Enhancing Visual Anomaly Detection with Auxiliary Information

by

Zhaoxiang Zhang

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

Master of Science

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering

University of Alberta

Abstract

This thesis delves into the advancements in visual anomaly detection (AD), a challenging task in identifying outliers in images such as defects and lesions, which is crucial in many applications including medical diagnosis and industrial manufacturing. This thesis addresses two main challenges: increasing the detection accuracy in unsupervised medical tumor detection and enhancing the performance of zero-shot anomaly detection (ZSAD) models, both with the assistance of auxiliary data.

In the first part, the thesis focuses on unsupervised AD in medical imaging. It introduces a novel pseudo-anomaly synthesis module designed to generate diverse anomalies in shape and intensity for pseudo-supervised learning. This approach leads to a two-stage training strategy aimed at fostering a well-generalized model that significantly improves tumor segmentation performance.

In the second part, the thesis presents the Dual-Image Enhanced CLIP ZSAD model. This innovative approach merges visual and semantic data to refine anomaly classification and localization. By leveraging unlabeled visual references and implementing test-time adaptation with pseudo anomalies, the model achieves a notable improvement in detection accuracy, surpassing current leading methods.

These contributions significantly enhance both unsupervised medical tumor segmentation and ZSAD accuracy through auxiliary data. The introduction of random-shape synthesized anomalies and two-stage training strategy, serves as a foundational framework for refining the pseudo anomaly generation and training methodology. Furthermore, by exploring a vision-language model framework in anomaly detection, this research lays the groundwork for future advancements in the field. These findings underscore the demand for robust, adaptable solutions and set a promising trajectory for ongoing research in AD systems.

Preface

The thesis is conducted under the supervision of Professor X. Li. Chapter 3 has been published as Z. Zhang, H. Deng, X. Li, “Unsupervised Liver Tumor Segmentation with Pseudo Anomaly Synthesis”, *Simulation and Synthesis in Medical Imaging (SASHIMI) 2023*. Chapter 4 of the thesis is the original work of myself and has been submitted to *European Conference on Computer Vision (ECCV) 2024*, as Z. Zhang, H. Deng, J. Bao, and X. Li, “Dual-Image Enhanced CLIP for Zero-Shot Anomaly Detection”, and the article is currently under review.

Acknowledgements

I wish to express my sincere thanks to Prof. Xingyu Li, my supervisor, for her invaluable mentorship during my Master's program. Her guidance through our regular meetings, willingness to entertain every inquiry, and provision of hands-on guidance in specialized research areas have been a significant source of inspiration for both my professional work and personal growth.

I'm also deeply thankful to my senior lab mates, Hanqiu Deng and Pengyue Hou, for their unwavering support and assistance. And to all my other lab mates, who have shared in this journey, your fellowship has been deeply appreciated.

Also, a special note of thanks to Jingxuan Zhu, my partner and fellow graduate student at the University of Alberta. Her encouragement and support have been a constant source of strength during my time as a Master's student.

Finally, I want to acknowledge my friends here in Canada - you know who you are. We've tackled projects, hackathons, and even the ski slopes. Thank you for creating unforgettable memories. Your companionship has provided immense support throughout my journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Anomaly detection (AD) aims to identify instances containing anomalous and defective patterns that deviate from normal data. This is crucial across various applications such as detecting defects in manufacturing [7], [94], analyzing medical images [8], [66], and monitoring video surveillance [47], [48], [87]. Anomalies are often characterized by subtle differences in texture, color, shape, or motion, blending seamlessly into normal surroundings. Due to their diverse nature, AD poses significant challenges and has been the focus of extensive research in real-world applications.

Initially, supervised methods [9] were employed, replying on both normal and abnormal instances as training data. However, their effectiveness heavily depends on the quality of annotations, and collecting diverse anomalies is difficult due to their rarity. Consequently, AD shifted towards unsupervised learning, aiming to model normal data without requiring abnormal samples.

Unsupervised AD models [16], [41] are often built under one-class setting, training unique models for each category of normal images. These models demonstrates superior performance in industrial benchmarks [7], [94] and exhibit promising potential in medical imaging tasks [6], [66]. However, challenges persist: (1) In medical domain, many proved successful model on industrial AD face limitations when applied to medical domain. The high heterogeneity of lesions restricts the generalizability of previous approaches, particularly in identifying rare lesion or anomalies. This discrepancy highlights the need for specific solution that improves the performance of medical AD. (2) Existing methods excel when abundant normal images are available, but in scenarios with limited or no normal training data, their performance significantly declines. Additionally, the one-class-one-model form poses scalability issues, especially in multi-class scenarios where multiple models are needed. While some strides have been made with the introduction of N-class-one-models [81], [88], they still necessitate retraining for new classes. To address this limitation, zero-shot anomaly detection emerges as a potential solution. However, current zero-shot methods have not fully met the efficacy required for diverse anomaly scenarios, despite additional training data or high computational overhead.

This thesis targets to tackle the above two challenges by leveraging auxiliary information, such as prior domain knowledge, pretrained models, and hidden normal spectrum in unlabeld test data. Specifically:

1. By leveraging prior knowledge on medical imaging, we proposes an anomaly synthesis method to generate pseudo-anomalies. The synthetic anomalies function as positive data for training AD models, effectively transforming unsupervised learning into a pseudo-supervised problem. We discussed the guidelines for pseudo-anomaly generation and associated training strategies. We validate our hypothesis on Liver Tumor Segmentation (LiTs) [8] dataset and achieve state-of-the-art (SOTA) tumor segmentation accuracy.

2. To comprehensively address the scalability and performance issues, we introduced a zeros-shot AD approach leveraging Contrastive Language-Image Pretraining (CLIP) [54] model and paired image referencing system. This framework enhances ZSAD performance through prompt-guided AD and auxiliary visual reference from unlabeld test images. Additionally, we introduce test-time adaptation with synthesized anomalies to refine anomaly localization capacity of the model.

1.2 Thesis Outline

The thesis is outlined as follows:

Chapter 2 provides an overview of the general background in anomaly detection. It delves into the characteristics of anomaly detection datasets, the rationale for approaching anomaly detection problems in an unsupervised context, and a review of existing methods and approaches in the field.

In Chapter 3, we delve into the details of my published paper, “Unsupervised Liver Tumor Segmentation with Pseudo Anomaly Synthesis”. This chapter focuses on the unique challenges of medical imaging in anomaly detection and discusses the novel approach developed for liver tumor segmentation.

Chapter 4 presents an in-depth discussion of my submitted paper titled “Dual-Image Enhanced CLIP for Zero-Shot Anomaly Detection”. This chapter elaborates on the methodologies, experiments, and findings that contribute to enhancing zero-shot learning in anomaly detection, particularly in industrial applications.

Finally, Chapter 5 concludes the thesis, summarizing key insights and discussing potential directions for future research.

Chapter 2

Background

As described in Chapter 1, the goal of this thesis is to develop accurate and efficient anomaly detection models, especially to overcome the challenges in medical tumor segmentation and zero-shot industrial AD. This chapter delves into the problem formulation specific to anomaly detection, providing examples of anomalous images. It further explores common unsupervised approaches employed in image anomaly detection and discusses the distinctive characteristics of medical AD.

2.1 Foundations and Overview of Anomaly Detection

Anomaly detection is the process of identifying out-of-distribution (OOD) examples, which essentially involves pinpointing data instances that deviate from the prevalent pattern within a dataset. In the domain of AD, a majority of data aligns with a "normal" class distribution. Anomalies, or outliers, are rare, often absent from training data, and their identification can be resource-intensive due to the complexity involved. An anomaly is flagged during testing if it significantly diverges from the normal data distribution. This detection is crucial for the maintenance of system integrity across various applications.

The challenge of visual AD is not only in detecting these outliers but in defining them. Anomalies can occur under various conditions: a flaw in an industrial process, an unusual tissue in medical imagery, or unexpected behavior captured by a surveillance system. These irregularities might stem from new or previously unseen changes in the environment, making AD a vital component in a broad spectrum of image analysis applications.

2.1.1 Problem Formulation

In the context of anomaly detection, we are given a set of unlabeld images, video frames, or pixels, denoted as $\mathcal{X}_{\mathcal{N}}$. It is assumed that the majority of $\mathcal{X}_{\mathcal{N}}$ conforms to the distribution of normal data, $p_{\mathcal{N}}$. However, in few-shot and zero-shot learning scenarios, we may not have a sufficient number of normal samples to establish $p_{\mathcal{N}}$ robustly. To tackle this, we utilize a feature extractor F that is pre-trained or fine-tuned to represent data in a feature space where anomalies inherently deviate from normal instances.

AD aims to determine whether a test sample y is an anomaly by measuring its conformity to the learned features of normal data. The AD function is defined as:

$$AD(y) = D(F(y), F(p_{\mathcal{N}})) \tag{2.1}$$

Here, D is a distance metric that quantifies the deviation of the test instance's features $F(y)$ from the features of the normal data $F(p_{\mathcal{N}})$. The feature extractor F maps the raw data to a set of discriminative features. In the context of few-shot and zero-shot learning, F can leverage transfer learning to encode meaningful representations without the need for a well-defined normal distribution $p_{\mathcal{N}}$.

In some cases where a yes or no discrimination result is needed, a threshold τ would serve as a decision boundary for determining the model's sensitivity to anomalies, adjustable based on the desired balance between false positives and false negatives. This formulation allows for AD even with limited normal samples, bridging the gap between traditional unsupervised learning and the emerging paradigms of few-shot and zero-shot learning.

AD can be categorized into two distinct levels:

- Sample Level: Also known as novelty detection or anomaly classification, this level involves image-level classification. The goal is to distinguish whether a query image is anomalous or contain anomalies.

- Pixel Level: Also referred to as anomaly segmentation or localization, this level deals with the detection and, ideally, segmentation of subtle anomalies within images. These are deviations that closely resemble the training data, often localized to small regions.

Whether referred to as anomaly detection, novelty detection, outlier detection, or one-class classification, the fundamental objective remains the same: to reliably and accurately identify the out-of-distribution samples.

2.1.2 Anomalous Samples

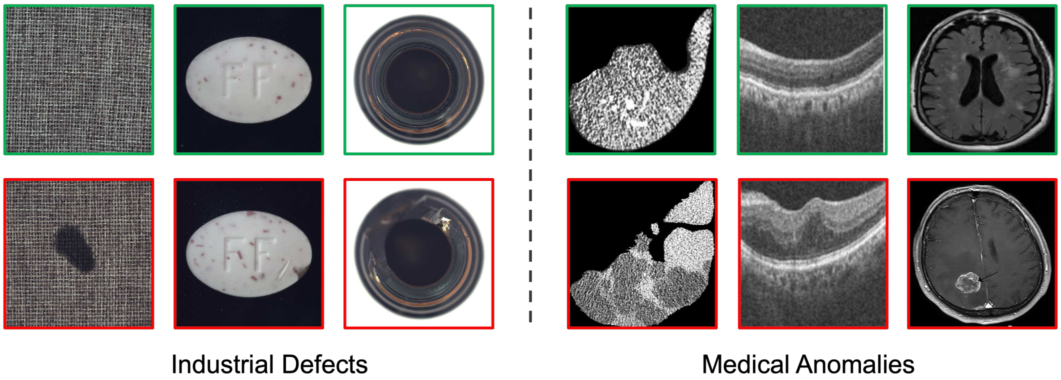


Figure 2.1: Image samples showcasing normal and anomalous instances. The top row with green borders depicts normal images, while the bottom row with red borders illustrates samples with defects or lesions. The left segment includes industrial images from MVTecAD [7], and the right segment features medical images from datasets such as LiTs [8], Retina OCT [66], and BraTs [6].

The visual array in Figure 2.1 provides a curated glimpse into the prevalent datasets utilized in industrial and medical anomaly detection. In industrial contexts, as exemplified by MVTecAD, anomalies manifest across a spectrum from textural deviations to outright object defects. The training sets consist entirely of normal images to establish a baseline of 'normalcy,' whereas test sets are designed to challenge models with a mixture of normal and anomalous samples. The provision of pixel-level annotations for anomalies is a definitive advantage, allowing for precise model evaluation and fine-tuning.

Medical datasets such as LiTs [8], Retina OCT [66], and BraTs [6] are more specialized, each focusing on distinct anatomical regions or imaging modalities. The anomalies here range from lesions in liver tissue to abnormalities in retinal OCT scans, and represent a critical need for accurate detection due to their implications for patient diagnosis and treatment. The datasets come with their own set of challenges, including variability in the manifestation of conditions and the subtlety of pathological changes against the backdrop of normal variations in human anatomy.

Each of these datasets reflects the intrinsic complexities of their respective domains. For instance, in industrial images, anomalies might be characterized by clear-cut contrasts against a structured backdrop, facilitating the task of anomaly segmentation. Medical images, conversely, require discerning often subtle differences between healthy and pathological states, a task complicated by the diverse presentations of diseases and the high stakes of medical diagnostics.

The analysis of these datasets underscores the diverse requirements of anomaly detection systems, industrial AD systems must be robust against a variety of defect types, while medical AD systems must be sensitive to the slightest indications of disease. Both domains benefit from the advancements in AD methods, with the goal of achieving precise anomaly detection and localization.

2.1.3 Evaluation Metrics

Accurate evaluation is essential in anomaly detection to ensure the effectiveness of a model in distinguishing between normal and anomalous instances. Different metrics are used to assess performance at the sample and pixel levels:

Area Under the Receiver Operating Characteristic Curve (AUROC)

The AUROC is one of the most widely used metrics for evaluating the performance of anomaly detection models. It measures the ability of a model to discriminate between the normal and anomalous classes across different threshold settings. The AUROC represents the likelihood that the model will rank a randomly chosen anomaly higher than a randomly chosen normal instance. It is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(u))\, du \tag{2.2}$$

Maximum F1 Score (F1Max)

The F1Max is the maximum F1 score achieved by a model over all possible thresholds. The F1 score is the harmonic mean of precision and recall, providing a balance between the model's sensitivity (recall) and its ability to only flag true anomalies (precision). F1Max is particularly useful when seeking a single metric to capture the trade-off between precision and recall.

$$\text{F1Max} = \max_t \left(2 \cdot \frac{\text{Precision}(t) \cdot \text{Recall}(t)}{\text{Precision}(t) + \text{Recall}(t)} \right) \tag{2.3}$$

Average Precision (AP)

Average Precision summarizes the precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. It provides a single figure of merit for evaluation across different recall levels, which is valuable when the cost of false negatives varies.

$$\text{AP} = \sum_n (\text{Recall}_n - \text{Recall}_{n-1}) \text{Precision}_n \tag{2.4}$$

where Precision_n and Recall_n are the precision and recall at the n^{th} threshold.

