, further advancing the field of medical anomaly detection. Recently, diffusion models[16, 88] have emerged as a powerful tool for medical anomaly detection, advancing the capabilities of GANs and demonstrating promising performance in multi medical domains.

Table 2.1: References of traditional image anomaly detection methods.

Method based	References
Statistical Modeling	In statistics, anomaly detection involves identifying observations that deviate from the norm. This is accomplished through statistical modeling using various methods, including parametric, normal distribution-based[19–21]. And non-parametric, distance-based approaches[22]. As well as advanced models like time series analysis and clustering[23, 24].
Bayesian Networks (BN)	BN has significant value in anomaly detection. As a probabilistic graphical model, it represents variable dependencies through nodes and edges, using conditional probabilities. This handles uncertainty, incompleteness, correlations well, making BN ideal for anomaly detection [25].
Support Vector Machines (SVM)	SVM primarily relies on an algorithm known as unsupervised SVM. SVM aims to learn a model from the training data that can distinguish between normal data and anomalous data, without the need for explicitly providing samples of anomalous data [26].

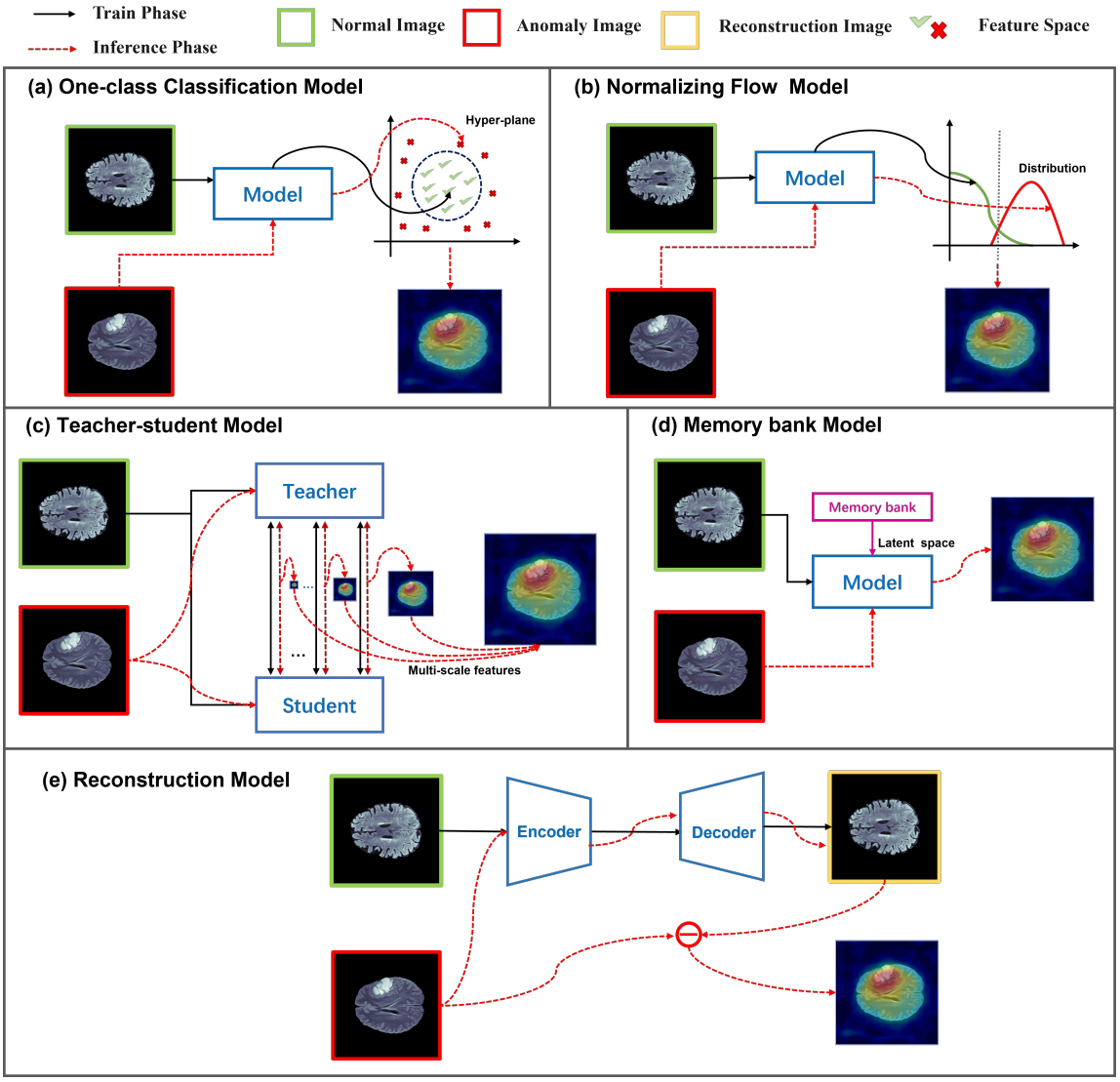

Figure 2.1: Conceptual illustration of various deep learning based AD models. The one-class classification model, normalizing flow model, teaching-student model and memory bank model detects anomalies in the embedding space, and the reconstruction based method takes a generative model as its backbone for pixel-level anomaly comparison between the original query and reconstruction.

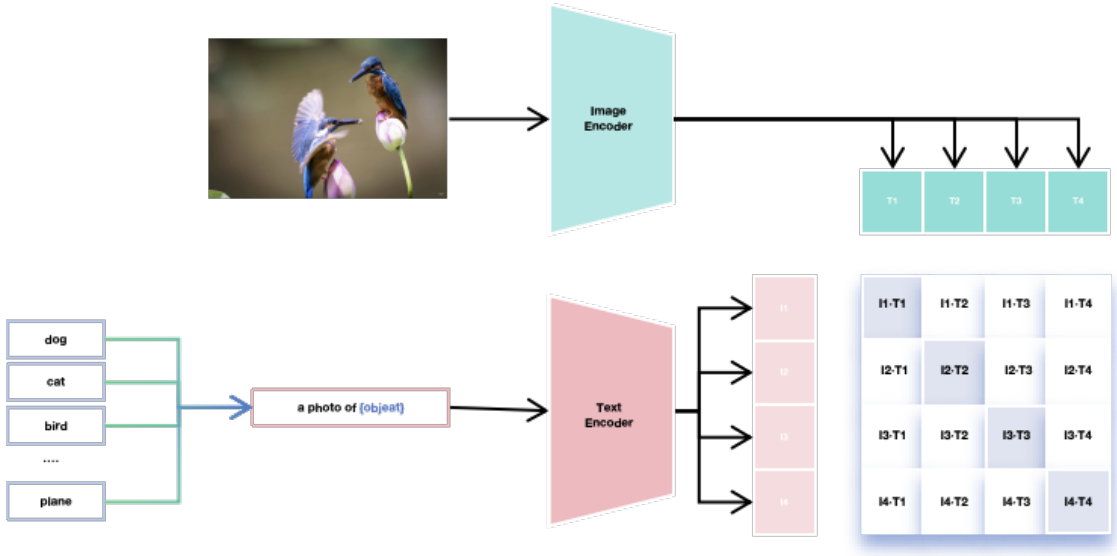

Figure 2.2: Diagram of the CLIP pretrain model.

Chapter 3

BMAD: Benchmarks for Medical Anomaly Detection

Anomaly detection (AD) stands as a cornerstone research challenge in machine learning and computer vision, with vital real-world applications spanning industrial inspection, video surveillance, and notably, medical diagnosis. Within the realm of medical imaging, AD holds paramount significance in pinpointing anomalies that may serve as indicators of rare diseases or conditions. Nonetheless, despite its pivotal role, a universal and equitable benchmark for evaluating AD techniques on medical images remains elusive, impeding the progress towards more generalized and resilient AD methods tailored to this domain.

To bridge this gap, we introduce a comprehensive evaluation benchmark tailored specifically for assessing AD methods on medical images. As shown in Figure 3.1, this benchmark encompasses six meticulously reorganized datasets, sourced from five diverse medical domains: brain MRI, liver CT, retinal OCT, chest X-ray, and digital histopathology. Furthermore, it incorporates three pivotal evaluation metrics and features an extensive lineup of fifteen cutting-edge AD algorithms. Our standardized and meticulously curated medical benchmark, accompanied by a well-structured codebase, empowers researchers to effortlessly compare and evaluate diverse AD approaches. This, in turn, fosters the development of more efficient and robust AD algorithms tailored for medical imaging, thereby advancing the state-of-the-art in this

crucial field. For additional insights into BMAD, please visit our GitHub repository at https://github.com/DorisBao/BMAD.

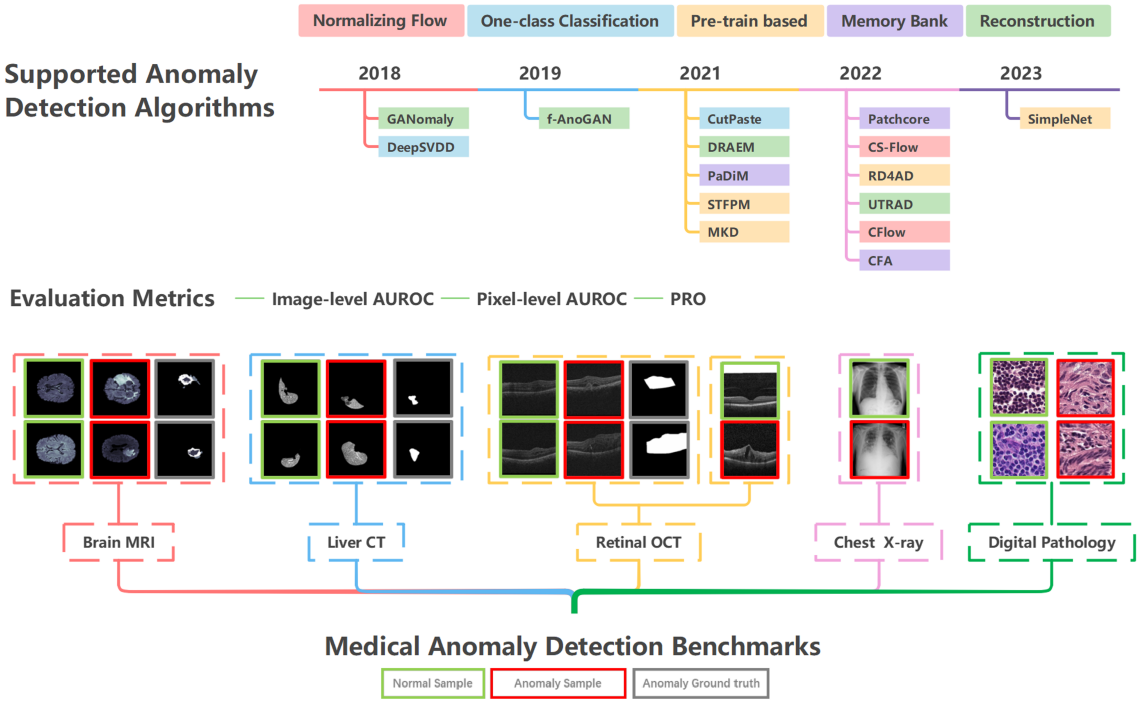

Figure 3.1: Diagram of the BMAD benchmarks. BMAD includes six datasets from five different domains for medical anomaly detection, among which three support pixel-level AD evaluation and the other three for sample-level assessment only. BMAD provides a well-structured and easy-used code base, integrating fifteen SOTA anomaly detection algorithms and three evaluation metrics.

3.1 Datasets and Benchmarks

When constructing this benchmark, we had following considerations in dataset selection: diversity of imaging modalities, diversity of source domains/organs, and license for data reorganization, remix and redistribution. Specifically, our BMAD includes six medical benchmarks from five different domains for medical anomaly detection, including brain MRI, retinal OCT, liver CT, chest X-ray, and digital histopathology. We summarize these benchmarks in Table. 3.1. Within these benchmarks, three supports pixel-level evaluation of anomaly detection, while the remaining three is for sample-level assessment only.

Table 3.1: Summary of the six benchmarks from five imaging domains in BMAD.

Domains	**Originations**	**Annotation Level**
Brain MRI	BraTS2021[89]	Segmentation mask
Liver CT	BTCV[90] + LiTs[91]	Segmentation mask
Retinal OCT	RESC[92]	Segmentation mask
Chest X-ray	RSNA[8]	Image label
Pathology	Camelyon16[93]	Image label
Retinal OCT	OCT2017[94]	Image label

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. To facilitate future research and potential benchmark extension, our benchmark 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.

3.1.1 Brain MRI Anomaly Detection Benchmark

Magnetic Resonance Imaging (MRI) imaging is widely utilized in brain tumor examination. The Brain MRI AD benchmark is reorganized using the flair modality of the latest large-scale brain lesion segmentation dataset, BraTS2021 [89].

The original BraTS2021 dataset is proposed for multimodal brain tumor segmentationm comprising a collection of the complete 3D volume of a patient's brain structure and corresponding brain tumor segmentation annotation. 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 tu-

mor segmentation annotation. The data size for each modality is 240 *240 *155. 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.

To adapt the data to AD, 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. In specific, we sliced both the brain scan and their annotation along the axial plane. Only slides containing substantial brain structures, usually with a depth of 60-100, were selected in this benchmark. Slices without brain tumor are labelled as normal. Each extracted 2D slice was saved in PNG format and has an image size of 240 * 240 pixels. With obtained image slides, to avoid data leakage in model evaluation, we leveraged the information of patient IDs for data partition and ensured that data from the same patient was contained by one set only. 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. 3.2 illustrates the specific procedure we followed for data preparation, and Fig. 3.3 provides examples of our brain MRI AD benchmark.

3.1.2 Liver CT Anomaly Detection Benchmark

Computed Tomography (CT) is commonly used for abdominal examination. We structure this benchmark from two distinct datasets, BTCV[90] and Liver Tumor Segmentation (LiTs) set[91]. The anomaly-free BTCV set is initially proposed for multi-organ segmentation on abdominal CTs and taken to constitute the train set in this benchmark. CT scans in LiTs is exploited to form the evaluation and test data.

BTCV [90] is introduced for multi-organ segmentation. It consists of 50 abdomi-

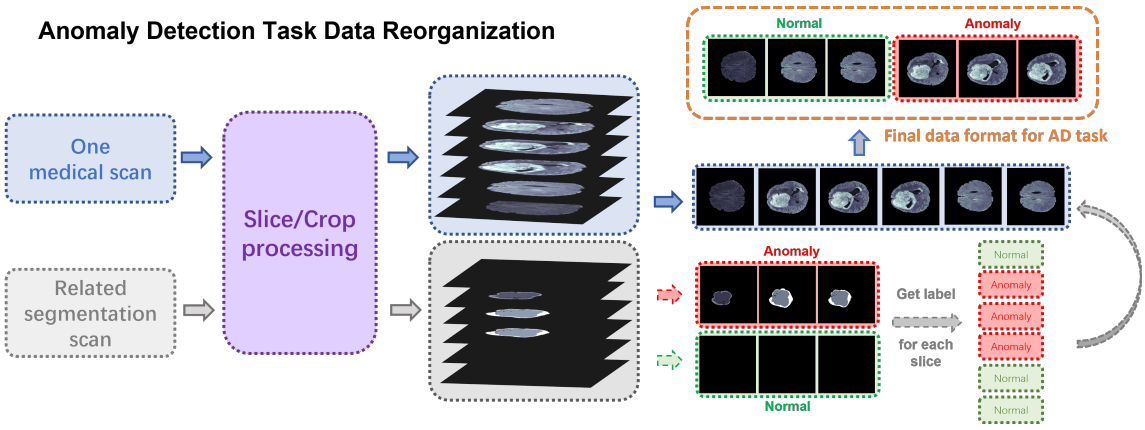

Figure 3.2: Diagram illustration of data preparation for the Brain MRI AD benchmark from 3D brain scans in BraTS2021.

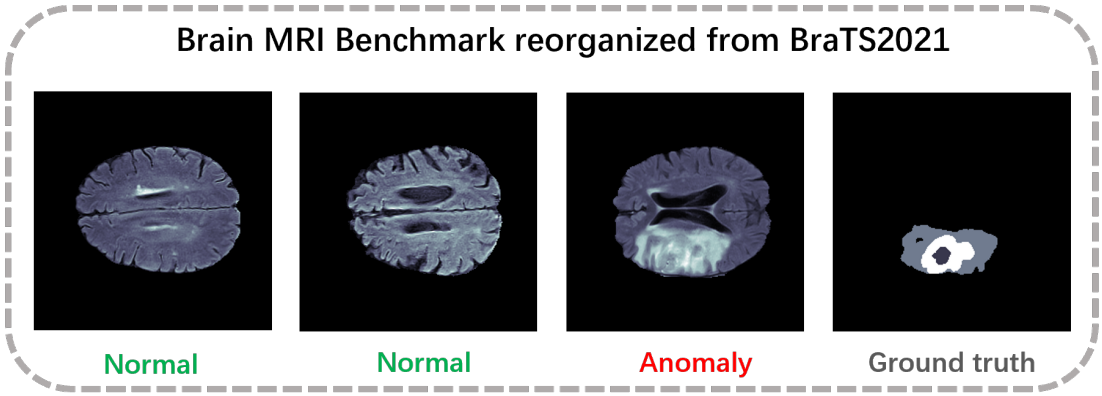

Figure 3.3: Visualization of our proposed Brain MRI benchmark.

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

LiTS [91] 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. The orignal LiTS dataset 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) [95].

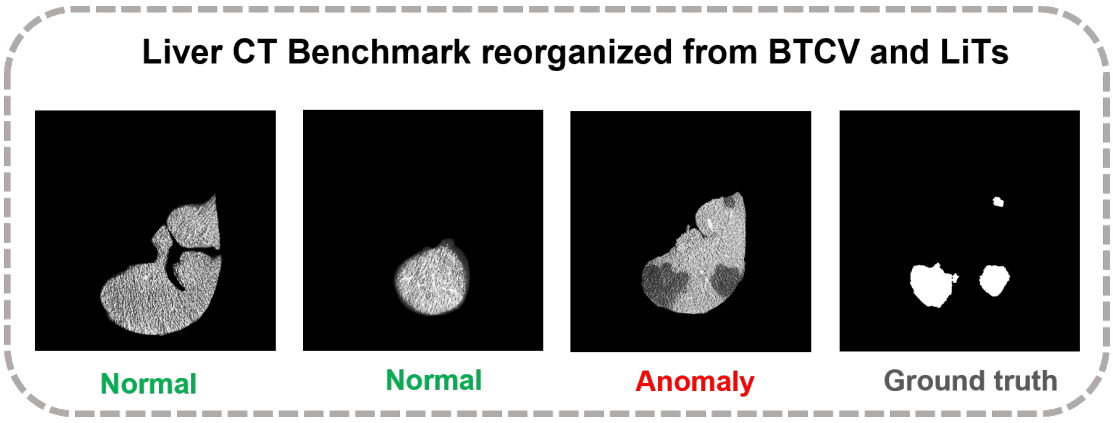

Figure 3.4: Visualization of our proposed Liver CT benchmark.

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. Following conversion in prior arts[96, 97], we further performed histogram equalization on each slide for image enhancement. 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 3.4 demonstrates several samples in the Liver CT AD dataset.

It should be noted that with the histogram equalization, some of the image content may be distorted due to the change of image intensity. Thus, for completeness, we also provide a version of the liver benchmark without any data processing in BMAD. This allow researchers to access both versions and make informed decisions based on their specific needs.

3.1.3 Retinal OCT Anomaly Detection Benchmark

Optical Coherence Tomography (OCT) is commonly used for scanning ocular lesions in eye pathology. To cover a wide range of anomalies and evaluate anomaly localization, the BMAD datasets includes two different OCT anomaly detection datasets. The first one is derived from the RESC dataset [92] and support anomaly localization evaluation. The second is constructed from OCT2017 [94], Which only support sample-level anomaly detection.

RESC (Retinal Edema Segmentation Challenge) dataset [92] specifically focuses on the detection and segmentation of retinal edema anomalies. The original training, validation, and test sets contain 70, 15, and 15 cases, respectively. Each case includes 128 slices, some of which suffer from retina edema. 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. The original RESC dataset 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.

OCT2017 [94] is a large-scale dataset initially designed for classification tasks.Images are categorized into 4 classes: normal, Choroidal Neovascularization, Diabetic Mac-

ular Edema, and Drusen Deposits. The images are continuous slices with a size of 512*496. The original OCT2017 data can be downloaded at: https://data.mendeley.com/datasets/rs Its usage is under a license of Creative Commons Attribution 4.0 International(CC BY 4.0).

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. To avoid data leakage, slices from the same subject only appear in either validation or test set. In specific, we select the normal samples from the original training dataset and adapt the original validation set into the AD setting for evaluation. On the other hand, to construct this sample-level anomaly detection benchmark from OCT2017, we use the disease-free samples in the original OCT2017 training set as our training data. For images in the original test set, images in the 3 diseased classes are labeled as abnormal. Stratified sampling is adopted to form the evaluation and test sets. Fig 3.5 demonstrates several examples in the two OCT AD datasets.

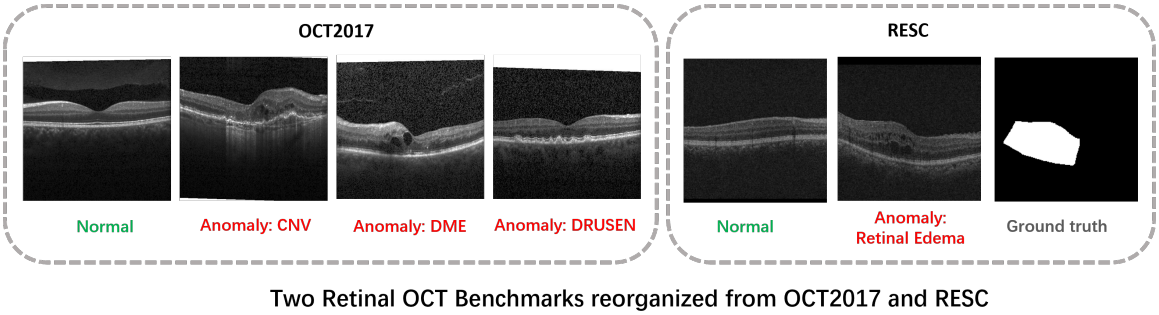

Figure 3.5: The Retinal OCT benchmarks consist of two separate datasets, each representing different anomaly types. These datasets are used to evaluate and benchmark various methods in the field of retinal OCT imaging. The datasets are designed to assess the performance of algorithms in detecting and localizing specific anomalies in retinal images.

3.1.4 Chest X-ray Anomaly Detection Benchmark

X-ray imaging is widely used for examining the chest and provides precise thoracic data. We reconstruct the X-ray anomaly detection benchmark from the RSNA dataset

that was orignally created for levering ML models for chest X-Ray diagnosis[8].

RSNA [8], short for RSNA Pneumonia Detection Challenge, is originally provided for a lung pneumonia detection task. The lung images are associated with nine labels: Normal, Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumothorax, which covers the eight common thoracic diseases observed in chest X-rays. All images are stored in DICOM format. The original RSNA data can be accessed by: https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/overview. As 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."

To reorganize the dataset for anomaly detection, we utilized the provided image labels for data re-partition. Specifically, "Lung Opacity" and "No Lung Opacity/Not Normal" were classified as abnormal data, and we labelled images in the abnormal categories as abnormal. We follow the original datasheet and split the data into train, test, and validation sets for anomaly detection. Consequently, 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 3.6.

3.1.5 Digital Histopathology Anomaly Detection Benchmark

Histopathology involves the microscopic examination of tissue samples to study and diagnose diseases such as cancer. We utilize Camelyon16[93], a digital pathology imaging breast cancer metastasis detection dataset, to build the histopathology benchmark.

Camelyon16 [93] was initially utilized in the Camelyon16 Grand Challenge to detect and classify metastatic breast cancer in lymph node tissue. It comprises 400 40x whole-slide images (WSIs) of lymph node sections stained with hematoxylin

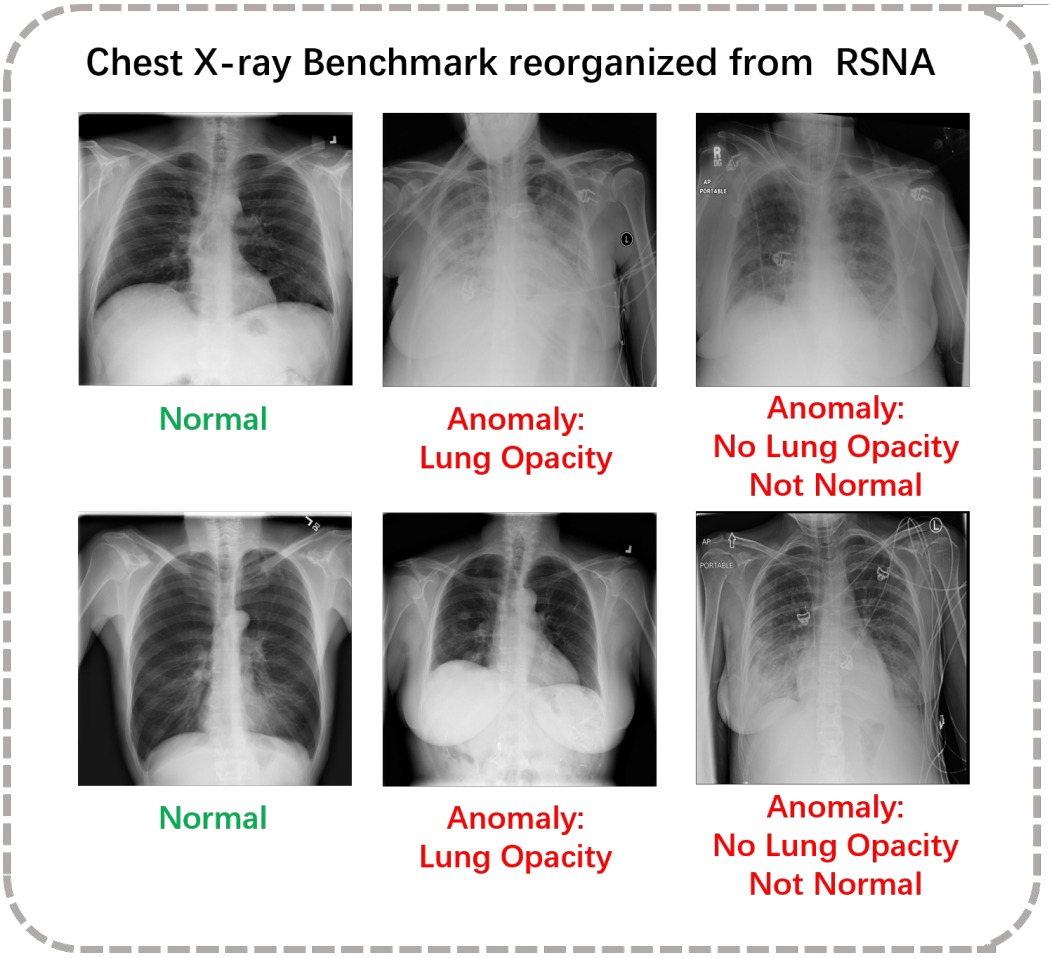

Figure 3.6: Our proposed chest X-ray benchmark consists two types of anomalies. These anomalies are clearly labeled in the images, and all of them are considered as anomaly samples.

and eosin (H&E) from breast cancer patients, accompanied by multiple versions in a lower magnification. Among these WSIs, 159 of them exhibit tumor metastases, which have been annotated by pathologists. The WSIs are stored in standard TIFF files. Note, in Camelyon16, the highest resolution available is on level 0, corresponding to a magnification of 40X. 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).

To construct the benchmark histopathology image dataset, considering their unique characteristics such as large size, we follow convention in prior arts[98–100] and opted

to assess AD models at the patch level in 40X. Specifically, 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 3.7 presents several examples in the constructed histopathology AD benchmark.