Area Under the Precision-Recall Curve (AUPR)

AUPR is another important metric for problems with a significant class imbalance, which is common in anomaly detection. It provides a comprehensive view of the trade-off between precision and recall without being dominated by the large number of true negatives, unlike AUROC.

$$\text{AUPR} = \int_0^1 \text{Precision}(\text{Recall}^{-1}(u))\, du \tag{2.5}$$

Per-Region-Overlap (PRO)

Proposed in [7], PRO considers the segmented regions in anomaly detection and measures the overlap of these regions with the ground truth. It is computed by applying a threshold to the anomaly map and identifying connected components or regions. Each region is then compared to the ground truth to calculate the overlap. The threshold is adjusted on a validation set such that the largest connected component is just smaller than a pre-defined minimum defect area. This threshold is then used to evaluate the anomaly maps of the test set

Dice Coefficient

The dice coefficient, often used for medical imaging, measures the overlap between the predicted and actual anomalies at the pixel level. It is particularly useful for medical diagnosis where binary decisions are common. The DICE coefficient is 2 times the area of overlap between the predicted and true anomalies divided by the total number of pixels in both the predicted and true anomalies, giving a value between 0 (no overlap) and 1 (perfect overlap).

$$\text{dice} = \frac{2 \times |Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \tag{2.6}$$

where Y is the set of pixels in the ground truth anomaly and $\hat{Y}$ is the set of pixels in the predicted anomaly.

Each of these metrics offers insights into different aspects of a model's performance. While AUROC and AUPR provide aggregate measures of performance across all thresholds, F1Max and AP offer insights into peak performance. The PRO metric, along with the dice coefficient, is particularly useful for anomaly localization tasks, as it assesses the precision of the detected regions against the true anomalies. The dice coefficient is especially valuable in medical contexts for its clear interpretation related to actual diagnostic decisions. Selecting the appropriate metric depends on the specific requirements and constraints of the application domain within anomaly detection.

2.2 Unsupervised Anomaly Detection Approaches

2.2.1 One-Class-One-Model Approaches

In the one-class-one-model paradigm, distinct models are independently trained to identify anomalies within specific categories. This method is efficient when a substantial amount of normal data is available for each class. These can be broadly categorized as reconstruction-based methods, pseudo anomaly synthesis, distillation-based methods, and representation-based methods.

Reconstruction-Based Methods

Reconstruction-based methods are a foundational approach in anomaly detection. They rely on the principle that normal data will have a certain pattern or structure that a neural network can learn to replicate. By training the network to encode and then decode images, it essentially learns to reconstruct the input image from a compact representation. The assumption is that the network trained exclusively on normal data, will not reconstruct anomalies as well since it hasn't learned their patterns.

During the testing phase, the network is presented with new images, and the reconstruction error is measured. Anomalies are expected to have a higher reconstruction error because they deviate from the normal pattern the network has learned. This error forms the basis of the anomaly score, where a higher error indicates a greater likelihood of the image being anomalous. Fig 2.2 shows the basic flows of reconstruction-based methods.

Classical reconstruction-based methods [2], [4], [27], [62], such as autoencoders (AE), variational autoencoders (VAE [37]), and generative adversarial networks (GANs [28]) have shown proficiency in this technique. AEs are simple yet effective at capturing the distribution of normal data. VAEs add a probabilistic twist, allowing for the generation of new samples from the learned data distribution, while GANs introduce an adversarial component, with one

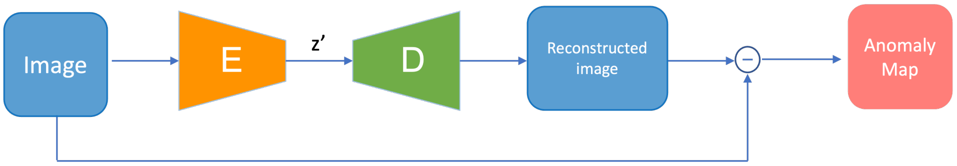


Figure 2.2: The workflow of reconstruction-based methods, illustrating the process of inputting images into a reconstruction network such as AE or GAN to produce a reconstructed image. The discrepancy between the original and reconstructed images is used to derive an anomaly map, facilitating anomaly detection.

network generating candidates and the other evaluating them, leading to a robust model of normality.

For instance, F-AnoGAN [62] trains the model to reconstruct only normal information, with the network purely trained on normal images. However, this assumption can fall short when abnormal features are still reconstructed accurately, blurring the distinction between normal and anomalous data. To address this, some methods have been developed to more strictly constrain the normal reconstruction. One such approach is MemAE [27], which introduces a memory bank that stores a collection of normal feature prototypes. This strategy ensures that reconstruciton is biased towards normality, creating a distinct discrepancy on abnormal test images.

Pseudo Anomaly Synthesis

To address the challenge of limited anomalous data samples, several algorithms introduce the concept of pseudo anomaly synthesis. Methods like CutPaste [41], DRAEM [83], and NSA [64] generate artificial anomalies using data augmentation techniques. Specifically, CutPaste [41] augment the normal image with in-distribution images patches. DRAEM [83] generates pseudo anomalies by overlaying random noise with texture patterns. These methods synthesize pseudo anomalies by employing various image manipulations like cropping, rotating, transforming, and overlaying. The result is a collection of artificially crafted anomalies that vary in shape and texture, closely resembling actual defects.

This synthetic augmentation serves a dual purpose. First, it enhances the

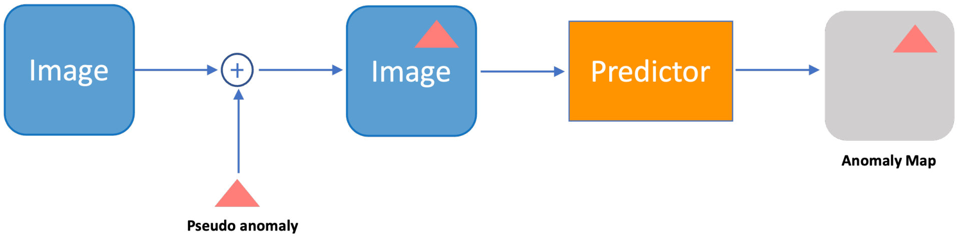


Figure 2.3: The principle of pseudo anomaly synthesis approaches, showing the integration of synthesized anomalies with normal images for training a model to predict the localization of pseudo anomalies.

variety of anomaly examples, allowing models to learn from a broader spectrum of defect appearances. Second, it ensures that models do not overfit to the limited real-world anomalies available, promoting better generalization when encountering novel anomalies.

The training process with pseudo anomalies is straightforward yet effective. Models are exposed to a mixture of normal images and those with superimposed synthetic anomalies. Through training, they learn to distinguish the subtle difference between anomalies and normal baselines. During testing, these models leverage their training to accurately identify and segment anomalies, even those that were not present in the training dataset.

Distillation-Based Methods

Knowledge distillation in anomaly detection [16], [61] leverages the asymmetry of information between two neural networks: a teacher and a student. The teacher, often a pre-trained and more complex network, imparts its high-level feature extraction capabilities to the student, a less complex or untrained network. The student's goal is to mimic these capabilities as closely as possible. Normal data serves as the training ground, providing a baseline of standard patterns against which anomalies can be detected.

During testing, the student's inability to replicate the teacher's performance on anomalous samples manifests as a discernible difference in outputs. This discrepancy is quantified to identify anomalies, thus capitalizing on the student's limited exposure to only normal data during training.

The teacher-student (T-S) framework generally employs similar architec-

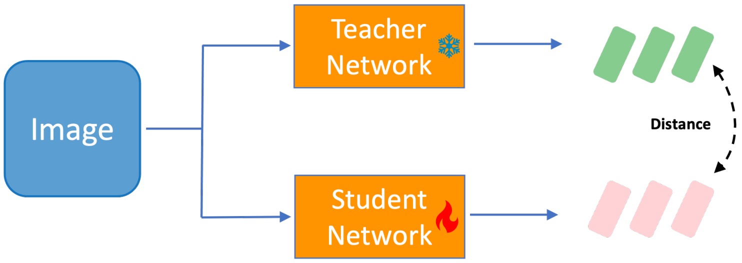


Figure 2.4: The basic architecture of T-S anomaly detection frameworks, depicting the knowledge transfer from the teacher network to the student network and the subsequent anomaly score calculation based on their output disparity. The snowflake denotes the model is frozen, flame icon represents the trainable modules.

tures, allowing for a direct comparison of feature representations. A novel twist to this paradigm is the reverse distillation [16] process. Here, the student network is trained not on raw images but on embeddings produced by the teacher network. The student's task is to reconstruct the teacher's multi-scale representations, effectively creating a reversed flow of information. This method enriches the student's feature space with nuanced details specific to the teacher's interpretation of 'normal,' making the detection of anomalies

Representation-Based Methods

Representation-based approaches utilize neural networks to derive compact feature vectors from images, which are then used to define normalcy. The anomaly score is typically a function of the distance between the test image's embedded vector and the normal reference vector. This methodology benefits from the absence of a training stage, requiring no parameters beyond the pre-trained network used as the backbone. This strategy aligns with the principles of metric learning and is akin to clustering in its operation.

In testing, the distance between the sample features and the normal features is computed to detect anomalies, with methods such as SPADE [13], PaDIM [14], and PatchCore [57] employing various distance metrics to ascertain anomaly scores and generate score maps. Among these, PatchCore [57] proposed an innovative "coreset" concept on optimizing the normal prototype selections, which greatly enhances the efficiency and effectiveness of the

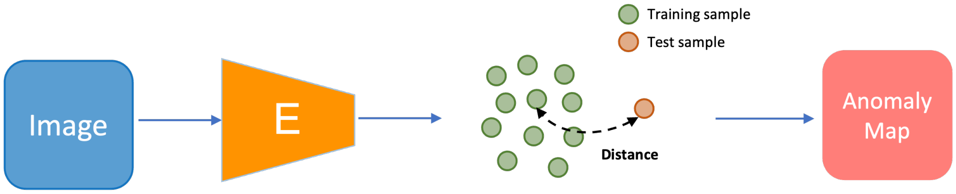


Figure 2.5: Illustration of the representation-based anomaly detection process. A pre-trained neural network acts as a feature extractor of the input image and embeds it into a feature vector. The anomaly score is derived by measuring the distance between this embedding and reference embeddings determined from training set.

detection process.

One-Class Classification Approaches

One-class support vector machine (OC-SVM) [65] and support vector data description (SVDD) [71] are foundational algorithms in one-class classification. SVDD transforms all the standard training data into a predefined kernel space, aiming to enclose the data within the smallest possible hypersphere. The training process is primarily concerned with determining the hypersphere's radius and center. Anomalies are identified when they fall outside the boundaries of this hypersphere.

However, these classic methods often struggle in scenarios involving high-dimensional, data-rich scenarios. To address these limitations, advancements have been made, such as Deep SVDD [60], which replaces the traditional kernel function with a neural network, enhancing the algorithm's ability to handle complex data structures. Furthermore, Patch-SVDD [80] extends this improvement to patch-wise detection, offering a more granular approach to identifying anomalies.

Normalizing Flow-Based Methods

A flow model trains a mapping that maximizes likelihoods for extracted features which are quantifiable in the latent space. Normal samples are naturally localized into the trained distribution range, while abnormal samples are projected onto a separate distribution range. As shown in Figure 2.7.

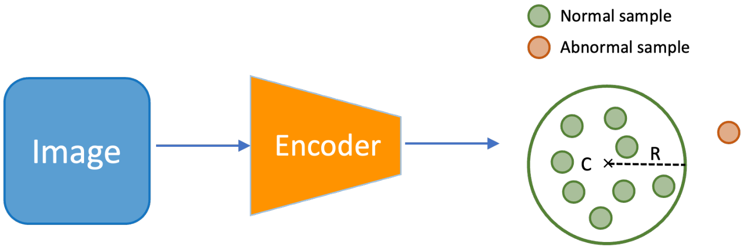


Figure 2.6: One-class classification approach trains a network that learns to transform most of the data representations into a minimum-volume hypersphere with center C and radius R. Normal samples are mapped within the hypershpere, while anomalous samples fall outside.

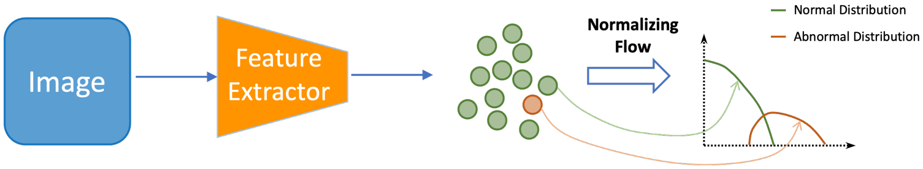


Figure 2.7: Flow-based approaches. The NF block learns the density estimation which transforms the anomaly-free samples into a Gaussian distribution, while the anomalous samples are projected to a distinct distribution.

DifferNet [58] processes vectors through a one-dimensional normalizing flow (NF) with the features extracted by a pretrained feature extractor. CFlow [29]incorporates the positional encoding as a conditional vector to improve the spatial relevance. Fast-Flow [82] implements the two-dimensional NF to enhance detection accuracy with the multi-scale aggregated results. CS-Flow [59] integrates information across scales with the addition of scaled volumes. UFlow [67] further boosts AD capabilities with a multi-scale Transformer-based feature extractor and a U-shaped NF block architecture, effectively managing complex data structures.