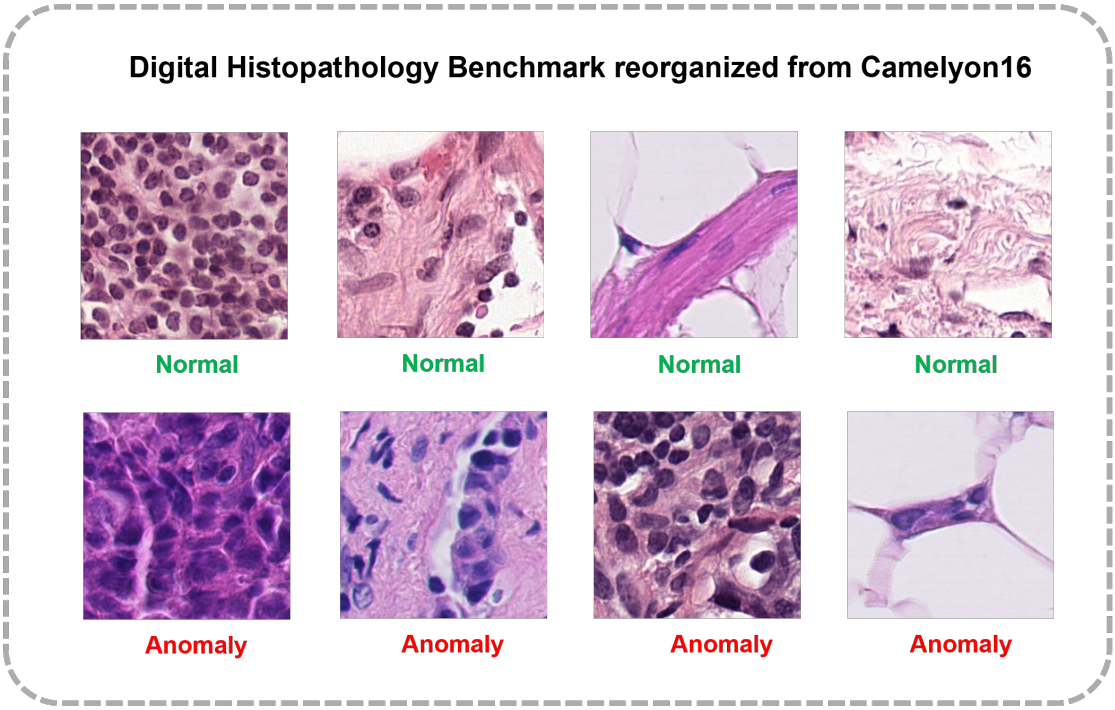

Figure 3.7: Examples of the digital histopathology AD benchmark. Unlike other medical image AD benchmarks, histopathology images shows higher diversities in tissue components.

3.1.6 Overall Remark

Among the 7 original datasets used to construct BMAD are from advanced countries. This gives rise to inherent geographical and sampling biases, which inevitably exerts some impact on the evaluation outcomes.

3.2 Evaluation Metrics

Anomaly detection can be evaluated from the sample level (i.e., detection rate) and the pixel level (i.e., anomaly localization). In accordance with established practices in previous literature, we employ **AUROC** (Area Under the Receiver Operating Characteristic Curve) for evaluations in both levels. Note that AUROC has limitations to evaluate small tumor localization, as incorrect localization of smaller defect regions has a minimal impact on the metric. To address this issue, we follow prior arts[70, 101, 102] and include another threshold-independent metric, **PRO** (Per-Region Overlap), for anomaly localization evaluation.

Area Under the Receiver Operating Characteristic Curve

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.

$$AUROC = \int_0^1 (TPR)\mathrm{d}(FPR) \tag{3.1}$$

- 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.

Per-Region Overlap

We utilized PRO, a region-level metric,to assess the performance of fine-grained anomaly detection. PRO treats anomaly regions of different size equally, up-weighting the influence of small-size abnormality localization in evaluation. Specifically, for each threshold, detected anomalous pixels are grouped into connected components and then PRO averages localization accuracy over all components. 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,

$$PRO = \frac{1}{N} \sum_{k} \frac{|A \cap C_k|}{|C_k|}, \tag{3.2}$$

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

DICE score

The Dice score is an important metric in medical image segmentation, evaluating the similarity between segmented results and reference standards. Though DICE is not included as the major metric due to its threshold dependence, our codebase still includes a Dice function for its potential usage. It measures the pixel-level overlap between predicted and reference regions, ranging from 0 (no agreement) to 1 (perfect agreement).

$$DICE(\tilde{M}_b, M_b) = \frac{2 \cdot |\tilde{M}_b \cap M_b|}{|\tilde{M}_b| + |M_b|}, \tag{3.3}$$

where $\tilde{M}_b, M_b$ are the binary value for $\tilde{M}, M$. And $\tilde{M}, M$ represent the prediction score for the whole test set.

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. Instead, we reported the DICE values of the 15 algorithms for reference.

3.3 Supported AD Algorithms

BMAD integrates fifteen SOTA anomaly detection algorithms, among which four are reconstruction-based methods and the rest eleven are feature embedding-based approaches. Among the reconstruction-based methods, **AnoGAN**[5] and **f-AnoGAN**[39] exploit the GAN architecture to generate normal samples. **DRAEM**[50] adopts an encoder-decoder architecture for abnormality inpainting. Then a binary classifier takes the original data and the inpainting result as input for anomaly identification. **UTRAD**[13]treated the deep pre-trained features as dispersed word tokens and construct an autoencoder with transformer blocks. Among the projection-based methods, **DeepSVDD** [55], **CutPaste**[72] and **SimpleNet**[67] are rooted in one-class classification. DeepSVDD searches a smallest hyper-sphere to enclose all normal embeddings extracted from a pre-tarined model. CutPaste and SimpleNet introduce abnormality synthesis algorithms to extend the one-class classification, where generated abnormality synthesis is taken as negative samples in model training. Motivated by the paradigm of knowledge distillation, **MKD** [11] and **STFPM** [69] leverage multi-scale feature discrepancy between the teacher-student pair for AD. Instead of adopting the similar backbones for the T-S pair in knowledge distillation, **RD4AD** [70] introduced a novel architecture consisting of a teacher encoder and a student decoder, which significantly enlarges the representation dissimilarity for anomaly samples. All of **PaDiM** [63], **PatchCore** [64] and **CFA** [65] rely on a memory bank to store normal prototypes. Specifically, PaDiM utilizes a pre-trained model for feature extraction and models the obtained features using a Gaussion distribution. PatchCore leverages core-set sampling to construct a memory bank and adopts the nearest neighbor search to vote for a normal or abnormal prediction. CFA improves upon PatchCore by creating the memory bank based on the distribution of image features on a hyper-sphere.

As notable from the name, **CFlow** [61] and **CS-Flow** [59] are flow-based methods. The former introduced positional encoding in conjunction with a normalizing flow module and the latter incorporates multi-scale features for distribution estimation.

Table 3.2: Model detail of One-class classification based methods.

Implementation Details
DeepSVDD[55] 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.
PatchSVDD [56], in contrast to DeepSVDD, examines images at the patch level, exhibiting commendable performance in localized detection. The proposed model introduces the utilization of hierarchical encoders, featuring hyper-parameters of 64 for feature dimensions and 0.9 for λ, which effectively balances the loss function.

Table 3.3: Model detail of Flow based methods.

Implementation Details
CS-Flow [59] 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.
CFLOW [61] is a normalizing flows-based method. We utilized Wide Resnet-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.

Table 3.4: Model detail of Mermory Bank based methods.

Implementation Details
Patchcore[64] 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 hyper-parameters of 0.1 for the coreset sampling ratio and 9 for the chosen neighbor number. These values were chosen based on the original implementation.
PaDiM [63] 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 Wide Resnet-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 Wide Resnet-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.
CFA [65] is also a memory bank-based algorithm. We employs a Wide Resnet-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.

3.4 Experiments and Discussions

3.4.1 Implementation Details

When evaluating the fifteen AD algorithms over the BMAD benchmarks, we follow their original papers and try their default hyper-parameter settings first. If a model doesn't converge during training and requires hyper-parameter tuning, we try the combination of following common settings, which include 3 learning rate (10^{-3},10^{-4} and 10^{-5}), 2 optimizer (SGD and Adam), and 3 thresholds for anomaly maps (0.5, 0.6, and 0.7). Please refer to the Table 3.2,3.3,3.4,3.5,3.6 for the specific hyper-

Table 3.5: Model detail of Reconstruction-based methods.

Implementation Details
DRAEM[50] 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.
CutPaste[72] 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[38] 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.
f-AnoGAN[39] 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.

parameter setting for each algorithm. For each converged model, we monitor the training progress and record the validation accuracy every 10 epochs. The final evaluation is carried out on the test set using the best checkpoint selected by the validation sets. To visualize anomaly localization results, we employ min-max normalization on the obtained anomaly maps. This ensures the effects of all algorithms appropriately displayed and facilitates the comparison of anomaly localization across different methods. Notably, for a reliable comparison, we repeat the training and evaluation five times, each with a different random seed, and report the mean and standard deviation of the numerical metrics. In this study, all experiments are performed on a workstation with 2 NVIDIA RTX 3090 GPU cards.

3.4.2 Results and Discussions

Experimental result overview. The numerical results of anomaly detection and localization over the BMAD benchmark are summarized in Tab. 3.7, 3.8, 3.9, 3.10, where the top three performance along each metric are highlighted by underlining. We also provide visualization examples of anomaly localization results in Fig. 3.8, where redness corresponds to a high anomaly score at the pixel level. Although no single algorithm consistently outperforms others, overall, the feature-based methods shows better performance than the reconstruction-based methods. We believe that two reasons may lead to this observation. First, applying generative models to anomaly detection usually relies on model's reconstruction residue in the pixel level. However, a well-trained generative model usually has good generalizability and it has been found in prior arts that certain anomalous regions can be well reconstructed. This issue hurts anomaly detection performance. Second, reconstruction residue in the pixel level may not well reflect the high-level, context abnormalities. In contrast, algorithms detecting abnormalities from the latent representation domain (such as RD4AD [70], PatchCore [64], etc.) facilitate identifying abstract structural anomalies. Therefore, these algorithms perform much better than the generative models. It should be noted that for benchmarks like Liver CT and Brain MRI, where the background consists mostly of black pixels and the distribution of normal and anomalous samples is imbalanced, the numerical results exist bias. Therefore, a high pixel-level AUROC score may indicate that the model correctly classifies the majority of normal pixels, but it does not necessarily reflect the model's ability to detect anomalies accurately. Besides, we have several interesting observations through this research that necessitate careful analysis in order to advance the field of medical anomaly detection. We elaborate our insights and discoveries as follows.

Anomaly localization analysis. Since different approaches generate the anomaly map in various ways, either relying on reconstruction error [13, 50, 70, 103], using

gradient-based visualization [11, 64], or measuring feature discrepancy [61, 63, 65], they shows distinct advantages and limitations. Generally speaking, both numerical data in Table. 3.8, Table. 3.7, Table. 3.10 and the visualization results in Fig. 3.8 demonstrate that knowledge-distillation methods, especially RD4AD [70], achieve better localization performance. Although memory bank-based algorithm, Patchcore [64], is more convincing at sample-level detection, its abnormality localization is very coarse. Reconstruction-based algorithms, DRAEM [50] and UTRAD [13], shows diverse performance. We hypothesize their distinct capability of anomaly localization is attributed to the different architecture of CNN and transformer. We notice that DRAEM [50] is particularly sensitive to texture information, often focusing on regions with significant variations in tumor texture. Since such variations may be distributed across all regions in medical imaging, it partially limits the effectiveness of the proposed approach. CFlow [61] shows bad anomaly localization performance and more investigation is needed for its improvement.

Model efficiency analysis. For all fifteen algorithms in BMAD, we conduct a comparative efficiency analysis, in terms of sample-level AD accuracy, inference speed and GPU usage. The results are summarized in Fig. 3.9, where the X-axis refers to the inference time per image and Y-axis denotes the performance of the anomaly detection result. The size of the circle denotes the GPU memory consumption during the inference phase. PatchCore[64], RD4AD[70], and CS-FLOW[59] emerge as the top 3 models across multiple benchmarks in terms of performance. It should be noted that though CS-Flow demonstrates comparable inference time to the other two models, it has lower efficiency to generate pixel-lever anomaly maps.

Anomaly synthesis is challenging. In unsupervised AD methods, one common approach is to synthesize abnormalities to augment model training. CutPaste[72] and DRAEM[50] are the examples. However,to address the variability in shape, texture, and color of medical anomalies across different domains, a customized synthesis algorithm is needed to simulate realistic tumor lesions and their distributions. It

is important to acknowledge the inherent difficulty in simulating the morphology of anomalies, and this challenge becomes even more pronounced when considering rare diseases.

We discovered that the Brain MRI and Liver CT benchmarks are better suited for low-level feature-based anomaly augmentation methods. This observation aligns with the characteristics of the Chest X-ray and Histopathology benchmarks, where abnormalities often exhibit distinct and observable changes in overall structure or appearance. Therefore, it is essential to develop domain-specific approaches that account for these factors when augmenting anomalies in medical image datasets.

Pre-trained networks significantly contribute to medical domain. Through there is a continuous debate on if information obtained from natural images is transferable to medical image analysis, our results show that the rich representations of pre-trained models would improve medical anomaly detection by careful algorithm design. Among the models evaluated, algorithms based on the knowledge-distillation paradigm (e.g. MKD [11] and RD4AD [70]) and memory bank (e.g. Patchcore [64]) leverage the powerful feature extraction capabilities of large pre-train models and exhibit better performance in anomaly localization, which plays a crucial role in clinical diagnosis. SimpleNet[67] utilized a pre-trained feature extraction module alongside a Gaussian denoising module, proving effective for enhancing low-level feature images. This also suggests that the denoising module operates optimally within the same repersatatation will fit the best for the detection.

Memory bank-based methods have shown promising performances. PatchCore[64] is a representative example. These methods possess the ability to incorporate new memories, effectively mitigating forgetting when learning new tasks. Hence, the memory bank serves as an ideal rehearsal mechanism. However, these methods have specific hardware requirements to ensure efficient storage and retrieval of stored information. Achieving high-capacity storage systems and efficient memory access mechanisms for optimal performance while minimizing interference time presents a

notable challenge. Furthermore, our observations indicate that memory-based methods, while sensitive to global anomalies, may not excel in terms of localizing and visualizing anomalies when compared to feature reconstruction methods. Accurate anomaly localization holds crucial practical value for AD algorithms and provides valuable insights to medical professionals. Therefore, memory bank-based methods may encounter challenges and limitations that impact their competitiveness in certain scenarios.