2.2.2 Multi-Class Anomaly Detection

While traditional one-class-one-model approaches are effective with plenty of normal data, they struggle in diverse and multi-class scenarios due to high memory demands and poor handling of intra-class variability. The push towards multi-class AD aims to address these limitations. These models, trained on normal samples from various categories, are tasked with detecting anomalies without specific adjustments for each category, and without categorical information during training and inference. UniAD [81] innovates with a layerwise query encoder and a neighbor masked attention module to avoid identical reconstructions and better model multi-class distributions. OmniAL [88] leverages a network enhanced with Dilated Channel and Spatial Attention blocks to increase reconstruction discrepancies, along with a DiffNeck feature to examine multi-level differences. Additionally, SNL [19] introduces spatial-channel distillation and intra-& inter-affinity distillation techniques for assessing structural distances in teacher-student network frameworks.

2.2.3 Zero-Shot Anomaly Detection

The scarcity of normal data caused by privacy concerns or lack of domain-specific training data prompts the exploration of ZS approaches. With no training data available, ZSAD necessitates a powerful models that can be well-generalized across varied visual features and backgrounds.

Advancements in data scale have led to significant strides in pretrained visual language models [5], [11], [22], [54], which demonstrate remarkable proficiency in a variety of downstream tasks [34], [35], [49], [53], [73], [90], [91]. A prime example is the CLIP model, through contrastive vision-language pretraining on a diverse array of internet-sourced image-text pairs, it exhibits exceptional generality and adaptability. This model is particularly adept at zero-shot inference, displaying a superior capacity for recognizing images beyond its training data. Recent explorations have extended the zero-shot capabilities of CLIP models to tasks like open-vocabulary semantic segmentation, achieved by harnessing intrinsic dense features [24], [45], [89]. Additionally, efforts to optimize CLIP's recognition performance have been fruitful, focusing on areas such as prompt engineering [90], [91], adapter modules [26], [85], and additional training for enhanced vision-language alignment [34], [35]. Importantly, CLIP's inherent ability to detect out-of-distribution data without additional training has catalyzed its application in zero-shot anomaly classification and localization.

The WinCLIP model [32] marks the first use of CLIP [54] for prompt-guided anomaly detection, setting text descriptions for normal and abnormal states and seeking matched images based on vision-language embedding correlations. Building on this, AnoCLIP [20] enhances localization representation and implements V-V attention introduced in [45]. However, vision-language models such as CLIP are primarily trained to align with the class semantics of foreground objects rather than the abnormality/normality in the images, and as a result, their generalization in understanding the visual anomalies is restricted, leading to weak ZSAD performance. Existing zero-shot prompt-guided AD models often lack robust visual representation as a basis for detecting anomalies. Addressing this, methods such as [10], [92] propose fine-tuning the pretrained CLIP model with auxiliary images for cross-set training/validation.

2.3 Anomaly Detection in Medical Domain

2.3.1 Current Challenges in Medical AD

In medical image analysis, unsupervised anomaly detection plays a pivotal role in identifying atypical features, such as abnormal structures or lesions, indicative of various medical conditions. While normality in biomedical images is usually well-defined and more straightforward to collect, anomalies present a significant challenge due to their heterogeneity. It's often impractical to gather a comprehensive dataset encompassing all possible abnormal cases, particularly for rare diseases or new anomalies. This open-set nature of medical data necessitates approaches beyond conventional supervised methods, which may struggle with unseen abnormalities.

Additionally, the study of medical anomaly detection encompasses a range of image modalities and body components, such as Retina OCT [66], Brain MRI [6], and Liver CT [8], etc. Each of these modalities presents its own unique set of characteristics, leading to distinct challenges:

Diversity in Normal Data: A comprehensive representation of normal data is crucial. However, acquiring and annotating such medical data is challenging and costly, which can cause incomplete representations.

Distinct Image Characteristics: Medical images differ significantly from natural daily datasets. Pretrained models on large general datasets like ImageNet [21] might not adequately capture the nuances of medical images, potentially missing subtle yet crucial signs of anomalies that often require expert analysis or context for identification.

High Sensitivity Requirement: Given the critical nature of medical diagnosis, test sensitivity is critical. Anomaly detection models in this domain must be highly accurate, taking into account individual patient differences and variations across different ages and geners.

Specific Preprocessing Needs: Medical images often require extensive preprocessing due to their unique sampling methods and inherent noise.

These factors highlight the unique and complex nature of anomaly detection in medical imaging, emphasizing the need for specialized research and study in this area.

2.3.2 AD Approaches on Medical Images

Medical anomaly detection methods often adapt techniques used in industrial settings. Reconstruction-based approaches are widely implemented in this domain [3], [63], [74], [76], [78], [93], however encounter significant performance limitation. Alternatively, pseudo-supervised methods play a significant role [33], [43], [69], [70], [86] that incorporates pseudo-positive samples to enhance the detection accuracy by overlaying color, texture, and semantic outliers to normal samples, a model is trained to segment the synthetic anomalous regions [43], [69], [70], [86] and reached promising AD capacity. Despite extensive research in pseudo anomaly synthesis and model training, two fundamental questions remain under-explored:

- *Should pseudo-anomalies approximate the queries in test phase?*
- *How to train the segmentation model on pseudo-synthesized data?*

Chapter 3 delves into these questions, exploring the impact of pseudo-anomaly synthesis on liver tumor segmentation performance. It also examines the limitations of industrial anomaly detection methods when applied to medical datasets.

2.4 Conclusion

While notable progress in anomaly detection in both industrial and medical imaging, certain challenges remain, especially in enhancing AD performance in ZSAD setting, and formulating more effective approach on medical lesion segmentation. In the realm of medical imaging, there's a need for investigating into the generation and training principle involving pseudo anomalies. Regarding zero-shot anomaly detection, the potential exists to develop multimodality joint inference system that uncover hidden normal patterns in textual and visual information. Our research is focused on overcoming these obstacles through the improved utilization of auxiliary information, aiming to make meaning substantial contributions to the evolution and refinement of AD in diverse domains.

Chapter 3

Unsupervised Liver Tumor Segmentation

3.1 Introduction

Liver tumors are one of the leading causes of cancer-related deaths, and accurately segmenting them in medical images such as computed tomography (CT) is crucial for early detection and diagnosis. While supervised tumor segmentation methods show promising results, their performance is heavily dependent on high-quality annotated data, which can be expensive to obtain. Furthermore, due to the high heterogeneity of tumors, the generalizability of supervised models may be limited in identifying rare lesions or anomalie. Recently, there is an increased interest in treating tumors as anomalies in medical images and exploring unsupervised learning approaches, i.e. anomaly segmentation, to address the aforementioned challenges. In the context of unsupervised anomaly segmentation, a model is expected to identify and segment potential abnormalities by learning from a healthy cohort of patients during model training.

Anomaly synthesis has emerged as a prominent approach that incorporates pseudo-positive samples to enhance anomaly segmentation. By overlaying color, texture, and semantic outliers on normal samples, a model is trained to segment the synthetic anomalous regions [18], [30], [33], [43], [69], [70], [79], [86]. Despite yielding promising results, there exists significant variation in methods for generating pseudo anomalies. For instance, [42], [69], [70] generate anomalies by utilizing in-distribution image patches, while [30], [43], [79], [86] focus on producing lesions that closely resemble real anomalies. Additionally, prior arts usually focus on model design [43], [84]. However, there is little study explicitly tackling the following two fundamental questions behind this paradigm.

- *Should pseudo-anomalies approximate the queries in the test phase?*
- *How should the segmentation model trained on the synthesis data?*

Addressing these questions necessitates an understanding of the medical mechanisms behind tumor appearances in imaging. Variations in Hounsfield Unit (HU) values in CT images are a key indicator of lesions, with tumor tissues typically exhibiting distinct densities compared to surrounding normal tissues. This difference in density affects X-ray attenuation, leading to variations in HU values, as exemplified by hepatocellular carcinomas, which are on average 11 HU lower than adjacent liver parenchyma in the portal venous phase [40]. By synthesizing pseudo anomalies that reflect these HU variations, we create a spectrum of anomalies that closely simulate real tumor appearances. This approach not only mimics the realistic presentation of liver tumors but also leverages auxiliary information, aiding unsupervised models in accurately identifying and segmenting actual tumors.

To explore these objectives and questions further, this chapter delves into unsupervised liver tumor segmentation. We utilize an adapted version of the Discriminative Joint Reconstruction Anomaly Embedding (DRAEM) [84], introducing a nuanced anomaly synthesis pipeline and a balanced two-stage training strategy. This method demonstrates impressive performance on the Liver Tumor Segmentation dataset (LiTs) [8], showcasing the efficacy of our approach in addressing the challenges of unsupervised anomaly detection in medical imaging.

3.2 Preliminaries

This section tackles two fundamental, yet under-explored questions in pseudo-supervised anomaly segmentation with synthetic abnormalities. The reasoning offers insights for designing the proposed solution.

Q1: About pseudo anomaly generation: Should pseudo-anomalies approximate the common queries in the test set?

Pseudo anomaly is introduced to establish the boundary that distinguishes abnormality, transforming the unsupervised problem into pseudo-supervision, which helps the model learn normal patterns by providing negative samples. Since there is no clear definition of what constitutes an anomaly, there shouldn't be any bound or limit on pseudo anomaly synthesis. Instead of focusing on creating pseudo anomalies that match known abnormal patterns in queries, we advocate generating a diversity of anomalies to facilitate a model to learn the comprehensive normal spectrum. In particular, when dealing with unsupervised tumor segmentation, we believe that generating a large diversity of pseudo anomalies in terms of intensity, shape, and textures facilitates addressing the high heterogeneity in tumors. This motivates the design of the proposed pseudo anomaly generation module.

Q2: About model training: Should the model training follow the exact supervised training principles on synthetic anomalies?

The success of supervised learning relies on the IID assumption that both the training and test data follow an identical distribution. Under this assumption, a model is usually well-trained on the training set with multiple iterations. However, we argue that one shouldn't follow the same philosophy to train a model on pseudo anomalies in anomaly detection and segmentation. According to the reasoning in **Q1**, a covariate shift is likely to exist between the synthesized and query anomalies. We visualize this covariate shift by 2-D TSNE in Figure 3.1(C), where both tumor samples and normal images are from the LiTs dataset [8]. Consequently, due to the potential covariate shift between the synthesized and the common anomalies in pseudo-supervised segmentation, training a model on the pseudo anomalies may cause a bias and harm

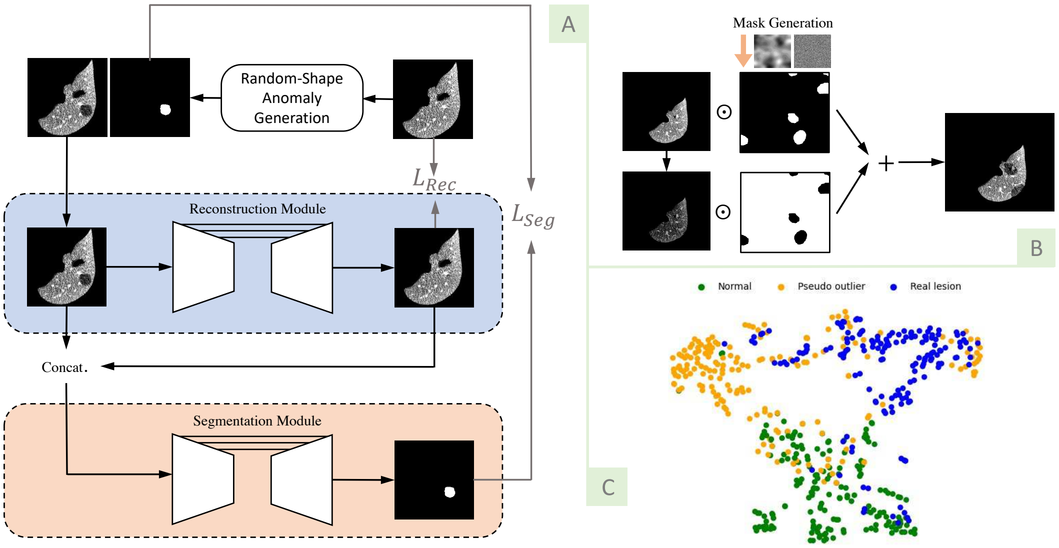


Figure 3.1: (A) Systematic diagram of the proposed unsupervised liver tumor segmentation scheme. During training, synthetic abnormalities are fed to a restoration net followed by a segmentation net. To avoid model overfitting on synthesis, the two models are trained in two phases represented by blue and orange, respectively. In inference, a query is directly passed to the two networks for segmentation. (B) Proposed synthesis pipeline based on Gaussian noise stretching. (C) Liver image embedding by 2-D t-SNE.

its performance on real queries. In another words, a good-fit model on the pseudo-anomaly data may fail on real testing data. Our ablation experiment shown in Figure 3.6 validates this hypothesis. Therefore, unlike conventional supervised learning that requires a relatively long training time, we argue that model optimization on anomaly synthesis for pseudo-supervised segmentation should stop early to preserve the model's generalizability on queries. Our answer to **Q2** inspires us to design the two-phase training strategy in this chapter.

3.3 Methodology

Toward unsupervised liver tumor segmentation, we incorporate our reasoning to **Q1** and **Q2** into the DRAEM-similar [84] architecture. As depicted in Figure 3.1(A), the framework comprises random-shape anomaly generation, a restoration network, and a segmentation network. Unlike DRAEM training both networks jointly, we propose a two-phase learning to avoid segmentation model over-fitting on synthetic abnormalities. In inference, only the reconstructive network and segmentation network are deployed on queries. Compared to DRAEM, our experiments show that both the proposed anomaly generation module and the two-phase learning strategy boost the liver tumor segmentation performance in terms of segmentation accuracy and model stability.

3.3.1 Pseudo Anomaly Generation

The anomalous training samples are simulated by the anomaly synthesis module, which generates masks of random shapes and sizes through Gaussian noise and morphological transformations. Initially, Gaussian noise is generated with the same resolution as a normal image and then blurred with a Gaussian kernel. The noise is then stretched and thresholded to produce a binarized mask. Subsequently, closing and opening operations with the elliptical kernel are applied to the binarized mask to obtain an anomaly segmentation mask. The detailed algorithm is shown in Algorithm 1.