Model degradation problem. The model degradation problem refers to the phenomenon where a deep neural network, trained on a large dataset, experiences a decline in performance as the network's depth increases. In our study, we have observed that this issue also exists within the BMAD benchmark. However, addressing this problem by adding appropriate preprocessing and data augmentation techniques to medical benchmarks poses a significant challenge. We propose that incorporating adversarial training for medical data could be a promising approach to enhance the robustness of the models. Adversarial training involves exposing the model to adversarial examples during the training process, which are inputs specifically designed to deceive the model. By training the model to resist such attacks, it can improve its ability to generalize and perform well on unseen data. This approach has shown promise in improving the robustness of deep learning models in various domains, and we believe it could be beneficial for medical anomaly detection as well.

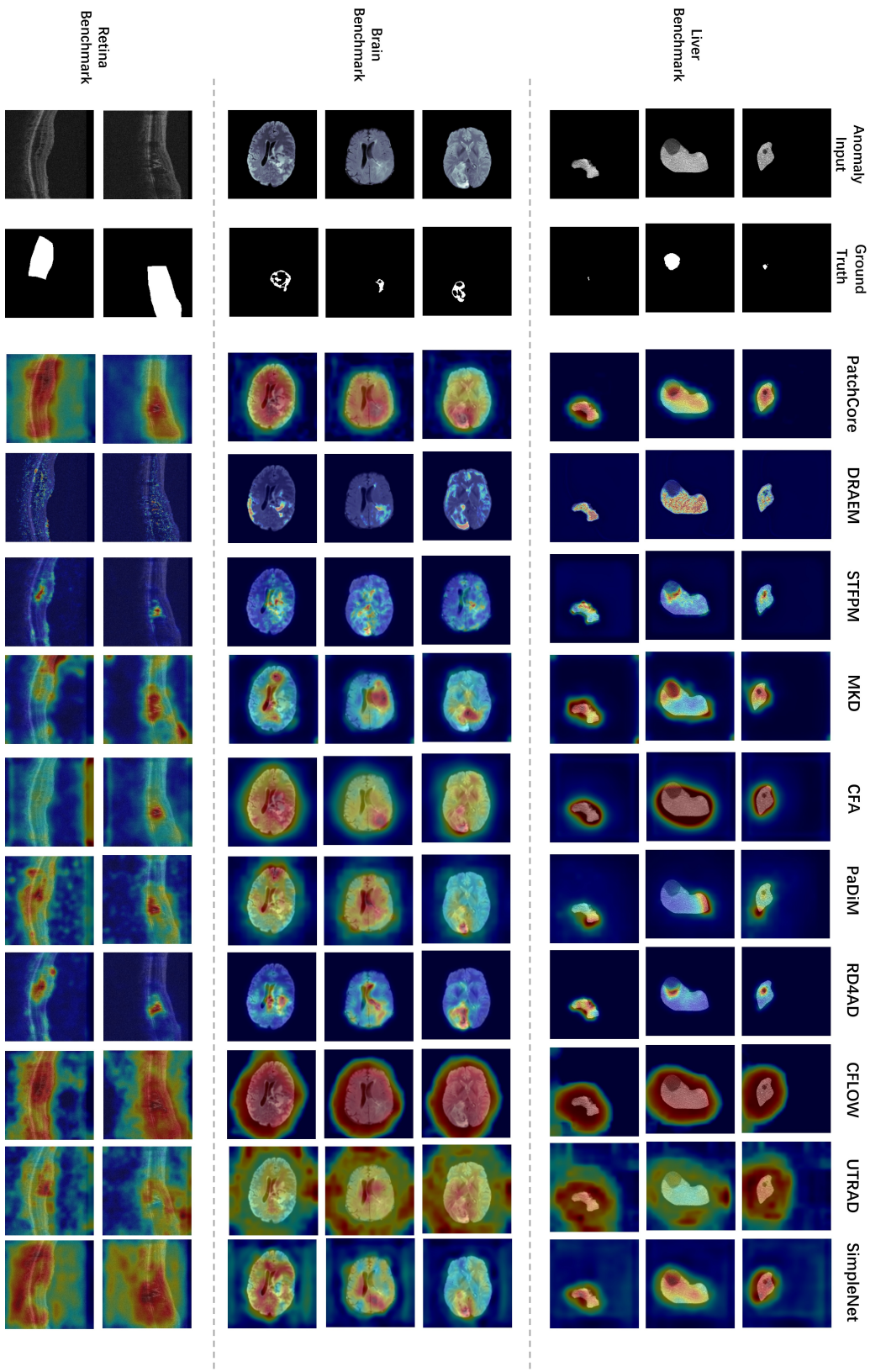

Figure 3.8: Visualization examples of anomaly localization on the three benchmarks that support pixel-level AD assessment.

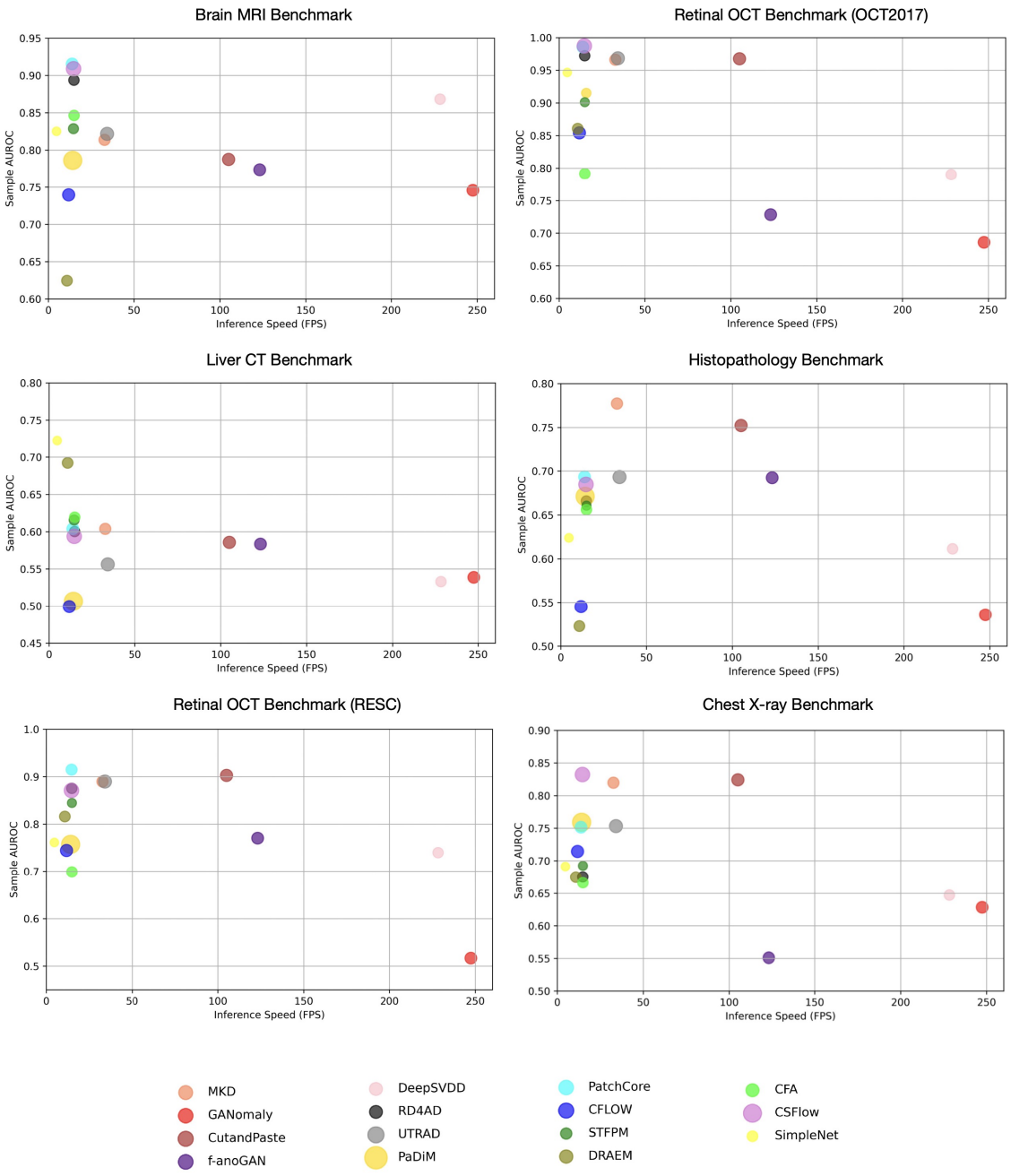

Figure 3.9: Model Efficiency Analysis. X-axis refers to the average inference time per image and Y-axis denotes anomaly detection accuracy. The size of the circle denotes the GPU memory consumption during the inference phase. In the sub-images, there may be slight variations in the results due to model adjustments like selecting specific parameters and backbones on each benchmark.

Table 3.6: Model detail of T-S based methods.

Implementation Details
RD4AD[70] 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.
STFPM[69] utilized feature extraction from a Teacher-student structure. In our experiments, we evaluated all benchmarks separately using two backbone networks: ResNet-18 and Wide Resnet-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.
MKD[11] 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[13] 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.
SimpleNet[67] 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

Methods	BTCV + LiTs		
	Image AUROC	**Pixel AUROC**	**Pixel Pro**
f-AnoGAN [39]	58.53 ± 0.21	NA	NA
GANomaly [38]	54.60 ± 3.06	NA	NA
DRAEM [50]	$\underline{69.95 \pm 3.86}$	87.45 ± 3.23	79.29 ± 5.66
UTRAD[13]	55.81 ± 5.66	87.88 ± 1.32	71.12 ± 3.46
DeepSVDD [55]	53.96 ± 1.84	NA	NA
CutPaste [72]	59.33 ± 4.86	NA	NA
SimpleNet [67]	$\underline{72.28 \pm 2.68}$	$\underline{97.51 \pm 0.56}$	$\underline{91.07 \pm 1.79}$
MKD [11]	60.72 ± 1.19	$\underline{96.06 \pm 0.27}$	$\underline{91.08 \pm 0.30}$
RD4AD[70]	60.38 ± 1.17	96.01 ± 1.19	$\underline{90.29 \pm 2.51}$
STFPM[69]	61.75 ± 1.58	91.18 ± 5.52	90.62 ± 6.87
PaDiM [63]	50.78 ± 0.61	90.94 ± 0.84	76.79 ± 0.41
PatchCore[64]	60.28 ± 0.76	$\underline{96.43 \pm 0.19}$	87.75 ± 0.49
CFA [65]	$\underline{62.00 \pm 1.08}$	$\underline{97.24 \pm 0.14}$	$\underline{92.75 \pm 0.21}$
CFLOW [61]	50.80 ± 4.47	92.41 ± 1.16	83.11 ± 1.28
CS-Flow [59]	59.37 ± 0.54	NA	NA

Table 3.7: Comparison of anomaly detection performance on liver CT benchmark. We report the mean and standard deviation over 5 random seeds for each measurement. Bold indicates the best performance.

Methods	BraTS2021		
	Image AUROC	**Pixel AUROC**	**Pixel Pro**
f-AnoGAN [39]	77.26 ± 0.18	NA	NA
GANomaly [38]	74.79 ± 1.93	NA	NA
DRAEM [50]	62.35 ± 9.03	82.29 ± 4.07	63.76 ± 4.16
UTRAD[13]	82.92 ± 2.32	92.61 ± 0.67	72.29 ± 2.12
DeepSVDD [55]	86.98 ± 0.66	NA	NA
CutPaste [72]	78.81 ± 0.67	NA	NA
SimpleNet [67]	82.52 ± 3.34	94.76 ± 1.04	78.38 ± 3.17
MKD [11]	81.47 ± 0.36	89.44 ± 0.24	67.59 ± 0.99
RD4AD[70]	$\underline{89.45 \pm 0.91}$	$\underline{96.45 \pm 0.17}$	$\underline{85.86 \pm 0.23}$
STFPM[69]	83.04 ± 0.67	95.62 ± 0.12	83.02 ± 0.44
PaDiM [63]	79.02 ± 0.38	94.37 ± 1.03	76.41 ± 0.84
PatchCore[64]	$\underline{91.65 \pm 0.36}$	$\underline{96.97 \pm 0.04}$	$\underline{85.68 \pm 0.24}$
CFA [65]	84.38 ± 0.87	$\underline{96.33 \pm 0.14}$	$\underline{83.78 \pm 0.51}$
CFLOW [61]	74.82 ± 5.32	93.76 ± 0.67	75.45 ± 3.53
CS-Flow [59]	$\underline{90.91 \pm 0.83}$	NA	NA

Table 3.8: Comparison of anomaly detection performance on brain MRI benchmark. We including sample-level and pixel-level results. We report the mean and standard deviation over 5 random seeds for each measurement. Bold indicates the best performance.

Methods	RESC		
	Image AUROC	**Pixel AUROC**	**Pixel Pro**
f-AnoGAN [39]	77.42 ± 0.85	NA	NA
GANomaly [38]	52.56 ± 3.95	NA	NA
DRAEM [50]	83.22 ± 8.21	86.79 ± 3.14	63.55 ± 4.62
UTRAD[13]	89.39 ± 1.92	94.54 ± 1.24	77.49 ± 4.30
DeepSVDD [55]	74.17 ± 1.29	NA	NA
CutPaste [72]	90.23 ± 0.61	NA	NA
SimpleNet [67]	76.15 ± 7.46	77.14 ± 4.76	49.07 ± 5.23
MKD [11]	87.77 ± 0.87	96.18 ± 0.15	85.62 ± 0.47
RD4AD[70]	87.77 ± 0.87	96.18 ± 0.15	85.62 ± 0.47
STFPM[69]	84.82 ± 0.50	94.68 ± 0.57	81.27 ± 1.49
PaDiM [63]	75.87 ± 0.54	91.44 ± 0.42	71.68 ± 0.81
PatchCore[64]	91.55 ± 0.10	96.48 ± 0.10	85.84 ± 0.25
CFA [65]	69.90 ± 0.26	91.10 ± 0.87	69.77 ± 0.41
CFLOW [61]	74.95 ± 5.81	93.78 ± 0.57	76.80 ± 1.72
CS-Flow [59]	87.34 ± 0.58	NA	NA

Table 3.9: Comparison of anomaly detection performance on retinal OCT benchmark. We including sample-level and pixel-level results. We report the mean and standard deviation over 5 random seeds for each measurement. Bold indicates the best performance.

Methods	OCT207	RSNA	Camelyon16
	Image AUROC	**Image AUROC**	**Image AUROC**
f-AnoGAN [39]	73.42 ± 1.85	55.15 ± 0.09	69.49 ± 1.98
GANomaly [38]	70.47 ± 9.98	62.90 ± 0.65	54.44 ± 2.57
DRAEM [50]	88.03 ± 8.36	67.70 ± 1.72	52.35 ± 0.77
UTRAD[13]	96.78 ± 0.56	75.64 ± 1.24	69.96 ± 4.64
DeepSVDD [55]	76.76 ± 1.37	64.48 ± 3.17	60.98 ± 1.82
CutPaste [72]	96.76 ± 0.62	$\underline{82.61 \pm 1.22}$	$\underline{75.18 \pm 0.41}$
SimpleNet [67]	94.68 ± 2.17	69.12 ± 1.27	62.38 ± 3.71
MKD [11]	96.74 ± 0.26	$\underline{82.01 \pm 0.12}$	$\underline{77.54 \pm 0.27}$
RD4AD[70]	$\underline{97.30 \pm 0.79}$	67.63 ± 1.11	66.81 ± 0.71
STFPM[69]	96.76 ± 0.23	72.93 ± 1.96	66.36 ± 1.01
PaDiM [63]	91.75 ± 0.96	77.49 ± 1.87	67.25 ± 0.32
PatchCore[64]	$\underline{98.57 \pm 0.03}$	76.14 ± 0.67	$\underline{69.34 \pm 0.21}$
CFA [65]	79.47 ± 0.56	66.83 ± 0.23	65.64 ± 0.10
CFLOW [61]	85.35 ± 2.11	71.53 ± 1.49	55.66 ± 1.97
CS-Flow [59]	$\underline{98.47 \pm 0.28}$	$\underline{83.20 \pm 0.46}$	68.38 ± 0.42

Table 3.10: Comparison of anomaly detection performance on retinal OCT benchmark, chest x-ray benchmark and digital histopathology benchmark. We including only sample-level results. We report the mean and standard deviation over 5 random seeds for each measurement. Bold indicates the best performance.

Benchmarks	BraTS2021	BTCV + LiTs	RESC
DRAEM [50]	19.31 ± 5.52	9.38 ± 0.78	33.51 ± 3.52
UTRAD [13]	7.27 ± 0.06	2.33 ± 0.06	22.81 ± 0.36
MKD [11]	28.89 ± 0.72	$\underline{14.92 \pm 0.23}$	43.53 ± 1.10
RD4AD [70]	28.28 ± 0.48	10.72 ± 2.50	33.51 ± 3.52
STFPM [69]	25.40 ± 0.82	8.87 ± 2.52	49.23 ± 0.23
PaDiM [63]	25.84 ± 1.20	4.50 ± 0.46	38.30 ± 0.89
PatchCore [64]	$\underline{32.82 \pm 0.59}$	10.49 ± 0.23	$\underline{57.04 \pm 0.21}$
CFA [65]	30.22 ± 0.32	$\underline{14.93 \pm 0.08}$	36.57 ± 0.18
CFLOW [61]	19.50 ± 2.73	7.58 ± 3.16	44.83 ± 1.78
SimpleNet [67]	28.96 ± 1.73	12.26 ± 2.41	30.28 ± 1.64

Table 3.11: 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.

Chapter 4

A Language-Enhanced Reconstruction Model for Medical Anomaly Detection

4.1 Introduction

Upon establishing the first medical anomaly detection benchmark, BMAD, we found that no single existing model possesses the universality required to comprehensively address detection tasks across various domains. These models may perform well on specific datasets but often lack the generalizability needed to adapt to different medical data environments and detection requirements. We argue that this observation is attributed to the significantly diverse gaps between medical domains, which poses challenges to conventional representations learning.

Inspired by the powerful zero-shot capability and impressive transfer ability of large vison-language foundation models like CLIP, we aim to leverage such multi-modal models to better represent the medical information in images to overcome this limitation,thereby assisting the model in performing anomaly detection and localization tasks more effectively. By incorporating specialized medical knowledge, prior information, and data characteristics, we can guide the model to more accurately identify and delineate the contours of normal biomedical signals. This enhanced representational capability enables the model to more effectively recognize abnormal

patterns that deviate from the normal mode.

In this chapter, we will elaborate on how we construct a model framework that integrates additional knowledge, as well as how we use this framework to improve the accuracy and robustness of medical anomaly detection. We will discuss data preprocessing steps, feature engineering strategies, and the process of model training and validation. Furthermore, we will present experimental results and conduct a comparative analysis of the performance of different models on multiple medical datasets to demonstrate the effectiveness and versatility of our proposed methods. Through these efforts, we aim to advance the field of medical anomaly detection and provide stronger technical support for healthcare.

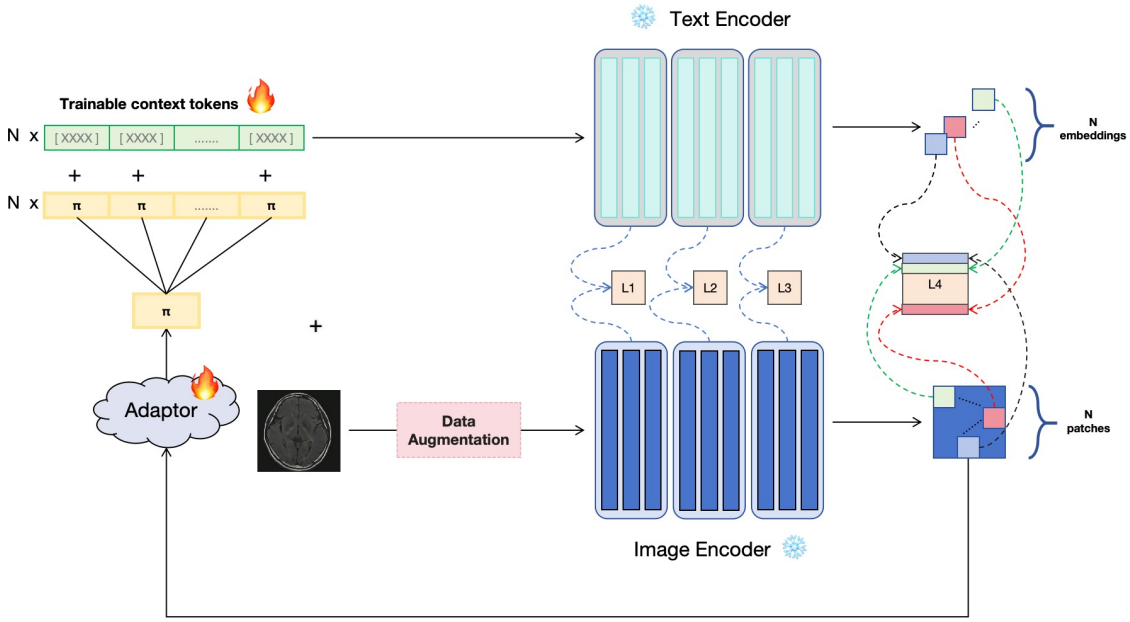

Figure 4.1: Diagram of the proposed network.

4.2 Methods and Procedure

4.2.1 Architecture Overview

Fig 4.1 illustrates the overall workflow of the proposed network, which can be roughly divided into two parts: Data Augmentation module and Language-Enhanced Reconstruction module. For an input image I, it first undergoes detail enhancement through

the Data Augmentation module. Then, the enhanced image is fed into the CLIP's frozen image encoder E_{image} to obtain the output feature f^{last} from the last layer. This feature f^{last} is subsequently input into a trainable adapter, which converts the image information into auxiliary prompt information that is comprehensible by text. The transformed information is then appended to a trainable blank prompt. The newly obtained prompt is then input into the CLIP's frozen text encoder E_{text} to acquire the encoded text features, which is then reconstructed with the previously obtained multi-dimensional image feature. Here, we employ cosine similarity as a constraint for this reconstruction process. Eventually, we obtain a set of trained prompts that describe the characteristic information of normal images, as well as a trained adapter that effectively assists the image encoder E_{image} in adapting to various medical domains.

It is noteworthy that the trainable components in the proposed method are the adaptor after the image encoder and the prompting tokens concatenating to visual tokens. The motivation behind this design is to bridge the domain gap between CLIP's pretrained dataset and the target data. Specifically, CLIP is pretrained on huge numbers of paired image-text data collected from internet, among which a large portion are about natural images. Thus, CLIP is powerful to represent the rich and diverse knowledge associated to natural images using text tokens. However, medical images from different domains exhibit different medical content, which poses challenges to the pretrained CLIP model for pairing the text encoder and image encoder. By leveraging trainable adaptor and text tokens on the target data (i.e. medical images in our study), we aim to bridge the domain gap and inject extra domain-specific pairing information to the CLIP model. Since the trainable components are adapted to the targeted medical domain, the model is flexible to tailor for different AD tasks.

4.2.2 Language-Enhanced Reconstruction

We trained the model exclusively on anomaly-free samples, so that the image branch and text branch are re-pairing for normal data, but show large pairing errors (or, reconstruction errors) for abnormal samples. In this study, we used E_{text} and E_{image} as our text encoder and image encoder in the pre-trained CLIP, respectively. The adaptor after the image encoder is parameterized by θ, denoted by h_θ, and the set of N trainable numerical text prompts are represented by T.

$$\boldsymbol{T_i} = [\mathrm{t}]_1[\mathrm{t}]_2 \dots [\mathrm{t}]_M, \tag{4.1}$$

where i represents the i-th blank prompt and M represents the number of the trainable tokens contained in one prompt T_i. M is a hyper-parameter and we keep the same M for all N prompts. The rationale underlying our N trainable text prompts stems from the process whereby, subsequent to an input image traversing the image encoder E_{image}, the resulting patch features undergo augmentation with position embedding. Correspondingly, we target to train an individual numerical prompt for each patch location. This integration introduces subtle distinctions, even among patches that exhibit similarity in their original patterns, thereby enriching the representation by the spatial information and ensuring the uniqueness of each patch embedding.

Concisely, E_{image} processes the input image I to yield $f_i^k = E_{image}^k(I)$, where f_i^k signifies the i-th feature from the k-th layer of E_{image}. To further enhance the capability of text branch in reconstructing normative patterns with prompts T and bridging the gap between the pre-trained CLIP encoders and target medical domain, we devised a tailored adapter module for style transfer learning. This module serves as a refinement tool to fine-tune the image feature after E_{image}, enabling it to better capture the intricacies of medical imagery. Here, we introduce $\boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{f_i^{last}})$ as the feature after the adapter, which updates each visual context token. Here, f_i^{last} denotes the distinctive feature output from the culminating layer of the E_{image} model. This architecture design ensures that the image encoder E_{image} is specifically adapted

to the target medical images. Meanwhile, the E_{text} model processes the i-th input P_i in an augmented form, incorporating the adapter's output:

$$p_i^k = E_{\text{text}}^k(T_i + h_{\boldsymbol{\theta}}(\boldsymbol{f_i^{last}})), \tag{4.2}$$

where p_i^k represents the text feature corresponding to the i^{th} patch extracted from the k-th layer of E_{text}, thereby integrating the visual nuances of medical images into the textual prompts.

Training objective: During training, the whole model concurrently update the textual trainable prompts $\boldsymbol{T_i}$ alongside the parameters $\boldsymbol{\theta}$ of the adapter. Our adapter module incorporates a two-layer bottleneck architecture, specifically Linear-ReLU-Linear, where the hidden layer effectively compresses the input dimension by a factor of 16×. To encourage the pairing and alignment of representations from the text branch and image branch, the cosine similarity between f_i^k and p_i^k is minimized with normal input data only. The obtained 2-D anomaly map $M^k \in \mathbb{R}^{H_k \times W_k}$ is calculate from the previous $f^k, p^k \in \mathbb{R}^{C_k \times H_k \times W_k}$, where C_k, H_k and W_k denote the number of channels, height and width of the k^{th} layer activation tensor.

$$M_i^k(h,w) = 1 - \frac{\left(f_i^k(h,w)\right)^T \cdot p_i^k(h,w)}{\left\|f_i^k(h,w)\right\| \left\|p_i^k(h,w)\right\|}, \tag{4.3}$$

Then the complete set of feature maps from one layer are accumulated:

$$M_{\text{overall}}^k = \sum_{i=1}^{C_K} M_i^k. \tag{4.4}$$

A final objective loss to optimize the multimodal model combines the multi-scale anomaly scores as follows.

$$\mathcal{L} = \sum_{k=1}^{K} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M_{\text{overall}}^k(h,w) \right\}, \tag{4.5}$$

where K indicates the number of feature layers used in the experiment.

Inference: At the inference stage, after obtaining a set of anomaly maps from the image-prompt pairs, we up-samples M^k to the original image size by a bilinear

up-sampling operation Ψ. Then a precise score map S is formulated as the pixel-wise accumulation of all anomaly maps, $S = \sum_{i=1}^{L} \Psi(M^i)$. Then the maximal value in S is taken as the sample-level anomaly score.

4.2.3 Data Augmentation

As in prior work, P-Net[12], the authors show that incorporating the structure information into the whole reconstruction network boosts medical anomaly detection. By training an additional segmentation network in advance, P-Net generates an retina OCT vessel structural images for each query picture and incorporates them into the OCT anomaly detection network. Despite using the pre-trained segmentation network, as P-net did, we found that the existing segmentation networks in medicine do not possess sufficient generalization performance to adapt to data from varying medical domains. Failing in finding an universal pre-trained model for all medical domains, we opt to enhance the image histological information in a simple way as shown in Fig. 4.3. Essentially, our training-free data augmentation leverages the well-known Laplacian pyramid in Fig. 4.2 to enhance the low-frequency information in medical images and removes high-frequency noise from an original image.