Algorithm 1 Random-shape Pseudo Anomaly Generation

```
Input: $Image, Threshold$
Output: $AnomalyMask, Label$
  $NoiseImage \leftarrow gaussianNoise(Image\_height, Image\_width)$
  $BlurImage \leftarrow gaussianBlur(NoiseImage, kernal\_size)$
  $StretchImage \leftarrow rescaleIntesity(BlurImage, (0, 255))$
  $AnomalyMask \leftarrow binarize(StretchImage, Threshold)$
  $AnomalyMask \leftarrow Morph\_open\_close(AnomalyMask, kernel\_ellipse)$
  if sum(AnomalyMask) $> 0$ then
      $Label \leftarrow 1$
  else
      $Label \leftarrow 0$
  end if
```

Using the generated anomaly mask M_s, we proceed to synthesize the abnormal sample I_s. In CT slides, unhealthy patterns in liver regions are demonstrated by abnormal HU values. Therefore, we propose to randomly shift the intensity of the slice and overlay the new intensity values on the original image I within the mask regions (as shown in Figure 3.1(B)). We formulate the proposed abnormality synthesis as

$$I_s = (1 - M_s) \odot (I + C) + M_s \odot I, \quad |C| \in (minRange, maxRange), \quad (3.1)$$

where I_s represents synthesized anomalies, $\odot$ is element-wise multiplication, and C is a random value drawn from a Gaussian distribution within a defined range.

It is noteworthy that unlike [30], [86] that aims to fabricate pseudo anomalies to approximate the common patterns of liver tumors, we follow our principle to **Q1**, leverage the stochastic nature in the proposed synthesis process to generate a wide spectrum of anomalies deviating from normal patterns (as shown in Figure 3.1(C)). We provide a demonstration of synthesized pseudo samples in Figure 3.2, our experiment shows that our method outperforms [86] by 12% in Dice.

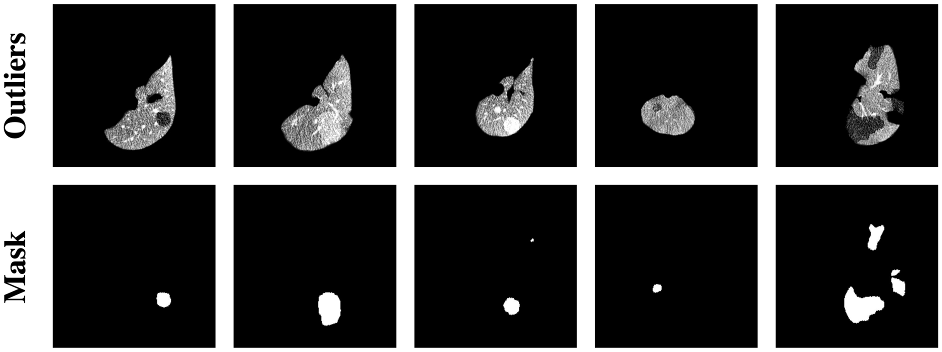


Figure 3.2: Pseudo anomalous samples, and the corresponding anomaly masks.

3.3.2 Model Architecture and Training Functions

The reconstruction network is trained to restore anomalous regions while preserving the normal regions. The segmentation network takes the concatenation of the restoration and pseudo-anomalous image as input and targets to estimate an accurate segmentation map for the anomaly. For the reconstruction network, we use U-Net [56] with 3 encoder and decoder blocks as backbones. The specific encoder block in the restoration network adopts the architecture proposed in [33], where it consists of 2 weight-standardized convolutions [52]

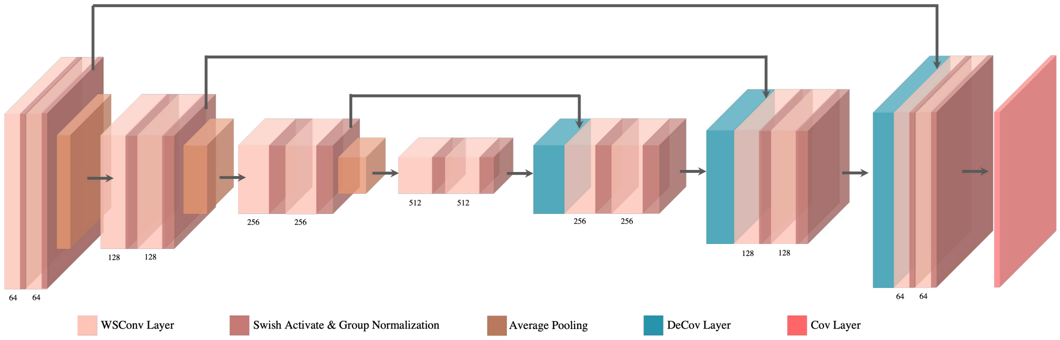


Figure 3.3: Reconstruction network architecture. The encoder consists of 4 blocks, each block contains two 3×3 weight-standardization convolutions, followed by the swish activation and group normalization. Symmetrically, each decoder block has deconvolution with 2×2 as kernel size, $stride = 2$, followed by weight-standardized convolution.

followed by swish activation [55] and group normalization [77]. We illustrate the model architecture in 3.3.

To address diverse levels of model optimization complexity, we train the two networks consequently in two phases. The reconstruction model is first trained to restore the anomalous region in synthetic abnormal images with L_1 loss:

$$L_{rec}(I_s, \tilde{I}_s) = |I_s - \tilde{I}_s|, \tag{3.2}$$

where $I_s, \tilde{I}_s$ are the pseudo outlier augmented sample and the reconstruction image. After freezing the well-trained generative module, we slightly train the segmentation model to avoid bias introduced by the covariance shift. To accommodate potential small tumors, Focal Loss [46] is adopted:

$$L_{seg}(M_s, \tilde{M}_s) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \alpha_j (1 - \tilde{m}_{s,ij})^{\gamma} \log(\tilde{m}_{s,ij}) \tag{3.3}$$

where $\tilde{m}_{s,ij}$ is the predicted probability of class j at pixel i and α_j is the weight for class j, and $M_s, \tilde{M}_s$ are the ground truth and the estimated anomaly masks.

3.4 Experiments

3.4.1 Experimental setting

Dataset Preparation: We evaluate the proposed method on the Liver Tumor Segmentation (LiTs) dataset [8] from MICCAI 2017 challenge. LiTs dataset consists of 131 abdominal CT scans with the paired liver and liver tumour ground truth labels. Unlike with previous works [23], [43], which perform the cross-fold validation on the LiTs dataset, we argue that training on the retrieved partial samples from an unhealthy CT scan is not ideal for the model to learn the complete liver feature distribution. Therefore, we train our model on an anomaly-free dataset BTCV [39], which provides 40 healthy CT abdomen scans and the corresponding organ masks.

For all CT volumes in training and test, HU values are transformed into grayscale and the liver Region of Interest (ROI) is extracted according to the organ annotations. Then 2D slices are obtained along the Axial plane, resized to 256×256, and normalized independently by histogram equalization.

Implementation Details: We run the experiments on dual Nvidia RTX-3090 GPUs. The threshold for pseudo mask generation is set to be 200, and the intensity range of the random intensity shift is $[-100, 100]$. The focal loss parameters are defined as $\alpha = 1$ and $\gamma = 2$. We use PyTorch [50] to implement the proposed method. The model is trained for 200 epochs for the first stage and just 1 epoch for the second stage to avoid bias introduced by pseudo anomalies. The learning rate is set to 0.0001, with a batch size of 8 using Adam [36] optimizer. We follow previous studies and use the Dice score as our evaluation metric.

3.4.2 Results and Discussion

Comparison to SOTA: We quantitatively compare the proposed method with state-of-the-art unsupervised liver tumor segmentation methods including Zhang et al. [86], Hu et al. [30] ASC-Net [23] both with and without manually-designed post-processing and report the results in Table 3.1. The fully supervised method is taken as performance upper bound. As shown in

Table 3.1: Liver tumor segmentation on LiTs [8]. Our method exhibits the best Dice with a standard deviation of 1.78. Results with $*$ are directly copied from original papers.

Methods	**Supervision**	**Dice**
Zhang et al. [86]	✓	61.91*
DRAEM [84]	X	14.75
Zhang et al. [86]	X	40.78*
ASC-Net [23]	X	32.24*
ASC-Net + postprocessing [23]	X	50.23*
Hu et al. [30]	X	**59.77***
Ours	X	53.03

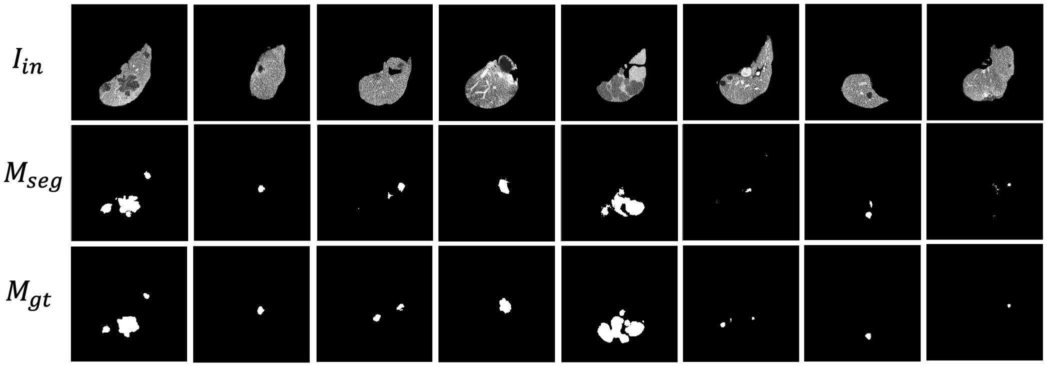


Figure 3.4: Tumor segmentation on real liver tumor data, from easy (left) to difficult (right). I_{in}: Input , M_{seg}: segmentation mask, and M_{gt}: Ground-Truth.

Table 3.1, our approach significantly outperforms the other methods, with the exception of [30] and shrinks the gap between unsupervised method and fully-supervision. Notably, [30] leverages extensive clinical prior knowledge to synthesize pseudo anomalies resembling real tumors. Furthermore, our approach achieves a substantial reduction in runtime at $0.018s/slice$, compared to $0.476s/slice$ in [30] which operates on 3D volume, incurring higher memory usage and slower inference time. In Figure 3.4, we show our segmentation results on real tumor data in the LiTs dataset.

Ablation on model components: The proposed method and DRAEM differ in three aspects: pseudo anomaly generation (corresponding to **Q1**), two-phase training (corresponding to **Q2**), and U-Net backbone in the restoration

Table 3.2: Ablation study of two-phase training (TP), pseudo anomaly (PA), and reconstructive network. The baseline is DRAEM model [84]. Asterisks indicate statistical significance (*: $p \leq 0.05$, **: $p \leq 0.001$) when using a paired Student's t-test compared to baselines.

Method	$+TP$	$+PA$	$+U\text{-}Net$	**Dice**
Baseline				14.75 ± 14.28
Baseline	✓			21.31 ± 12.54
Baseline	✓	✓		$30.17 \pm 5.50^{*}$
Baseline		✓	✓	$40.06 \pm 6.85^{*}$
Baseline	✓	✓	✓	$\mathbf{53.03 \pm 1.78}^{**}$

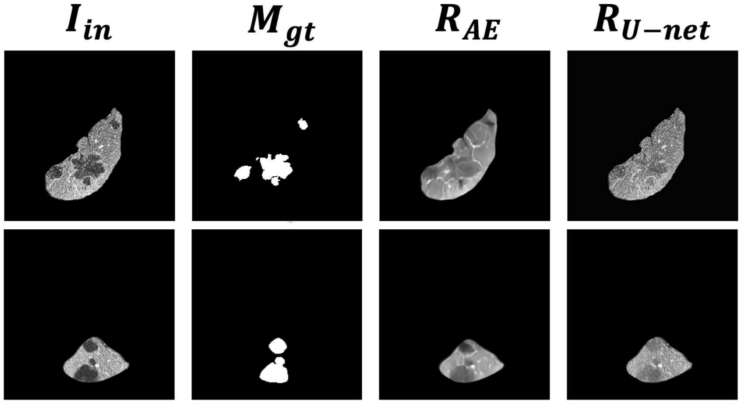


Figure 3.5: Visualization of image reconstruction by AE and U-Net.

net. In this ablation study, we take the DRAEM as baseline, decouple these factors, and evaluate their impact in terms of tumor detection (by AUROC) and segmentation (by Dice) on LiTs. We run this ablation 3 times, and the performance is reported in the Table 3.2. The selection of the anomaly size mask threshold selection is shown in Figure 3.7 and Table 3.3.

The two-phase training strategy improves Dice by 6.5% on the baseline. When combined with U-Net and PA, there's a significant 13% performance boost compared to using only U-Net and PA, validating our hypothesis that light segmentation training on pseudo anomalies helps address the covariant shift between synthetic anomalies and real tumors. We further extended the training of the segmentation net to 200 epochs and captured tumor detection performance (by AUROC) and segmentation quality (by Dice) every 5 epochs. It's worth emphasizing that these experiments incorporate the synergistic ap-

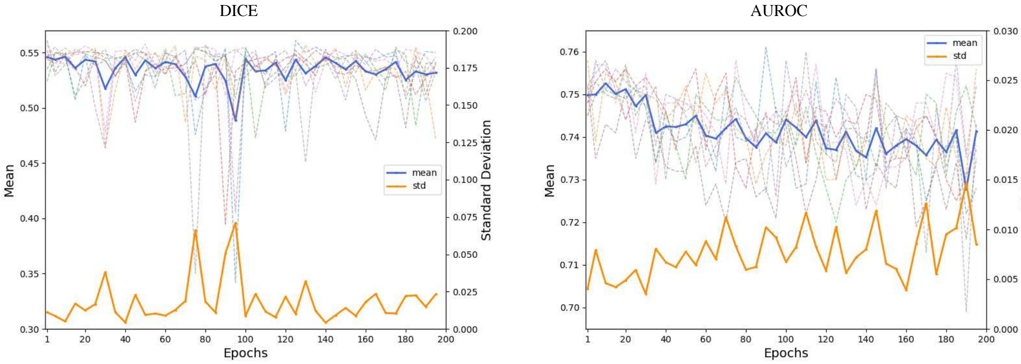


Figure 3.6: An illustrative depiction of the evaluation performance of the segmentation network reveals a tendency to overfit shortly after a short period of training. Throughout this process, the reconstruction network maintains a frozen state.