To get the low-frequency residual image, we can formulate Laplacian Pyramid by first apply a Gaussian Pyramid. We define K to represent the level of our Gaussian operations and I is a single image. For each I, it has a set of band-pass images:

$$G(\boldsymbol{I}) = [\boldsymbol{I}_1, \ldots, \boldsymbol{I}_K] \tag{4.6}$$

Where $\boldsymbol{I}$ should be $\boldsymbol{I_0}$. The Gaussian pyramid is essentially a downsampling process that convolves an image with a Gaussian kernel, denoted by $F_{\downarrow}(\cdot)$ here. With downsampleing, the size of the image will be decimated by:

$$\boldsymbol{I}_{k+1} = F_{\downarrow}(\boldsymbol{I}_k) \tag{4.7}$$

Where a $j \times j$ input image $\boldsymbol{I_K}$ will be blurred and decimated to $\boldsymbol{I_{K+1}}$ with size of $j/2 \times j/2$. For low-frequency restoration to the original image dimension, the

low-freuqnecy residue are upsampled by K-times,

$$I_k = F_{\uparrow}\left(\boldsymbol{I}_{k+1}\right) + L_k, \tag{4.8}$$

where $\boldsymbol{I_K}$ can be seen as a invertible combination of different resolutions,$F_{\uparrow}\left(\boldsymbol{I}_{k+1}\right)$ and L_k. L_k is well known as Laplacian Pyramid. And for $F_{\uparrow}\left(\boldsymbol{I}_{k+1}\right)$, it is the low-resolution residual image which contains the key structure information.

We define the target of our proposed low-resolution residual image as S_K, which K indicates the level of Gaussian operations and also represents the number of times we need to upsampling. So our target structure information is computed as follows:

$$S_k = F_{\uparrow}\left(F_{\downarrow}\left(\boldsymbol{I}_k\right)\right) = F_{\uparrow K}\left(F_{\downarrow K}\left(\boldsymbol{I}\right)\right) \tag{4.9}$$

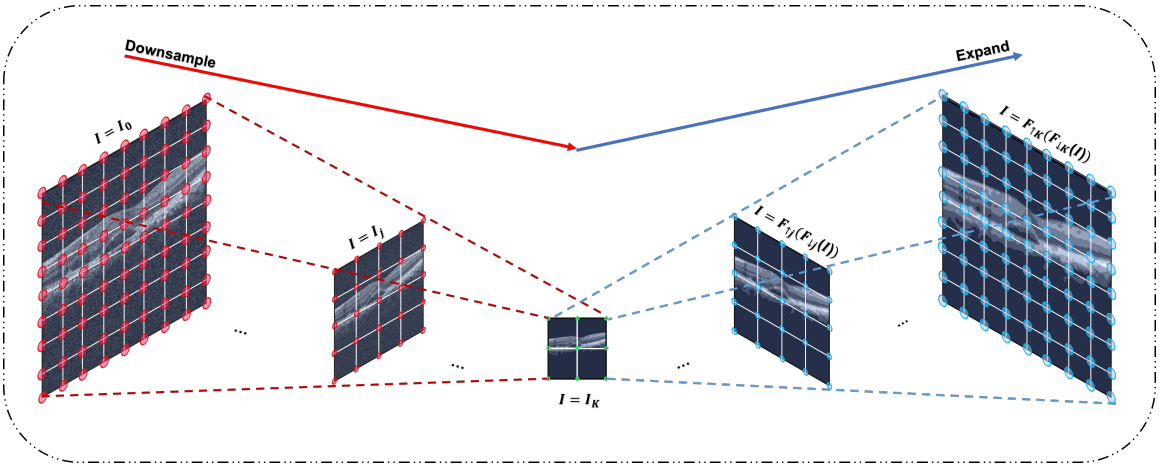

Figure 4.2: Diagram Illustrating the Process of Downsampling and Upsampling. Use retinal image as a example.

Finally, the obtained structure information is added to the original image for an enhanced image version for our multimodal analysis.

4.3 Results and Discussion

4.3.1 Dataset

We conduct medical anomaly detection and localization experiments on previous banchmark work which contains seven datasets from six different domains: BraTS2021[89],

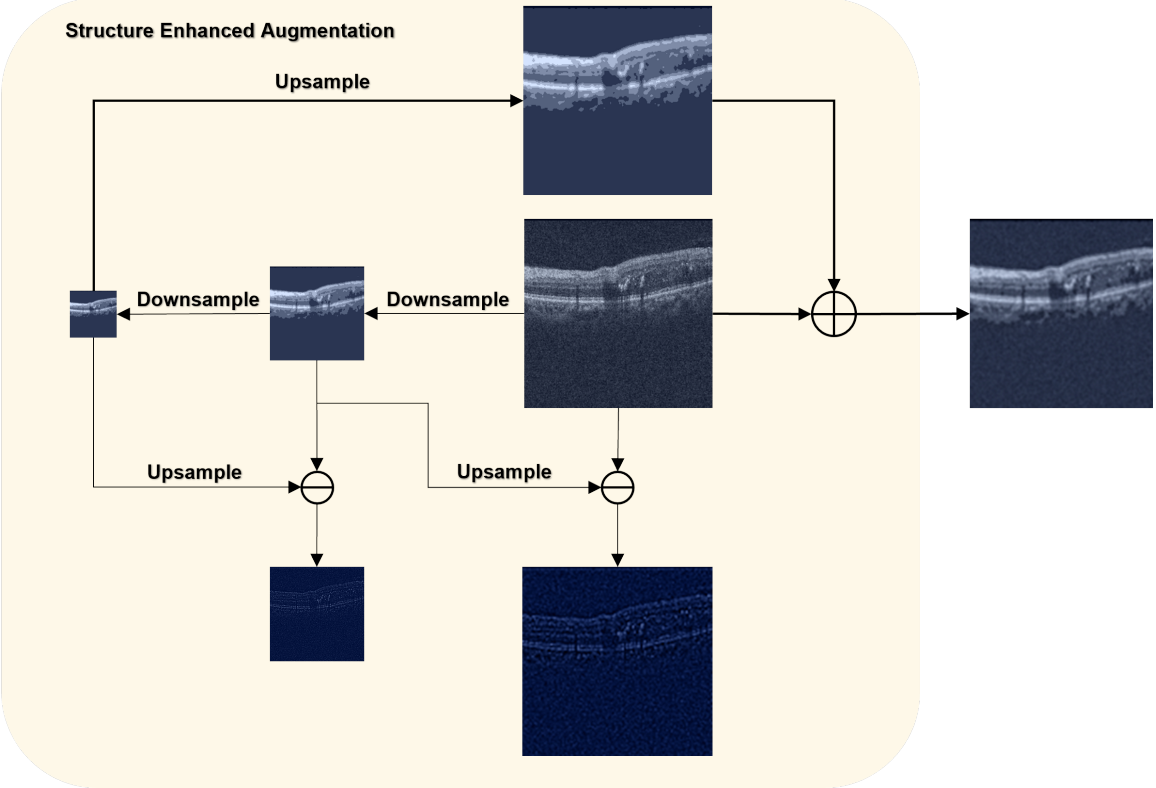

Figure 4.3: Diagram of the our image agumentation process.

BTCV[90], LiTs[91], RESC[92], RSNA[8], Camelyon16[93], OCT2017[94]. We have already give a detail intorduction in Table 3.1. In this work, we used the training data for language reconstruction. The test data in the mentioned benchmarks are used for evaluating anomaly detection and localization.

4.3.2 Metrics

We used three methods for evaluation. In accordance with established practices in previous literature, we employ AUROC (Area Under the Receiver Operating Characteristic Curve), and PRO (Per-Region Overlap) metrics to assess the performance of anomaly detection.

4.3.3 Implementation

We adopt ViT-B-16+[104] as the visual encoder and the transformer [105] as the text encoder by default from the public pre-trained CLIP model. We also provide

results under ResNet [106]. we set the image size to 256 × 256. And we added a extra medical image augmentation module. For the text encoder, we use 16 empty tokens for reconstruction. For traning stage, we use the AdamW optimizer and set the learning rate to 0.001, and the adaptor is optimized with 400 epochs. We report the mean and variance of the results of MedCLIP over 5 random seeds.

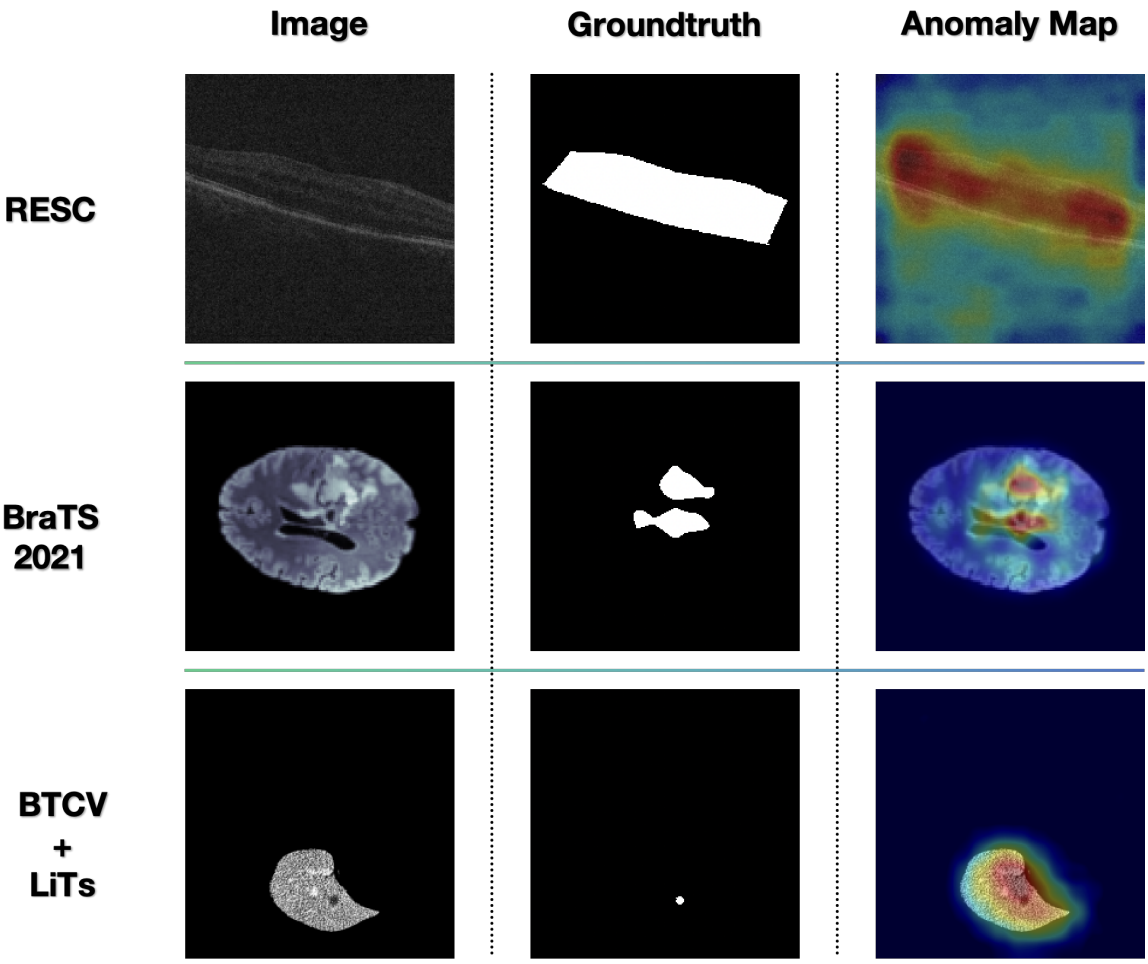

Figure 4.4: In this diagram,we present the anomaly localization results of our model.

4.3.4 Performance and Ablation Studies

We present out results in Table 4.1, Table 4.2, Table 4.3 and Table 4.4 present the performance of our results of anomaly localization and detection on different medical domains and datasets, respectively. We compare our proposed methods with prior vision-language based works for medical AD. We also conducted ablation experiments

simultaneously, experimenting with different setups of vision encoders: ResNet-18, ResNet-50, and ViT, respectively. Regarding the DA module, we performed ablation experiments as well. The experimental results demonstrated that the combination of ViT and DA configuration could enable our model to achieve the best performance. Additionally, we conducted ablation experiments on the number of the reconstructed tokens, setting them to 8, 16, 24, and 32, respectively. The model achieved the best performance at 16. We also conducted ablation experiments on the value of k mentioned in Equation 4.2. The experimental results, showing in Fig.4.8 demonstrate that our k should be set to a combination of four layers, specifically 1, 2, 3, and 4. Our method can also effectively locate the position of abnormal detection. As for abnormal areas, our method is more sensitive and can achieve better boundary as shown in Fig 4.4.

At the beginning of this chapter, we raised a challenging issue, which is that, according to Chapter 3, no existing model can perfectly address the problem of anomaly detection in various medical domains, as shown in Fig 4.5. They often overlook the rich semantic information contained in medical images as well as the unique imaging modalities of medical images. To avoid excessive manual annotation, which is often the most costly aspect in medicine, we chose to use blank tokens to reconstruct the multi-dimensional information contained in normal images. By adopting a text-description perspective, we aimed to restore finer-grained normal information and identify the location of anomalies through differences detected during the inference stage. Our results achieved the best performance among all existing methods.

We conducted ablation experiments on the hyper-parameters M in Fig 4.6 and K in Fig 4.8. Specifically investigating the influence of the length of trainable tokens on the reconstruction performance. The results indicated that when M was set to 16, the reconstruction effect was optimal, leading to the best performance in medical anomaly detection task. Additionally, we performed ablation experiments on the hyper-parameter K for our selected multi-layer anomaly map. The conclusion

Methods	BraTS2021		
	Image AUROC	**Pixel AUROC**	**Pixel Pro**
WinCLIP [39]	86.76 ± 0.31	94.72 ± 0.26	76.46 ± 2.86
CLIP [38]	64.43 ± 2.67	76.14 ± 1.79	54.57 ± 5.92
DDPM [16]	89.74 ± 3.12	96.14 ± 0.69	86.16 ± 3.24
Ours(ResNet-18 based with DA)	87.23 ± 1.34	82.54 ± 0.45	75.60 ± 2.28
Ours(ResNet-50 based with DA)	89.68 ± 2.12	87.73 ± 0.67	78.46 ± 2.32
Ours(ViT based without DA)	91.12 ± 0.84	92.40 ± 2.76	80.36 ± 1.83
Ours(ViT based with DA)	92.74 ± 0.84	97.01 ± 0.46	85.71 ± 0.58

Table 4.1: We compare our results on Brain MRI benchmark with prior works and our work with different settings.

Methods	BTCV + LiTs		
	Image AUROC	**Pixel AUROC**	**Pixel Pro**
WinCLIP [39]	72.48 ± 2.38	96.12 ± 0.48	86.26 ± 2.26
CLIP [38]	57.81 ± 3.58	91.64 ± 1.64	74.37 ± 2.32
DDPM [16]	90.92 ± 2.68	92.32 ± 2.70	74.65 ± 0.73
Ours(ResNet-18 based with DA)	65.86 ± 2.92	90.14 ± 1.48	85.63 ± 1.28
Ours(ResNet-50 based with DA)	71.28 ± 4.10	91.62 ± 2.25	90.24 ± 3.38
Ours(ViT based without DA)	70.32 ± 0.24	92.40 ± 2.76	88.16 ± 2.83
Ours(ViT based with DA)	74.94 ± 1.34	98.84 ± 0.46	69.60 ± 1.74

Table 4.2: We compare our results on Liver CT benchmark with prior works and our work with different settings.

Methods	RESC		
	Image AUROC	**Pixel AUROC**	**Pixel Pro**
WinCLIP [39]	84.45 ± 0.21	86.79 ± 3.14	74.38 ± 0.42
CLIP [38]	74.61 ± 3.06	72.18 ± 1.43	45.01 ± 0.34
DDPM [16]	64.43 ± 2.67	76.14 ± 1.79	54.57 ± 5.92
Ours(ResNet-18 based with DA)	80.12 ± 2.67	90.32 ± 3.15	79.34 ± 1.82
Ours(ResNet-50 based with DA)	89.24 ± 1.53	94.54 ± 1.24	80.14 ± 0.72
Ours(ViT based without DA)	87.72 ± 1.34	90.42 ± 1.73	76.74 ± 2.16
Ours(ViT based with DA)	93.98 ± 2.46	95.40 ± 1.26	78.64 ± 3.64

Table 4.3: We compare our results on Retinal OCT benchmark with prior works and our work with different settings.

Methods	OCT2017	RSNA	Camelyon16
	Image AUROC	**Image AUROC**	**Image AUROC**
WinCLIP [39]	94.36 ± 0.78	79.72 ± 2.52	70.14 ± 2.08
CLIP [38]	86.62 ± 1.28	66.82 ± 1.34	64.34 ± 1.22
DDPM [16]	95.62 ± 2.34	76.14 ± 1.79	64.72 ± 3.64
Ours(ResNet-18 based with DA)	90.47 ± 1.71	74.54 ± 1.11	64.78 ± 0.23
Ours(ResNet-50 based with DA)	96.76 ± 0.62	80.23 ± 2.62	75.18 ± 0.41
Ours(ViT based without DA)	94.34 ± 2.36	84.36 ± 3.62	72.14 ± 0.96
Ours(ViT based with DA)	98.94 ± 2.46	85.70 ± 2.60	77.86 ± 1.18

Table 4.4: We compare our results on Retinal OCT, Chest X-ray and Histopathology benchmark with prior works and our work with different settings.

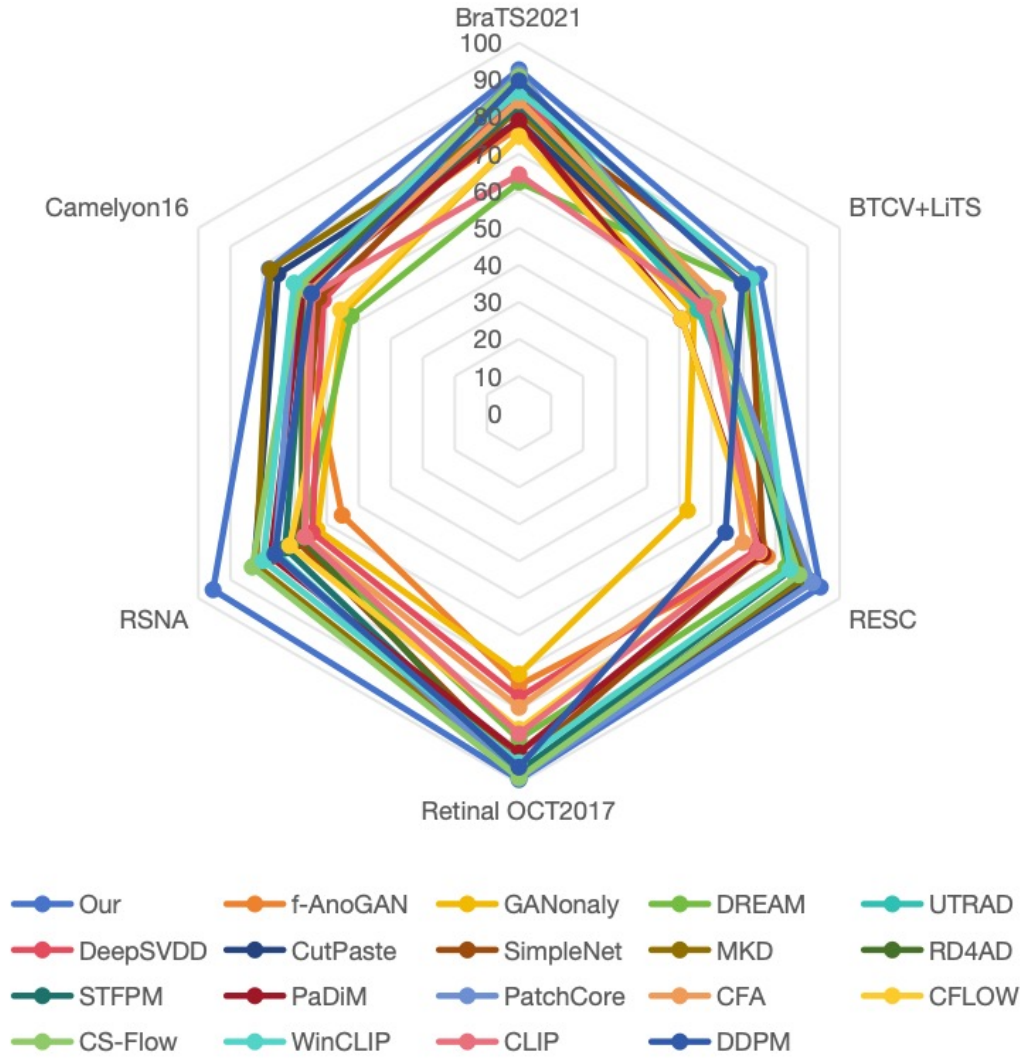

Figure 4.5: In this diagram,we present the performance outcomes of our replicated experimental results on established medical benchmarks. While our model did not attain exceptional performance in all evaluated metrics, it consistently achieved the relatively optimal results across various domains, underscoring its potential for generalization within the medical domain.

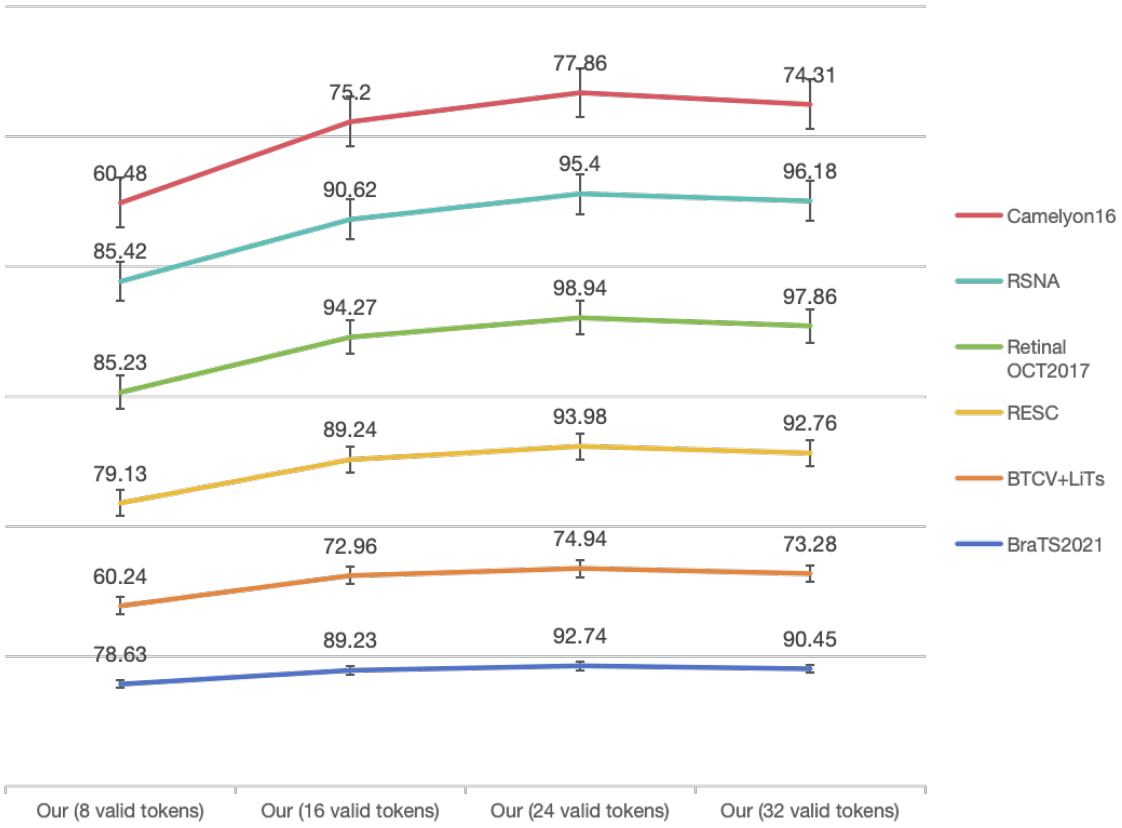

Figure 4.6: In this diagram, we showcase the anomaly detection outcomes stemming from our ablation experiments, specifically targeting the hyper-parameters related to the quantity of tokens M.

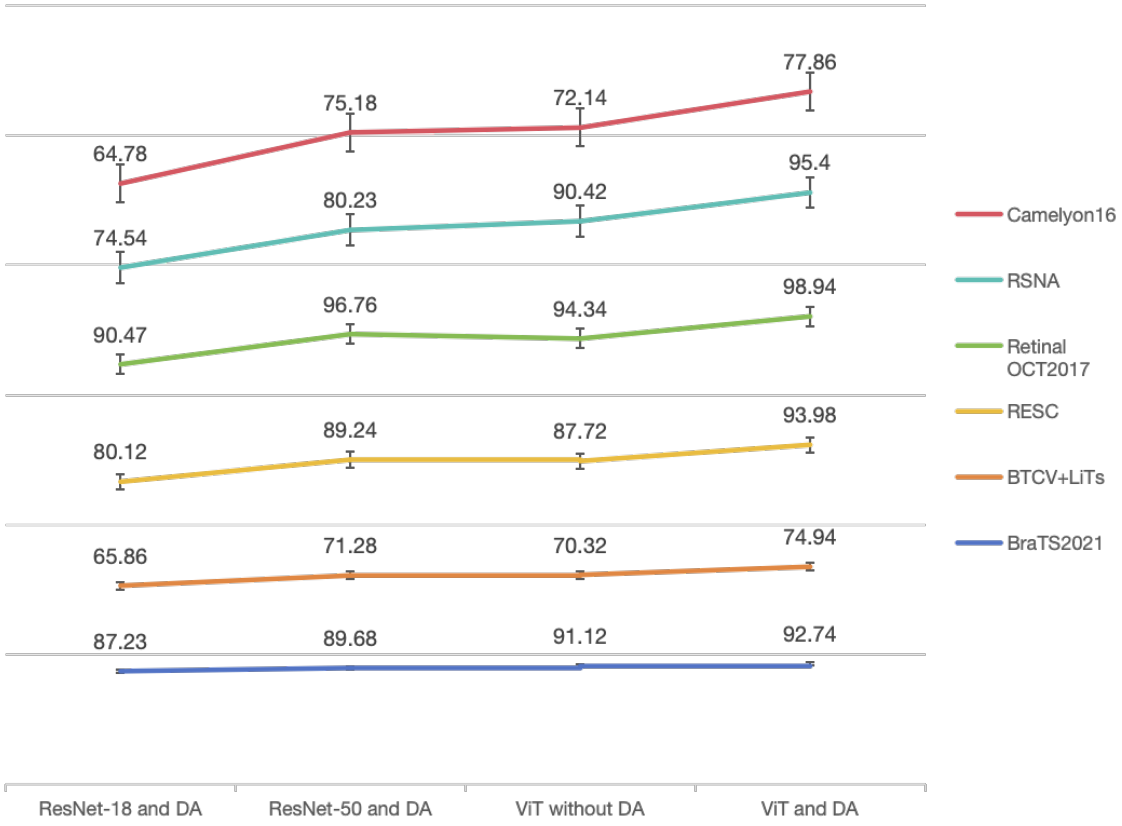

Figure 4.7: In this diagram,we present the anomaly detection results of our ablation experiments on different module combinations.