Table 3.3: Mask threshold ablation numeric results. Changing the threshold (TH) would generally alter the shape of the pseudo anomalies, and decreasing the threshold would allow larger outliers to be produced.

TH	**AUROC**	**Dice**
180	72.678 ± 3.117	40.756 ± 8.577
190	72.189 ± 2.767	46.856 ± 5.490
200	$\mathbf{75.854 \pm 1.538}$	$\mathbf{52.562 \pm 2.288}$
210	70.933 ± 7.195	38.844 ± 15.658
220	74.100 ± 2.723	46.200 ± 6.4274

plication of TA, U-Net and PA, as this combination has proven to demonstrate optimal outcomes with TA. Therefore, Figure 3.6 results diverge from the Dice score, where TA solely influences the baseline, yielding a comparetively less pronounced impact on reducing training perturbation As shown in Figure 3.6, the mean AUROC keeps decreasing, and the standard deviation keeps increasing. The perturbation also occurs in Dice after 50 epochs. We attribute this to model overfitting on the pseudo data, which hurts model's generalizability on queries.

Additionally, the proposed anomaly synthesis module and U-Net designed in our restoration net significantly boost the segmentation performance. Figure 3.5 presents a visualization comparison of reconstructions generated by autoencoder (AE) and U-Net. Compared to the AE-based network, the skip-

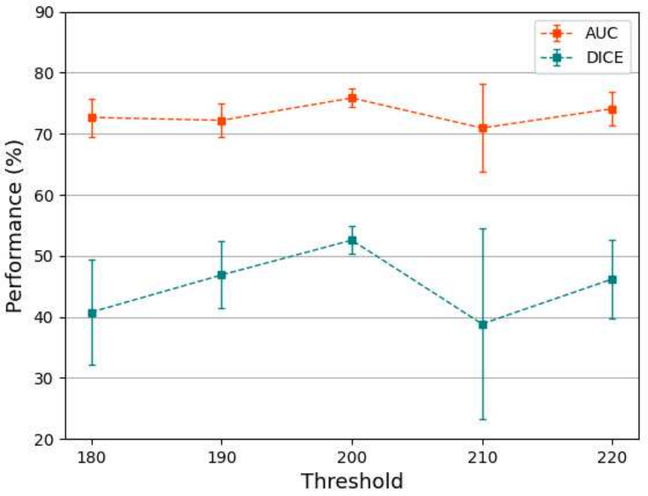


Figure 3.7: Mask threshold ablation results. $Threshold = 200$ gives the model the best and most stable accuracy on AUROC and DICE. Error bars show the standard deviation across 9 runs.

connection in U-Net helps preserve the texture details in liver reconstruction images, which facilitates the downstream segmentation task.

3.5 Conclusion

In this chapter, we tackled the challenging problem of unsupervised liver tumor segmentation and proposed a two-stage pseudo-supervision solution with synthetic anomalies. By generating anomalies spreading over a large spectrum, the synthesis data facilitated the model in finding normal sample boundary in embedding space. The two-stage training strategy mitigated the impact of covariant shift between synthesis data and actual tumor data on model optimization, and thus avoid segmentation model's overfitting on synthetic anomalies. Experimentation suggested that the proposed method performs comparably to SOTA methods. Looking ahead, we aspire to extend our exploration of model performance to encompass various other diseases and data modalities and investigate the integration of both real and synthetic tumor within the model training pipeline.

Chapter 4

Dual-Image Enhanced CLIP on Industrial Zero-Shot Anomaly Detection

4.1 Introduction

In this chapter, we explore the realm of Zero-shot Anomaly Detection (ZSAD), a critical and emerging area in anomaly detection. ZSAD presents the unique challenge of identifying anomalies without relying on training samples from the specific target dataset, making it a crucial approach in situations where such data is scarce or unavailable. This context highlights the need for models capable of detecting anomalies with high adaptability and robustness.

To tackle ZSAD, MuSc [44] was proposed to leverage unlabeld images in the test set as references for the query images. It operates under the argument that a rich amount of normal information implicit in unlabeld test images is underutilized. Even if the test image is anomalous, it still contains some normal patches that can serve as references. MuSc achieves SOTA performance, but as it requires knowledge of the test set distributions before inference, it aligns more with transductive rather than inductive learning. Also, its extensive comparison with all test set images can be time-consuming and computationally intensive.

Alternatively, WinCLIP [32] deployed CLIP [54] model and used text prompts for anomaly measurement, significantly improving over other category-agnostic methods in a zero-shot AD setup and extending the capabilities of the

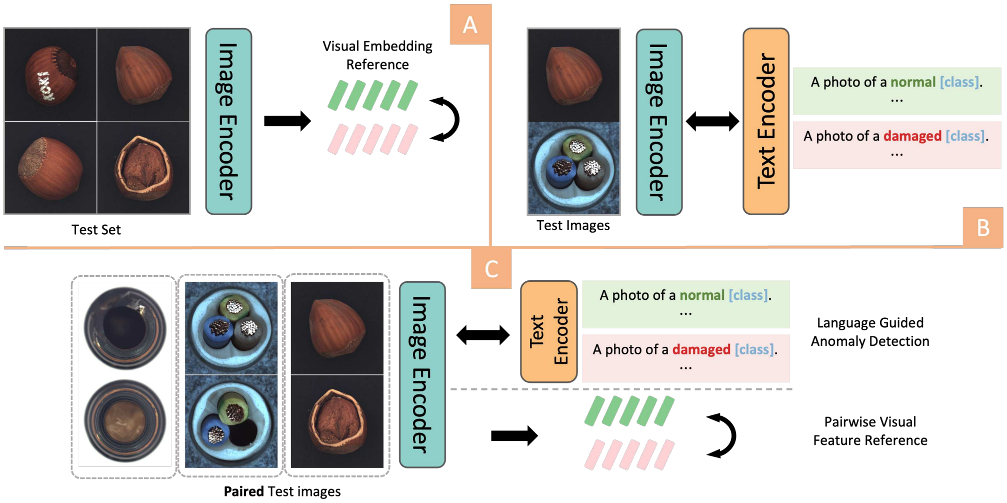

A
B
C
Test Set
Image Encoder
Visual Embedding Reference
Test Images
Image Encoder
Text Encoder
A photo of a normal [class].
...
A photo of a damaged [class].
...
Paired Test images
Image Encoder
Text Encoder
A photo of a normal [class].
...
A photo of a damaged [class].
...
Language Guided Anomaly Detection
Pairwise Visual Feature Reference

Figure 4.1: Overview of the Dual-Image Enhanced CLIP Zero-Shot Anomaly Detection Model. Traditional approaches often depend on a single modality for anomaly detection, where (A) demonstrates the use of image embeddings, and (B) illustrates reliance on text prompts. Our proposed method, shown in (C), integrates both visual and textual information, utilizing a dual-image input to enrich the feature space for a more robust and comprehensive anomaly detection framework.

CLIP model [54]. Subsequent approaches like [10], [20], [68], [92] further enhanced ZSAD capabilities. Recent works [10], [92] have begun fine-tuning the pretrained CLIP model with auxiliary anomalous image and, testing it on the target datasets. Alternatively, AnoCLIP [20] introduced a test-time adaptation (TTA) module to alter the visual representation space of the CLIP model. These studies underscore the importance of adding training parameters to the pretrained CLIP model to strengthen its anomaly localization ability. However, solely incorporating semantic information from text prompts may not fully exploit the potential of large vision-language models. Since the vision-language space isn't perfectly aligned, many visual anomalies implicitly defined in the visual distribution remain uncovered. Additional visual references need to be incorporated to assist the language-based ZSAD, especially for misplaced objects, rare-seen objects, and complicated scenes, whose anomaly information is usually hard to obtain from large pretrain datasets.

To address these issues, we propose a novel framework (see Fig. 4.2) that utilizes a pair of unlabeld images during testing. Our framework comprises a pretrained CLIP model, a test-time adaptation module, and an input path for image pairs to leverage the additional visual reference information into the language-vision AD. The anomaly score of a query image depends not only on its textual zero-shot score but also on the score derived from its randomly paired reference image. Additionally, we enhanced the model's AD capability by adding a TTA module involving pseudo anomaly synthesis to improve the agnostic ability to locate anomalies.

In summary, our contributions are threefold:

- We propose a novel ZSAD method that processes a pair of images, enhancing existing CLIP-based AD methods. This approach incorporates an additional reference image, operates without the need for further training and significantly boosts AD performance.

- We developed a TTA module that includes pseudo anomaly synthesis methods adopted from DRAEM [83], effectively refining the AD capabilities of the pretrained CLIP model.

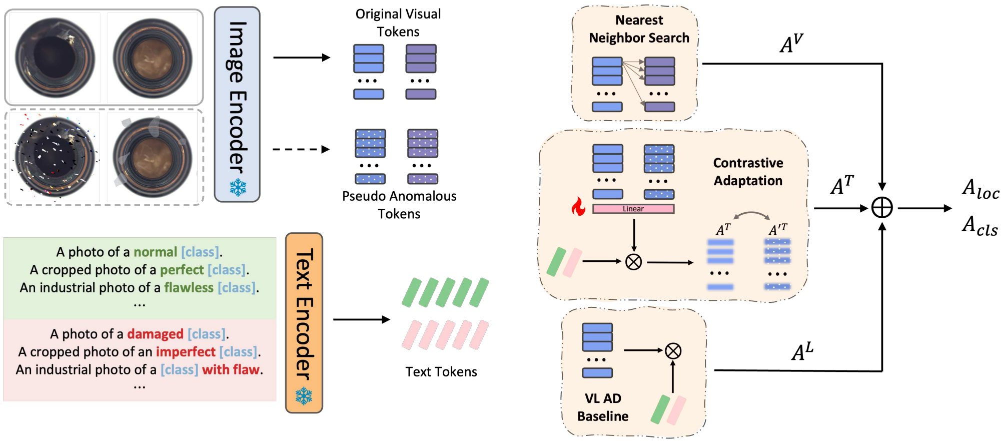
- Comprehensive experiments on MVTecAD [7] and VisA [94] reveal that our methods achieve comparable performance with current SOTA ZSAD methods in both anomaly classification and anomaly localization.

4.2 Methodology

In this section, we first introduce the CLIP-based baseline model for zero-shot anomaly classification and localization. Following this, we delve into details of our dual-image enhancement model. Lastly, we specify our test-time adaptation mechanism to refine the model's AD capability. Fig. 4.2 provides a comprehensive overview of our framework.

4.2.1 CLIP for Zero-Shot Anomaly Detection

CLIP's zero-shot visual recognition, trained on a multi-million image-text pair dataset, aligns images with textual descriptions through a visual encoder and a text encoder. These encoders respectively transform images and text prompts



A photo of a normal [class].
A cropped photo of a perfect [class].
An industrial photo of a flawless [class].
...

A photo of a damaged [class].
A cropped photo of an imperfect [class].
An industrial photo of a [class] with flaw.
...

Figure 4.2: Overview of our framework for Dual Image Enhanced CLIP. The left part shows the feature extraction process from the vision and text encoder, and the right section shows the inference process. The snowflake denotes the modules are frozen, and the flame icon represents trainable modules.

(e.g., *a photo of a [class]*) into visual and text tokens in a shared feature space. The model's ability to compare these tokens via cosine similarity allows it to identify class concepts within images.

For anomaly detection, CLIP utilizes semantic concepts of "normal" and "anomalous" states. Multiple prompts with varied descriptors (like "perfect", "broken", *etc.*) are used to create averaged text tokens representing these states, t_n and t_a for normal and anomalous text tokens, respectively. Anomaly score for an image is computed based on the similarity between its visual token and these averaged text tokens. Specifically, given a text prompt and the corresponding class token v, the sample-level anomaly score A^L_{cls} is computed as:

$$A^L_{cls} = F(v, t_a, t_n) = \frac{\exp(\langle v, t_a \rangle)/\tau}{\exp(\langle v, t_n \rangle/\tau)) + \exp(\langle v, t_a \rangle/\tau)} \tag{4.1}$$

where τ is the temperature hyperparameter. Note that no visual information is injected into the model, but rather unknown anomalies are detected through the powerful open-world generalization of CLIP.

The computation is extended from global visual embeddings to patch-level visual embeddings to derive the corresponding segmentation maps $A^L_{loc} \in \mathbb{R}^{H\times W}$, the final layer of the visual encoder has a set of patch tokens $p_{(j,k)} \in \mathbb{R}$

that potentially contain image local information in the patch level. For a patch token $p_{(j,k)}$, the local anomaly score is computed as:

$$A_{loc}^{L} = \left\{F(p_{(j,k)}, t_a, t_n)\right\}_{j=0,k=0}^{h-1,w-1} \tag{4.2}$$

$$= \left\{\frac{\exp(\langle p_{(j,k)}, t_a\rangle/\tau)}{\exp(\langle p_{(j,k)}, t_n\rangle/\tau) + \exp(\langle p_{(j,k)}, t_a\rangle/\tau)}\right\}_{j=0,k=0}^{h-1,w-1} \tag{4.3}$$

However, since CLIP was primarily trained to align the class tokens with the text token for global classification, there's a lack of alignment between local patch tokens and text embeddings that leads to limited performance in segmenting anomalous regions. Hence, after iterative explorations [20], [45], [92], V-V attention was adopted to produce the local-aware patch tokens.

In original Q-K-V attention, the attention score can be disproportionately influenced by specific tokens, leading to a representation that is disturbed by unrelated local features, which can weaken the model's localization ability to detect anomalies. The V-V attention mechanism is proposed as an alternative that enhances the local features without additional training. This novel attention mechanism replaces the queries and keys with values.

$$V^{l} = Proj.(Attention(V^{l-1}, V^{l-1}, V^{l-1})) + V^{l-1} \tag{4.4}$$

By focusing on self-attention within the values themselves, V-V attention avoids bias introduced by the query and key interactions in Q-K-V attention. It reduces the disturbance caused by other tokens, ensuring that each value contributes significantly to its own representation. As a result, attention maps produced by V-V attention exhibit a pronounced diagonal pattern, indicating that each token predominantly attends to itself, thereby preserving its local information.

In our model, the architecture remains the same with AnoCLIP [20], which follows a 2-way forward path. The original Q-K-V attention path was kept to produce the class token, which was used to calculate sample level anomaly score A_{cls}^{L}. The patch tokens used for localization score A_{loc}^{L} are all computed by the V-V attention path.

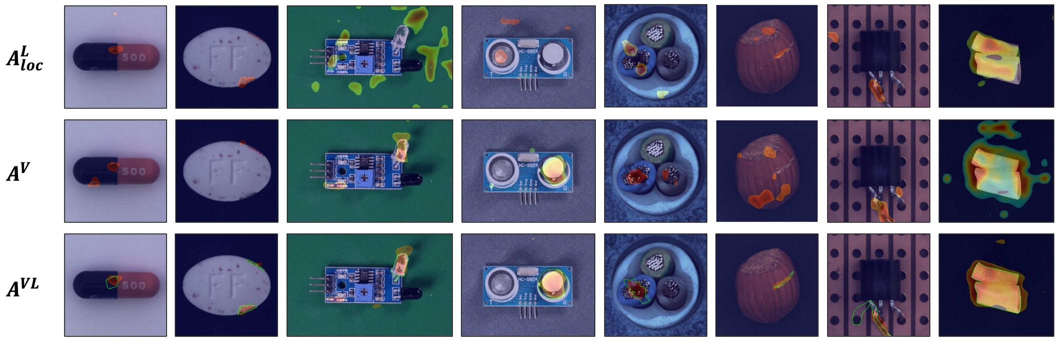


Figure 4.3: Qualitative illustration of the comparison with AD results on MVTecAD and VisA. The top row illustrates the result solely using textual information. The middle row depicts detection results through paired queries' visual feature comparison. The bottom row showcases more robust results achieved by integrating both language and visual features, and the ground truth is marked with green boundaries.

4.2.2 Dual Image Feature Enhancement

As shown in Fig. 4.2, we proposed a novel approach that inputs a pair of images in test-time. Unlike previous CLIP-based AD works [20], [32], [68], [92] which predominantly rely on text prompts for inference, we incorporate additional visual information to facilitate a more comprehensive joint vision-language prediction. To highlight the effectiveness of our approach, we provide a comparative analysis in Fig. 4.3.

Fig. 4.3 demonstrates a significant observation: the exclusive dependence on either textual or visual information alone proves inadequate for the accurate detection of certain anomalies. The limitation in leveraging text stems from the constraints inherent in utilizing state descriptions within prompts, with terms like "broken" or "damaged" falling short of encapsulating the full spectrum of potential anomalies. Making inferences on a single image invites biases and misinterpretations, emphasizing a more extensive visual intra-class diversity to form a baseline for normalcy. For instance, consider the "PCB" example in column 3 of Fig. 4.3, where a misplaced LED is an anomaly; however, using text description alone is insufficient for its detection. Moreover, logical anomalies, global distortions, rare objects, or complicated scenes are more challenging to discern from by text-based method. This emphasizes the importance of incorporating additional visual context for a varied example to enhance the detection accuracy. Conversely, relying solely on pairwise visual comparisons also presents limitations, as the reference image could itself be anomalous. Consequently, this paves the way for an integrated approach that combines textual and visual data to overcome these challenges. As depicted in Fig. 4.3, employing dual-image inputs within the CLIP-based framework mitigates these issues, contributing to improved anomaly localization accuracy.

In response to these findings, our framework introduces a novel strategy that capitalizes on both textual and visual features. This is achieved by a unique process of randomly selecting pairs of test images to serve as the query and reference images. For these image pairs, we extract patch tokens, denoted as $q_{(j,k)}$ for the query and $r_{(m,n)}$ for the reference. These patch tokens form the basis for the pairwise visual feature comparison.

In this pairwise feature comparison strategy, each patch token $q_{(j,k)}$ from the query image undergoes a nearest neighbour search with the patch tokens from the reference image, effectively using the latter as a memory repository. The anomaly score for each query patch token $q_{(j,k)}$ is determined by calculating its cosine similarity with all reference patch tokens set $\mathcal{S}_r = \{r_{(m,n)} \mid m \in \{1, \ldots, H\}, n \in \{1, \ldots, W\}\}$. The maximum similarity score, indicating the minimum deviation, is then designated as the anomaly score for the query patch:

$$A^V_{(j,k)} = \min_{r_{(m,n)} \in \mathcal{S}_r} \left(1 - \text{sim}\left(q_{(j,k)}, r_{(m,n)}\right)\right) \tag{4.5}$$

In the equation above, $A^V_{(j,k)}$ represents the visual reference anomaly score of the query patch $q_{(j,k)}$, the overall anomaly score $A^V \in \mathbb{R}^{H \times W}$ for the query image, is a composition of all the patch scores across the entire image. sim represents the cosine similarity between the patch tokens of the two samples.

As a result, the vision-language joint anomaly score can be computed as:

$$A^{VL}_{loc} = A^V + A^L_{loc} \tag{4.6}$$

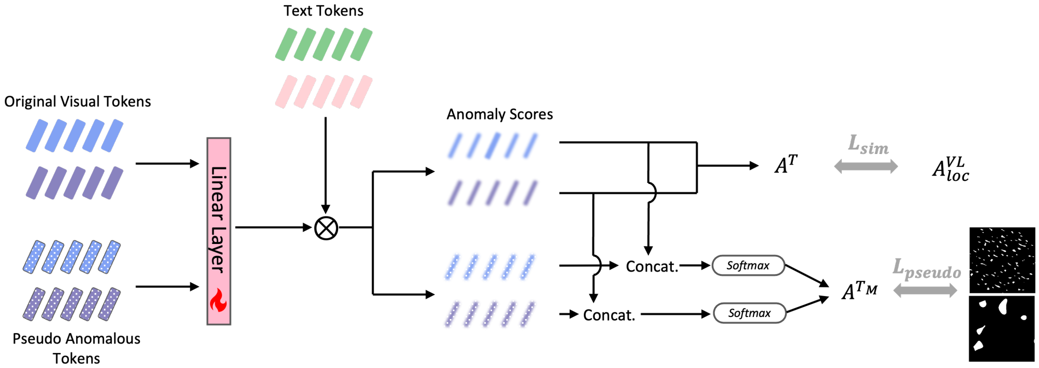


Figure 4.4: Workflow of the test-time adaptation module. The module inputs patch tokens through a linear layer, aligning predictions on the adapted token with the zero-shot vision-language joint anomaly score. Pseudo-anomalous samples are compared with original samples to predict pseudo-anomaly masks. The flame icon denotes trainable components. A^{T_M} denotes the prediction for the pseudo anomalies.

4.2.3 Test-Time Adaption with Pseudo Anomaly Synthesis

As the visual-language alignment needed to be refined for AD, we proposed a test-time adaptation module to boost the CLIP-based model's AD capabilities. Our TTA module is achieved through a linear adapter, as depicted in Fig. 4.4. For the original image, we utilize the pseudo anomaly synthesis technique from DRAEM [83] to introduce image corruptions. DRAEM creates random-shaped pseudo anomaly masks using Perlin noise [51] and overlays textures from [12] onto the original image at masked locations. The resultant pseudo-anomalous patch tokens, denoted as $q'_{(j,k)} \in \mathbb{R}$, encapsulate pseudo-anomalous features.

The online adaptation of the original and synthesized patch tokens is mathematically represented as:

$$q^T_{(j,k)} = \frac{1}{2}\left(G(q_{(j,k)}) + q_{(j,k)}\right) \tag{4.7}$$

$$q'^T_{(j,k)} = \frac{1}{2}\left(G(q'_{(j,k)}) + q'_{(j,k)}\right) \tag{4.8}$$

Here $G(\cdot)$ denotes the linear computation. Subsequently, these adapted patch tokens are aligned with text tokens to compute the anomaly score:

$$A^T = \left\{F(q^T_{(j,k)}, t_a, t_n)\right\}_{j=0,k=0}^{h-1,w-1} \tag{4.9}$$

To optimize the weights of the linear layer, we establish self-supervised tasks using pseudo anomalies. For the original and adapted patch tokens $q^T_{(j,k)}$, $q'^T_{(j,k)}$ from queries, we design two discriminative self-supervised tasks for TTA:

(1) For predicting pseudo anomaly masks M_a, we define the L_{pseudo} loss as:

$$L_{\text{pseudo}} = \frac{1}{|\mathcal{S}_a|} \sum_{(j,k)\in\mathcal{S}_a} \left(-M_a \cdot \log \left(\frac{\exp(A'^T)}{\exp(A^T) + \exp(A'^T)} \right) \right)_{j,k} \tag{4.10}$$

Here, $\mathcal{S}_a$ represents the set of indices (j,k) where $M_a \neq 0$, indicating regions augmented by the pseudo mask M_a. L_{pseudo} prompts the adapter to retain abnormal features and recognize pseudo anomalies, aiding in the subtle detection of real anomalies.

(2) To encourage the adapter to preserve normal features and uphold general anomaly detection capabilities, we utilize the similarity loss L_{sim} to ensure that adapted anomaly scores A^T are consistent with the zero-shot vision-language joint localization:

$$L_{\text{sim}} = \text{sim}\left(A^{VL}_{loc}, A^T\right) \tag{4.11}$$

The aggregate learning objective to train our adapter is $L = L_{\text{pseudo}} + \beta L_{\text{sim}}$. This TTA process is efficient and does not require any training data or annotation. Finally, the overall anomaly classification and localization score for the query image should be computed as:

$$A_{loc} = \lambda_1 A^V + \lambda_2 A^T \tag{4.12}$$

$$A_{cls} = \lambda_3 A^L_{det} + \lambda_4 \max_{j,k} A^V + \lambda_5 \max_{j,k} A^T \tag{4.13}$$

4.3 Experiment

4.3.1 Experimental Setup

Datasets.

In our chapter, we conducted experiments using the MVTecAD [7] and VisA [94] datasets. Both of these datasets offer a wide array of subsets featuring various objects and textures. MVTecAD includes high-resolution images with dimensions varying from 700^2 to 1024^2, while the VisA comprises rectangular images with resolutions around $1.5K \times 1K$, each accompanied by corresponding anomaly ground truth masks. Specifically, MVTecAD encompasses 5 texture categories and 10 object categories, whereas VisA is composed of 12 subsets, each dedicated to different objects. In this chapter, we exclusively utilized the test dataset to evaluate zero-shot anomaly classification and localization, without the acquisition of additional datasets.

Data Preprocessing

We adopted the OpenCLIP's [31] outlined preprocessing methods. The process commenced with the bilinear resizing of the images to a standard height dimension of 240 pixels, coupled with a subsequent channel-wise normalization process. The VisA dataset posed a unique challenge due to its assortment of non-square images, which did not conform to the desired $(240, 240)$ dimension post-resizing. To address this discrepancy and ensure compatibility with the CLIP model's training dataset dimensions, we deployed the image tiling technique, which involved segmenting each image into two equal parts of $(240, 240)$. These segments were later merged back into a single image. Post-inference, the overlapping areas are averaged to maintain consistency in the final image representation.

Metrics.

We assess the efficacy of our model by utilizing the Area Under Receiver Operator Characteristics (AUROC) image-level AUROC is used for anomaly detection, and pixel-level AUROC is measured for evaluating anomaly localization. However, the metric is dominated by a large number of normal pixels and is thus kept high despite false detections. We thus additionally report the F1Max score and Area Under Precision-Recall (AUPR) as a balanced calculation of the precision and recall to overcome the class imbalance. In addition to that, we compute the Per-Region-Overlap (PRO) to measure anomaly localization, which weights each connected component within the ground truth of varying sizes equally, making it more robust than simple pixel measurement.

Implementation.

We adopt ViT-B-16+ [25] as the visual encoder and the transformer [72] as the text encoder by default from the public pretrained CLIP model [31]. For the text encoder, following previous work of AnoCLIP [20], employing 22 base templates collected from CLIP [54], 7 pairs of state prompts, and 4 domain-aware prompts to generate sufficient prompts. We adhered to the data preprocessing pipeline outlined in OpenCLIP [31] for both MVTecAD and VisA benchmarks, standardizing image sizes to $(240, 240)$. Regarding the scoring coefficients, we configured $\lambda_1, \lambda_3, \lambda_4, \lambda_5$ to 1, and set λ_2 to 1.5. For TTA, we use the AdamW [38] optimizer and set the learning rate to 0.001, $\beta = 0.5$, the adaptor is optimized with 2 training steps. We report the mean and variance of the results over 6 random seeds.

4.3.2 Performance

Tab. 4.1 presents the performance of zero-shot anomaly detection on MVTecAD and VisA datasets. Our proposed method is compared with prior ZSAD based works, including CLIP [54], WinCLIP [32], AnoCLIP [20], and MuSc [44]. Notely, MuSc utilized the entire test set for visual reference, aligning with transductive rather than inductive learning. For a fair comparison, we adapted MuSc to our pairwise image setting and used ViT-B-16+ as the backbone, denoted as MuSc-2. From the table, our proposed methods exhibit exceptional performance, significantly outperforming AnoCLIP by margins of $2.2\%, 6.0\%, 6.2\%$ in AUROC, F1Max, and PRO for anomaly localization. We also achieve advancements in anomaly classification, surpassing other methods by substantial margins. This trend of exceptional performance is consistent on the VisA dataset. Qualitative results for ZSAD are further detailed in Fig. 4.3, illustrating our model's capacity to effectively classify and localize the anomalies across varied samples.