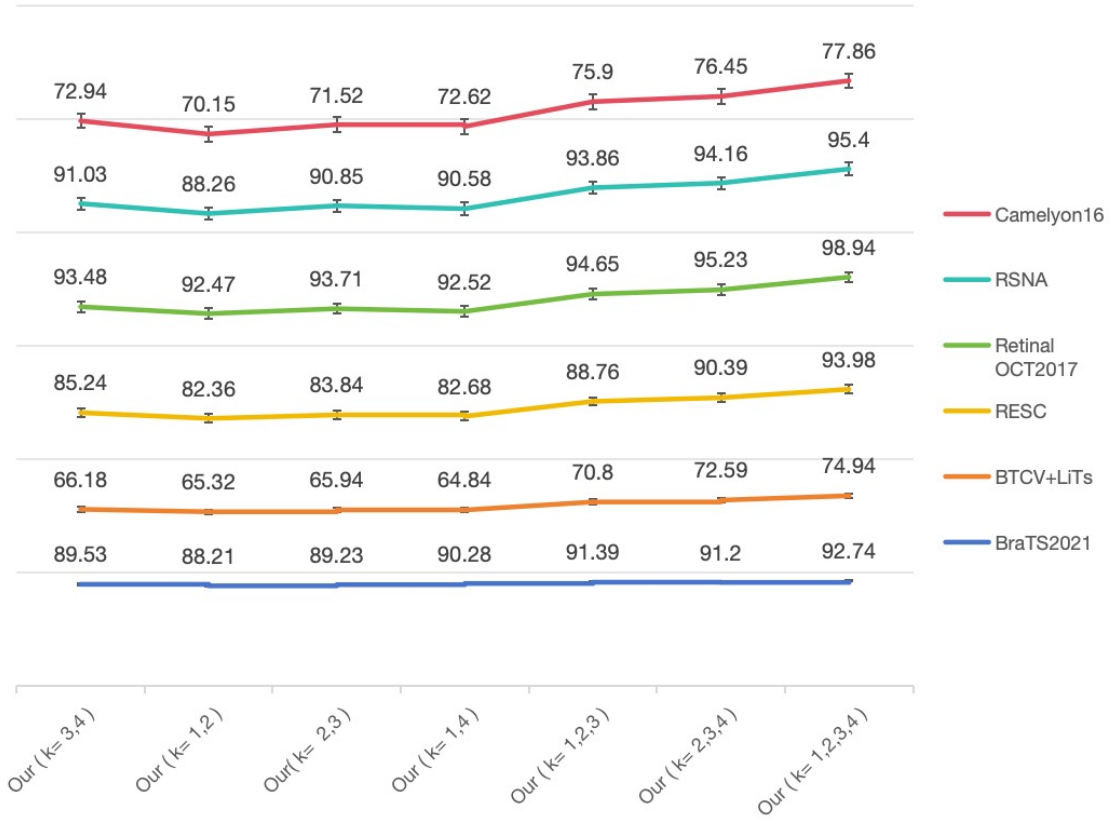

Figure 4.8: In this diagram,we present the anomaly detection results of our ablation experiments on hyper-parameters K.

revealed that when K was set to include the first, second, third, and fourth layers in superposition, the model achieved optimal performance. We also discussed the roles of our various modules in Fig4.7 to demonstrate that the final design enables the model to achieve optimal performance.

Chapter 5

Conclusions, Recommendations, & Future Work

5.1 Conclusion

In this work, we have presented two significant contributions to the field of medical anomaly detection. Firstly, we have introduced a comprehensive medical anomaly detection benchmark, comprising six diverse datasets across five key medical domains and integrating 15 state-of-the-art anomaly detection algorithms. This benchmark not only represents the most extensive collection to date but also provides a rigorous evaluation framework for medical anomaly detection algorithms, enabling thorough assessment from multiple perspectives. Secondly, we proposed a novel multimodal approach leveraging CLIP for unified anomaly detection and localization on medical benchmarks. Our method achieves state-of-the-art performance, demonstrating the potential of vision-language models to address anomaly detection tasks beyond traditional training data constraints.

5.2 Limitation

5.2.1 Data Bias and Representativeness

The medical datasets used in our benchmark are predominantly collected from advanced countries, which may introduce inherent geographical and sampling biases.

This limitation can affect the generalizability of the evaluation results to other populations and regions.

5.2.2 Hyper-parameter Optimization

While we carefully adhered to the hyper-parameter settings proposed in the original works for the evaluated algorithms, it is possible that some hyper-parameters did not achieve their optimal values for specific datasets in our experiments. This could limit the full potential of some algorithms.

5.2.3 Evaluation Framework

Our benchmark currently follows the one-for-one anomaly detection paradigm, where a separate model is trained for each subject or class. However, recent research has shown the advantages of unified one-for-N models that can handle multiple classes with a single model. Evaluating such unified models on our benchmark could provide further insights.

5.3 Remark and Future Work

To mitigate the limitations of data bias, we plan to expand the benchmark by including datasets from a wider range of geographical locations and demographic groups. This will enhance the representativeness and generalizability of the evaluation results. Meanwhile, we will develop more systematic approaches to optimize hyper-parameters for each algorithm and dataset combination. This could involve techniques such as grid search, random search, or Bayesian optimization to find the optimal settings that maximize performance.

Inspired by the success of our multimodal model for generalized anomaly detection cross multi-domains, we will further explore the potential of vision-language integration for medical anomaly detection and localization. This may involve developing new prompting techniques, incorporating domain-specific knowledge, or leveraging

advances in vision-language pre-training models.

Bibliography

[1] G. Xie *et al.*, "Im-iad: Industrial image anomaly detection benchmark in manufacturing," *arXiv preprint arXiv:2301.13359*, 2023.

[2] Y. Zheng, X. Wang, Y. Qi, W. Li, and L. Wu, "Benchmarking unsupervised anomaly detection and localization," *arXiv preprint arXiv:2205.14852*, 2022.

[3] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 142–32 159, 2022.

[4] J. Yang *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in neural information processing systems,Track on Datasets and Benchmarks*, 2022.

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

[6] K. Zhou *et al.*, "Sparse-gan: Sparsity-constrained generative adversarial network for anomaly detection in retinal oct image," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2020, pp. 1227–1231.

[7] J. Zhang *et al.*, "Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 879–890, 2020.

[8] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

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

[10] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Fusing unsupervised and supervised deep learning for white matter lesion segmentation," in *International Conference on Medical Imaging with Deep Learning*, PMLR, 2019, pp. 63–72.

[11] 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.

[12] K. Zhou *et al.*, "Encoding structure-texture relation with p-net for anomaly detection in retinal images," in *European conference on computer vision*, Springer, 2020, pp. 360–377.

[13] L. Chen, Z. You, N. Zhang, J. Xi, and X. Le, "Utrad: Anomaly detection and localization with u-transformer," *Neural Networks*, vol. 147, pp. 53–62, 2022.

[14] W. H. L. Pinaya *et al.*, "Unsupervised brain anomaly detection and segmentation with transformers," *arXiv preprint arXiv:2102.11650*, 2021.

[15] H. M. Rai, K. Chatterjee, and S. Dashkevich, "Automatic and accurate abnormality detection from brain mr images using a novel hybrid unetresnext-50 deep cnn model," *Biomedical Signal Processing and Control*, vol. 66, p. 102 477, 2021.

[16] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2022, pp. 35–45.

[17] D. Chen, X. Shao, B. Hu, and Q. Su, "Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra," *Analytical Sciences*, vol. 21, no. 2, pp. 161–166, 2005.

[18] P. H. Torr and D. W. Murray, "Outlier detection and motion segmentation," in *Sensor Fusion VI*, SPIE, vol. 2059, 1993, pp. 432–443.

[19] S. E. Guttormsson, R. J. Marks, M. A. El-Sharkawi, and I Kerszenbaum, "Elliptical novelty grouping for on-line short-turn detection of excited running rotors," *IEEE Transactions on Energy Conversion*, vol. 14, no. 1, pp. 16–22, 1999.

[20] E. J. Keogh, P. Smyth, *et al.*, "A probabilistic approach to fast pattern matching in time series databases.," in *Kdd*, vol. 1997, 1997, pp. 24–30.

[21] M. Desforges, P. Jacob, and J. Cooper, "Applications of probability density estimation to the detection of abnormal conditions in engineering," *Proceedings of the institution of mechanical engineers, part c: Journal of mechanical engineering science*, vol. 212, no. 8, pp. 687–703, 1998.

[22] S. Byers and A. E. Raftery, "Nearest-neighbor clutter removal for estimating features in spatial point processes," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 577–584, 1998.

[23] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *2007 IEEE symposium on computational intelligence and data mining*, IEEE, 2007, pp. 504–515.

[24] S. Singh and M. Markou, "An approach to novelty detection applied to the classification of image regions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 396–407, 2004.

[25] C. P. Diehl and J. B. Hampshire, "Real-time object classification and novelty detection for collaborative video surveillance," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 3, 2002, pp. 2620–2625.

[26] Q. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, vol. 32, no. 4, pp. 440–448, 2002.

[27] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, 2014, pp. 4–11.

[28] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 665–674.

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

[30] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3379–3388.

[31] X. Li, M. Radulovic, K. Kanjer, and K. N. Plataniotis, "Discriminative pattern mining for breast cancer histopathology image classification via fully convolutional autoencoder," *IEEE Access*, vol. 7, pp. 36 433–36 445, 2019.

[32] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.

[33] A. Kascenas, N. Pugeault, and A. Q. O'Neil, "Denoising autoencoders for unsupervised anomaly detection in brain mri," in *Medical Imaging with Deep Learning*, 2022.

[34] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 372–14 381.

[35] J. Hou, Y. Zhang, Q. Zhong, D. Xie, S. Pu, and H. Zhou, "Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8791–8800.

[36] K. Zhou *et al.*, "Memorizing structure-texture correspondence for image anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[37] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2898–2906.

[38] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*, Springer, 2018, pp. 622–637.

[39] 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.

[40] X. Yan, H. Zhang, X. Xu, X. Hu, and P.-A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 3110–3118.

[41] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *Proceedings of the IEEE/CVF Winter Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 650–656.

[42] Y. Teng, H. Li, F. Cai, M. Shao, and S. Xia, "Unsupervised visual defect detection with score-based generative model," *arXiv preprint arXiv:2211.16092*, 2022.

[43] Z. You, K. Yang, W. Luo, L. Cui, Y. Zheng, and X. Le, "Adtr: Anomaly detection transformer with feature reconstruction," in *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part III*, Springer, 2023, pp. 298–310.

[44] W. Jin, F. Guo, and L. Zhu, "Incremental self-supervised learning based on transformer for anomaly detection and localization," *arXiv preprint arXiv:2303.17354*, 2023.

[45] Y. Lee and P. Kang, "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," *IEEE Access*, vol. 10, pp. 46 717–46 724, 2022.

[46] V. L. Cao, M. Nicolau, and J. McDermott, "A hybrid autoencoder and density estimation model for anomaly detection," in *International Conference on Parallel Problem Solving from Nature*, Springer, 2016, pp. 717–726.

[47] V. L. Cao, M. Nicolau, and J. McDermott, "One-class classification for anomaly detection with kernel density estimation and genetic programming," in *European Conference on Genetic Programming*, Springer, 2016, pp. 3–18.

[48] Z. Li *et al.*, "Superpixel masking and inpainting for self-supervised anomaly detection," in *Bmvc*, 2020.

[49] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107 706, 2021.

[50] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8330–8339.

[51] H. Deng and X. Li, "Self-supervised anomaly detection with random-shape pseudo-outliers," in *International Conference of the IEEE Engineering in Medicine & Biology Society*, 2022.

[52] J. Tan, B. Hou, J. Battern, H. Qiu, and B. Kainz, "Detecting outliers with foreign patch interpolation," *Journal of Machine Learning for Biomedical Imaging*, vol. 13, 2022.

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

[54] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[55] L. Ruff *et al.*, "Deep one-class classification," in *International conference on machine learning*, PMLR, 2018, pp. 4393–4402.

[56] J. Yi and S. Yoon, "Patch svdd: Patch-level svdd for anomaly detection and segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[57] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*, PMLR, 2015, pp. 1530–1538.

[58] J. Yu *et al.*, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows," *arXiv preprint arXiv:2111.07677*, 2021.

[59] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1088–1097.

[60] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but differnet: Semi-supervised defect detection with normalizing flows," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1907–1916.

[61] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 98–107.

[62] N. Li, K. Jiang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Anomaly detection via self-organizing map," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 974–978.

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

[64] 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.

[65] 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.

[66] 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.

[67] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 402–20 411.

[68] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2592–2602.

[69] S. Yamada and K. Hotta, "Reconstruction student with attention for student-teacher pyramid matching," *arXiv preprint arXiv:2111.15376*, 2021.

[70] 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.

[71] R. Zhang *et al.*, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.

[72] 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.

[73] M. Shu *et al.*, "Test-time prompt tuning for zero-shot generalization in vision-language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 274–14 289, 2022.

[74] P. Gao *et al.*, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[75] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616.

[76] H. Deng, Z. Zhang, J. Bao, and X. Li, "Anovl: Adapting vision-language models for unified zero-shot anomaly localization," *arXiv preprint arXiv:2308.15939*, 2023.

[77] X. Li *et al.*, "Promptad: Learning prompts with only normal samples for few-shot anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 838–16 848.

[78] Y. Li *et al.*, "Myriad: Large multimodal model by applying vision experts for industrial anomaly detection," *arXiv preprint arXiv:2310.19070*, 2023.

[79] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

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

[81] S. N. Marimont and G. Tarroni, "Anomaly detection through latent space restoration using vector quantized variational autoencoders," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2021, pp. 1764–1767.

[82] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[83] M. M. R. Siddiquee *et al.*, "Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 191–200.

[84] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using wasserstein gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8309–8319.

[85] J. Wolleb, R. Sandkühler, and P. C. Cattin, "Descargan: Disease-specific anomaly detection with weak supervision," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, Springer, 2020, pp. 14–24.

[86] J. Pirnay and K. Chai, "Inpainting transformer for anomaly detection," in *International Conference on Image Analysis and Processing*, Springer, 2022, pp. 394–406.

[87] F. Meissen, G. Kaissis, and D. Rueckert, "Challenging current semi-supervised anomaly segmentation methods for brain mri," in *International MICCAI brainlesion workshop*, Springer, 2021, pp. 63–74.

[88] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[89] U. Baid *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[90] B. Landman, Z. Xu, J Igelsias, M. Styner, T Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, vol. 5, 2015, p. 12.

[91] P. Bilic *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[92] J. Hu, Y. Chen, and Z. Yi, "Automated segmentation of macular edema in oct using deep neural networks," *Medical image analysis*, vol. 55, pp. 216–227, 2019.

[93] B. E. Bejnordi *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[94] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[95] P. Bilic *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102 680, 2023.

[96] Y. Dey Raunak andAn Accurate Unsupervised Liver Hong, "Asc-net: Adversarial-based selective network for unsupervised anomaly segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 236–247.

[97] H. Li, Y. Iwamoto, X. Han, L. Lin, H. Hu, and Y.-W. Chen, "An accurate unsupervised liver lesion detection method using pseudo-lesions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 214–223.

[98] Y. Li and W. Ping, "Cancer metastasis detection with neural conditional random field," *arXiv preprint arXiv:1806.07064*, 2018.

[99] Y. Tian *et al.*, "Computer-aided detection of squamous carcinoma of the cervix in whole slide images," *arXiv preprint arXiv:1905.10959*, 2019.

[100] Y. He and X. Li, "Whole-slide-imaging cancer metastases detection and localization with limited tumorous data," in *Medical Imaging with Deep Learning*, 2023.

[101] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.

[102] T. D. Tien *et al.*, "Revisiting reverse distillation for anomaly detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023.

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

[104] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[105] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[106] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.