Additionally, Fig. 4.2 presents a comparison of our method against other AD models by AUROC scores on MVTecAD. Here, we categorize current zero-shot anomaly detection methodologies into three paradigms: Auxiliary Data

Table 4.1: Zero-shot Anomaly Localization (AL) and Anomaly Classification (AC) on MVTecAD and VisA datasets. Bold indicates the best performance and underline indicates the runner-up unless otherwise noted. MuSc-2 denotes inference with 2 images.

Methods	MvTecAD						VisA					
	AL			AC			AL			AC		
	AUROC	F1Max	PRO	AUROC	F1Max	PRO	AUROC	F1Max	PRO	AUROC	F1Max	PRO
CLIP [54]	19.5	6.2	1.6	74.0	88.5	89.1	22.3	1.4	0.8	59.3	74.4	67.0
WinCLIP [32]	85.1	31.7	64.6	91.8	91.9	96.5	79.6	14.8	59.8	78.1	79.0	81.2
AnoCLIP [20]	90.6	36.5	77.8	92.5	93.2	**96.7**	91.4	17.4	75.0	79.2	79.7	81.7
MuSc-2 [44]	92.4	41.2	76.5	81.7	89.1	90.3	92.6	**26.7**	63.2	69.4	75.1	73.3
Ours	92.6	41.8	82.6	93.1	94.0	96.6	**94.8**	23.5	78.6	82.6	**81.0**	84.2
Ours+	**92.8**	**42.5**	**84.0**	**93.2**	**94.1**	**96.7**	94.2	24.1	**79.7**	**82.9**	80.9	**84.7**

Table 4.2: Comparative analysis of AD methods in Full-shot and Zero-shot settings.

Setting	Full Shot		Zero-shot					
			Aux. Data		Transd.	Induct.		
Methods	DRAEM [83]	UniAD[81]	AprilGAN [10]	AnomalyCLIP[92]	MuSc [44]	WinCLIP [32]	AnoCLIP [20]	Ours
A.C.	88.1	96.5	87.6	91.5	97.8	91.2	92.5	93.2±0.8
A.L.	87.2	96.8	86.1	91.1	97.3	85.1	90.6	92.8±0.2

approaches that utilize additional anomalous data for training, Transductive Learning Methods that infer from the extensive portion of the test set, and Inductive Learning Approaches that assess each query independently, without prior knowledge of the overall distribution. Our method, an exemplar of the inductive approach, outperforms Auxiliary Data Approaches without requiring exposure to anomalies during training.

Contrasting our method with MuSc, the state-of-the-art transductive learning setting of ZSAD, MuSc requires accessing the entire test set distribution before inference, making it highly dependent on the data distribution, and potentially limiting in real-world applications like online real-time inference. In Tab. 4.1, our method surpasses MuSc's performance on the pair-image setting, especially in anomaly classification. This derives advantages from the utilization of the class token in CLIP embeddings, and emphasizes the efficiency and robustness of our joint language-vision prediction.

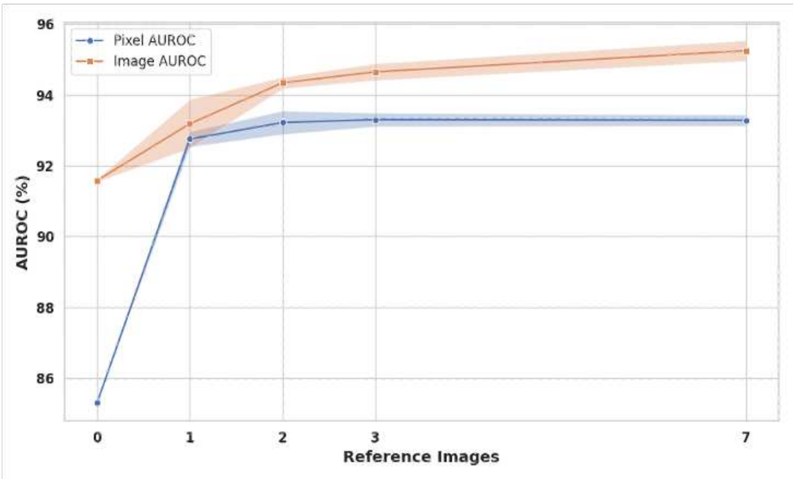


Figure 4.5: AUROC on MVTecAD with an increasing number of reference images.

4.4 Ablation Studies

4.4.1 Reference Images Quantity

From Fig. 4.5, a clear trend is observed: with the increasing number of reference images, the anomaly detection performance improves, reflected from both pixel-level and sample-level AUROC. This supports the hypothesis that unlabeld images can still offer valuable comparative references for anomaly identification.

Significantly, the most pronounced performance leap occurs when the number of reference images is increased from 0 to 1, indicating that even a single reference image can substantially improve the model's AD ability. However, the subsequent performance gains from 1 to 7 reference images are present but exhibit diminishing returns. This trend implies that considering the factors of inference time and memory efficiency, utilizing a large pool of reference images might not always be feasible or optimal, particularly in real-world applications where access to a wide array of suitable samples is often limited. In this context, a pairwise approach emerges as a balanced solution, optimizing the trade-off between improved detection performance and computational resource efficiency. Moreover, the observed differences in the degree of improvement between pixel AUROC and sample AUROC point that while visual

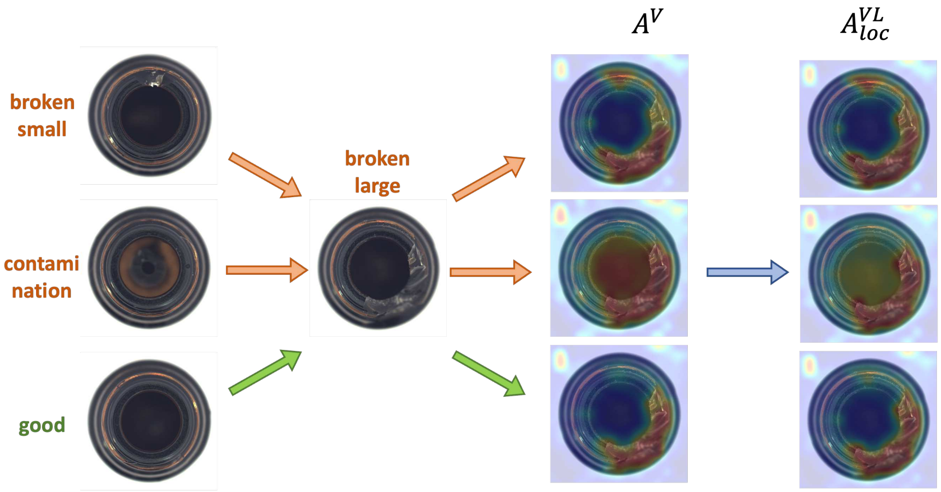


Figure 4.6: Impact of Reference Image Selection, illustrating variance in anomaly score on the choice of normal samples and various anomalous samples.

details are paramount in pinpointing anomalies, they may be less influential in the broader context of classifying an entire sample as anomalous.

4.4.2 The Choice of Pairing Samples

Since the reference images are randomly sampled from unlabeld images, they can either be normal or anomalous. This leads to the question: how does an anomalous reference image affect the precision of anomaly detection?

Fig. 4.6 illustrates how different reference images can significantly influence the anomaly score. In the case of the "bottle", it is evident that using a normal reference image generally guarantees the accuracy of anomaly detection, as it provides a clear baseline for identifying outliers. Conversely, when the reference image contains anomalies, such as breaks or contamination, these imperfections can misleadingly provide a false reference for the query image, erroneously highlighting the reference image's damaged region in the query. This phenomenon suggests that the abnormal condition of the reference image can "pollute" the anomaly score of the query image.

Interestingly, integrating textual cues with visual data can mitigate this negative effect. By leveraging textual features from prompts, the model can effectively counter the false prediction associated with the anomalous refer-

Table 4.3: Ablation studies on the TTA module.

	$+A^V$	+TTA	A.L.			A.C		
			AUROC	F1Max	PRO	AUROC	F1Max	AP
base.			85.3 ± 0.0	29.1 ± 0.1	71.8 ± 0.4	91.6 ± 0.0	92.9 ± 0.1	96.4 ± 0.0
		✓	88.7 ± 0.2	35.6 ± 0.2	80.1 ± 0.4	92.1 ± 0.3	93.1 ± 0.2	96.6 ± 0.1
	✓		92.6 ± 0.2	41.8 ± 0.8	82.6 ± 0.4	93.1 ± 0.6	94.0 ± 0.2	96.6 ± 0.3
	✓	✓	**92.8±0.2**	**42.4±0.7**	**84.0±0.4**	**93.2±0.8**	**94.1±0.2**	**96.7±0.4**

ences. As depicted in Fig. 4.6, the joint predictions that combine both visual and language information exhibit a notable increase in accuracy, underscoring the potential of language-vision joint anomaly detection.

4.4.3 Test-Time Adaption Module

We also studied the impact of various training steps on model performance, Fig. 4.8 demonstrates the AUROC and PRO for training steps ranging from 1 to 6. We can see both the pixel and sample AUROC and PRO scores reach the optimal when the training step is set to 2, and start to decrease. Therefore, we opted for a training step = 2. In Tab. 4.3, we showcase the performance enhancements by the TTA module. Here we take the text-only approach as the baseline. The table shows that implementing the TTA module on the baseline yields a notable increase in performance. When the TTA module operates alongside a paired reference image, the results are further amplified. In this scenario, the AUROC for AL climbs to 92.8%, and AC reaches 93.2%. Further, the improvement in the F1Max and PRO indicates a more balanced and effective model, particularly in terms of its localization capabilities. The influence of the TTA module is further presented in Fig. 4.7, where a marked distinction in anomaly scores between normal and anomalous patches is observed post-adaptation. Prior to adaptation, the “missing cable” region was not adequately identified, and the adaptation process leads to a refined alignment between visual perception and language context, resulting in superior AL and AC performance.

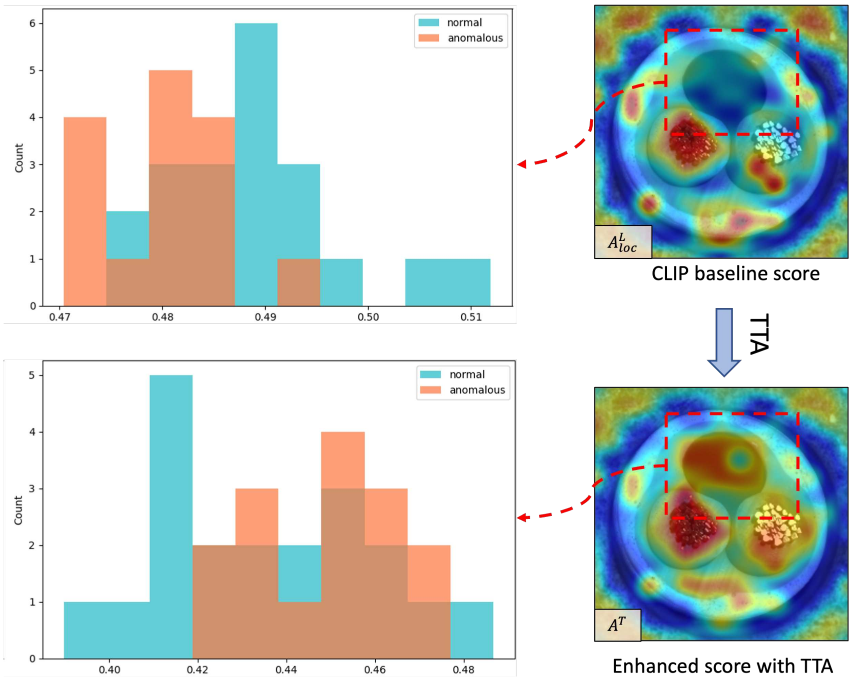


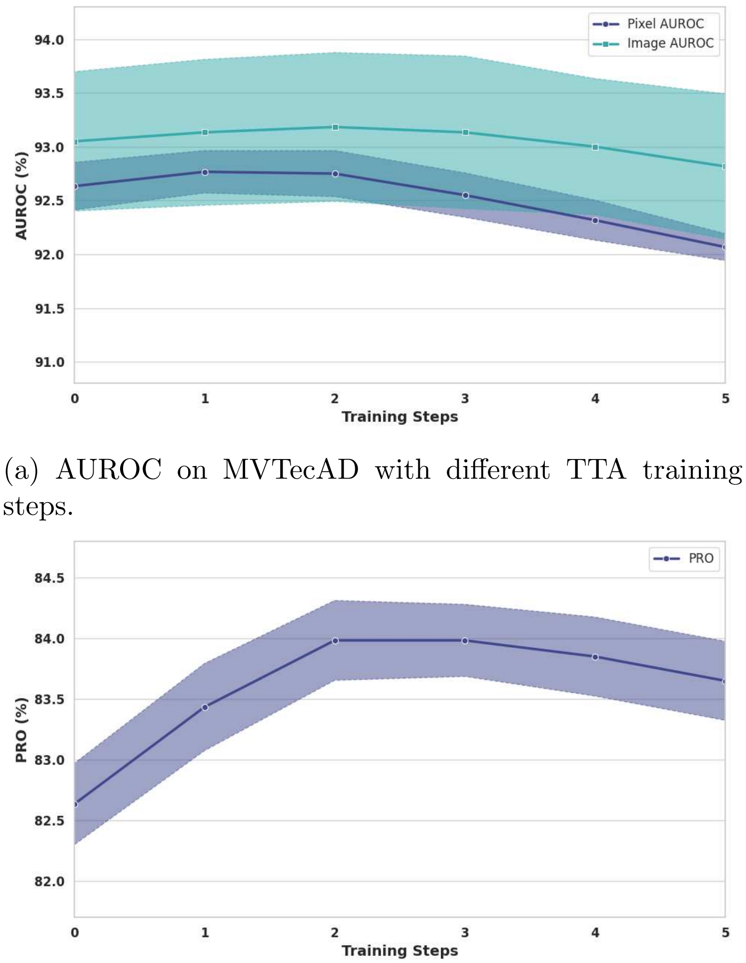
Figure 4.7: Left: Histogram of the TTA anomaly score from the highlighted red box region. Right: Heatmap of the anomaly score for “missing cable” before and after adaptation.

4.4.4 Ablation Study on Hyperparameters

We conducted a comprehensive performance comparison across various settings of hyperparameters λ_1 through λ_5. These experiments were executed using 6 different random seeds, and we report the results as mean values with standard deviations to provide a clear understanding of variability and reliability.

Table 4.4: Comparative study of the zero-shot anomaly localization (AL) performance on MVTecAD with various λ_1 and λ_2 settings. Bold values indicate the best results.

λ_1	λ_2	AUROC	F1Max	PRO
1	0.5	92.5 ± 0.2	41.7 ± 0.8	82.2 ± 0.3
1	1	92.7 ± 0.2	**42.4±0.7**	83.4 ± 0.4
1	1.5	**92.8±0.2**	**42.4±0.7**	84.0 ± 0.4
2	2	92.7 ± 0.2	42.3 ± 0.6	**84.1±0.3**



(a) AUROC on MVTecAD with different TTA training steps.

(b) PRO on MVTecAD with different TTA training steps.

Figure 4.8: Ablation studies on test-time adaptation.

4.5 Limitations and Conclusion

4.5.1 Limitations & Future work

Our approach, while robust in many scenarios, is not without its limitations. One notable constraint is the requirement of inputting two images during inference, which may not be feasible with certain scenarios where single-image processing is crucial. Despite this, our method still demonstrates a significant performance enhancement in most cases.

Moreover, while our framework achieves SOTA performance in the zero-shot inductive learning setting, it reveals a gap when compared to SOTA models trained under full-shot regimes and zero-shot transductive learning

Table 4.5: Comparative study of the zero-shot anomaly classification (AC) performance on MVTecAD with various λ_3, λ_4, and λ_5 settings. Bold values indicate the best results.

λ_3	λ_4	λ_5	AUROC	F1Max	AP
1	1	1	**93.2±0.8**	**94.1±0.2**	96.7 ± 0.4
1	1	2	93.1 ± 0.8	94.0 ± 0.3	**96.8±0.4**
1	2	1	91.9 ± 0.8	93.2 ± 0.2	94.8 ± 0.4
2	1	1	93.1 ± 0.6	94.0 ± 0.1	96.7 ± 0.3

approaches. As Tab. 4.2 shows, our method outperforms many existing models in unified AC and AL. However, it falls short of the benchmarks set by UniAD [81] and MuSc [44], particularly in scenarios where MuSc excels using visual features alone. This discrepancy suggests that there is substantial untapped potential for further exploration of visual reference features.

Additionally, our study offers novel insights into the application of the CLIP model for fine-grained anomaly detection: We demonstrate that joint visual and textual discrimination is a key contributor to enhancing fine-grained anomaly localization capabilities within the CLIP framework. Our findings also indicate that even when the visual reference images are anomalous, they can still serve as references for accurate anomaly scoring. These insights not only affirm the effectiveness of our proposed method but also open avenues for future research in refining visual-language models for more precise and versatile anomaly detection tasks.

4.5.2 Conclusion

In this chapter, we introduced an innovative framework, the Dual-Image Enhanced CLIP for anomaly classification and localization, in the realm of zero-shot learning. Our approach leverages pairs of unlabeld images utilizes the pseudo anomaly in the TTA module, and demonstrates remarkable enhancement in performance, outperforming several SOTA methods. This advancement was achieved without the need for additional training, showcasing the framework's practicality and efficiency. Our findings also highlighted the untapped potential in combining textual features and visual references, suggesting room for further exploration in this domain.

Chapter 5

Conclusion & Future Work

5.1 Conclusions

Anomaly detection has seen significant advancements in recent years with a trend of transitioning from unsupervised methods to multi-class and zero-shot scenarios, and developing specialized solution in distinct domains. This thesis showcases the strategic use of auxiliary information as a strategy to overcome the limitations of traditional models in these novel contexts, demonstrating how auxiliary data can enhance AD capabilities in varied and complex tasks.

The first part focus on Liver Tumor Segmentation for medical anomaly detection, we propose a random-shape anomaly synthesis algorithm and a two-stage training strategy to address the performance variability during training. Our findings suggest that a discriminative model should not be overly trained on synthetic anomalies to preserve generalizability. This work establishes a framework for generating and training with pseudo anomalies, and demonstrates substantial improvements over baseline methods.

In the second part, we present Dual-Image Enhanced CLIP, a method that leverages both visual and textual information to predict anomalies. By using pairs of images where one serves as the visual reference for the other, our method significantly leverages the hidden normality in unlabeld images, and improves accuracy over existing SOTA ZSAD methods.

5.2 Future Works

The increasing interest in zero-shot anomaly detection opens up new avenues for incorporating large Vision-Language models like Flamingo [5] into AD. These models, with their unique semantic-vision alignment capabilities, hold promise for identifying anomalous patterns. Our research demonstrates the effectiveness of integrating visual and textual features, utilizing auxiliary information to enhance anomaly localization within the CLIP framework. This insight lays the groundwork for future research aimed at refining vision-language models for more precise AD tasks.

In the medical domain, with the advent of medical vision-language models [75], there is a vast potential for applying these models to medical AD. Our future endeavours will extend to various diseases and data modalities, incorporating both real and synthetic tumors within the training pipeline, pushing forward the capabilities of medical AD. This approach harnesses the potential of auxiliary data to improve model performance in complex medical scenarios.

References

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

[2] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, *Ganomaly: Semi-supervised anomaly detection via adversarial training*, 2018. arXiv: `1805.06725 [cs.CV]`.

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

[4] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, *Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection*, 2019. arXiv: `1901.08954 [cs.CV]`.

[5] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, *Flamingo: A visual language model for few-shot learning*, 2022. arXiv: `2204.14198 [cs.CV]`.

[6] S. Bakas, M. Reyes, A. Jakab, *et al.*, *Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge*, 2019. arXiv: `1811.02629 [cs.CV]`.

[7] 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.

[8] P. Bilic, P. F. Christ, E. Vorontsov, *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[9] R. Chalapathy, E. Z. Borzeshi, and M. Piccardi, *An investigation of recurrent neural architectures for drug name recognition*, 2016. arXiv: `1609.07585 [cs.CL]`.

[10] X. Chen, Y. Han, and J. Zhang, *April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad*, 2023. arXiv: `2305.17382 [cs.CV]`.

[11] Y.-C. Chen, L. Li, L. Yu, *et al.*, *Uniter: Universal image-text representation learning*, 2020. arXiv: `1909.11740 [cs.CV]`.

[12] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[13] N. Cohen and Y. Hoshen, *Sub-image anomaly detection with deep pyramid correspondences*, 2021. arXiv: `2005.02357 [cs.CV]`.

[14] 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.

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

[16] 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.

[17] 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.

[18] H. Deng and X. Li, "Self-supervised anomaly detection with random-shape pseudo-outliers," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 4768–4772.

[19] H. Deng and X. Li, *Structural teacher-student normality learning for multi-class anomaly detection and localization*, 2024. arXiv: `2402.17091 [cs.CV]`.

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

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[22] K. Desai and J. Johnson, *Virtex: Learning visual representations from textual annotations*, 2021. arXiv: `2006.06666 [cs.CV]`.

[23] R. Dey and Y. 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.

[24] X. Dong, J. Bao, Y. Zheng, *et al.*, *Maskclip: Masked self-distillation advances contrastive language-image pretraining*, 2023. arXiv: 2208.12262 [cs.CV].

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV].

[26] P. Gao, S. Geng, R. Zhang, *et al.*, *Clip-adapter: Better vision-language models with feature adapters*, 2021. arXiv: 2110.04544 [cs.CV].

[27] D. Gong, L. Liu, V. Le, *et al.*, *Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection*, 2019. arXiv: 1904.02639 [cs.CV].

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[29] D. Gudovskiy, S. Ishizaka, and K. Kozuka, *Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows*, 2021. arXiv: 2107.12571 [cs.CV].

[30] Q. Hu, Y. Chen, J. Xiao, *et al.*, "Label-free liver tumor segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7422–7432.

[31] G. Ilharco, M. Wortsman, R. Wightman, *et al.*, *Openclip*, version 0.1, If you use this software, please cite it as below., Jul. 2021. DOI: 10.5281/zenodo.5143773. [Online]. Available: https://doi.org/10.5281/zenodo.5143773.

[32] 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.

[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*, 2021.

[34] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, *Maple: Multi-modal prompt learning*, 2023. arXiv: 2210.03117 [cs.CV].

[35] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, *Self-regulating prompts: Foundational model adaptation without forgetting*, 2023. arXiv: 2307.06948 [cs.CV].

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

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

[38] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: `1412.6980 [cs.LG]`.

[39] 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.

[40] K. Lee, M. O’Malley, M. Haider, and A. Hanbidge, “Triple-phase mdct of hepatocellular carcinoma,” *American Journal of Roentgenology*, vol. 182, no. 3, pp. 643–649, 2004.

[41] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, *Cutpaste: Self-supervised learning for anomaly detection and localization*, 2021. arXiv: `2104.04015 [cs.CV]`.

[42] 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.

[43] 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.

[44] X. Li, Z. Huang, F. Xue, and Y. Zhou, “Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images,” *arXiv preprint arXiv:2401.16753*, 2024.

[45] Y. Li, H. Wang, Y. Duan, and X. Li, *Clip surgery for better explainability with enhancement in open-vocabulary tasks*, 2023. arXiv: `2304.05653 [cs.CV]`.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[47] W. Liu, W. Luo, D. Lian, and S. Gao, *Future frame prediction for anomaly detection – a new baseline*, 2018. arXiv: `1712.09867 [cs.CV]`.

[48] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727. DOI: `10.1109/ICCV.2013.338`.

[49] F. Michels, N. Adaloglou, T. Kaiser, and M. Kollmann, “Contrastive language-image pretrained (clip) models are powerful out-of-distribution detectors,” *arXiv preprint arXiv:2303.05828*, 2023.

[50] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[51] K. Perlin, "An image synthesizer," *ACM Siggraph Computer Graphics*, vol. 19, no. 3, pp. 287–296, 1985.

[52] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille, "Micro-batch training with batch-channel normalization and weight standardization," *arXiv preprint arXiv:1903.10520*, 2019.

[53] Z. Qin, H. Yi, Q. Lao, and K. Li, *Medical image understanding with pretrained vision language models: A comprehensive study*, 2023. arXiv: `2209.15517 [cs.CV]`.

[54] A. Radford, J. W. Kim, C. Hallacy, *et al.*, *Learning transferable visual models from natural language supervision*, 2021. arXiv: `2103.00020 [cs.CV]`.

[55] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[57] 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.

[58] M. Rudolph, B. Wandt, and B. Rosenhahn, *Same same but differnet: Semi-supervised defect detection with normalizing flows*, 2020. arXiv: `2008.12577 [cs.CV]`.

[59] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, *Fully convolutional cross-scale-flows for image-based defect detection*, 2021. arXiv: `2110.02855 [cs.CV]`.

[60] L. Ruff, R. Vandermeulen, N. Goernitz, *et al.*, "Deep one-class classification," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 4393–4402. [Online]. Available: `https://proceedings.mlr.press/v80/ruff18a.html`.

[61] 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.

[62] 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.

[63] 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.

[64] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, *Natural synthetic anomalies for self-supervised anomaly detection and localization*, 2022. arXiv: `2109.15222 [cs.CV]`.

[65] 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.

[66] P. Seeböck, S. Waldstein, S. Klimscha, *et al.*, *Identifying and categorizing anomalies in retinal imaging data*, 2016. arXiv: `1612.00686 [cs.LG]`.

[67] M. Tailanian, Á. Pardo, and P. Musé, *U-flow: A u-shaped normalizing flow for anomaly detection with unsupervised threshold*, 2023. arXiv: `2211.12353 [cs.CV]`.

[68] M. Tamura, *Random word data augmentation with clip for zero-shot anomaly detection*, 2023. arXiv: `2308.11119 [cs.CV]`.

[69] J. Tan, B. Hou, J. Batten, H. Qiu, and B. Kainz, “Detecting outliers with foreign patch interpolation,” *arXiv preprint arXiv:2011.04197*, 2020.

[70] J. Tan, B. Hou, T. Day, J. Simpson, D. Rueckert, and B. Kainz, “Detecting outliers with poisson image interpolation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 581–591.

[71] D. M. Tax and R. P. Duin, “Support vector data description,” *Machine learning*, vol. 54, pp. 45–66, 2004.

[72] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: `1706.03762 [cs.CL]`.

[73] Y. Vinker, E. Pajouheshgar, J. Y. Bo, *et al.*, *Clipasso: Semantically-aware object sketching*, 2022. arXiv: `2202.05822 [cs.GR]`.

[74] M. Wang, J. Li, Z. Li, *et al.*, “Unsupervised anomaly detection with local-sensitive vqvae and global-sensitive transformers,” *arXiv:2303.17505*, 2023.

[75] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, *Medclip: Contrastive learning from unpaired medical images and text*, 2022. arXiv: `2210.10163 [cs.CV]`.

[76] 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*, 2022, pp. 35–45.

[77] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[78] 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 Conference on Computer Vision and Pattern Recognition*, 2022, pp. 650–656.

[79] Q. Yao, L. Xiao, P. Liu, and S. K. Zhou, "Label-free segmentation of covid-19 lesions in lung ct," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2808–2819, 2021.

[80] 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.

[81] Z. You, L. Cui, Y. Shen, *et al.*, *A unified model for multi-class anomaly detection*, 2022. arXiv: `2206.03687 [cs.CV]`.

[82] J. Yu, Y. Zheng, X. Wang, *et al.*, *Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows*, 2021. arXiv: `2111.07677 [cs.CV]`.

[83] 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.

[84] 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.

[85] R. Zhang, R. Fang, W. Zhang, *et al.*, *Tip-adapter: Training-free clip-adapter for better vision-language modeling*, 2021. arXiv: `2111.03930 [cs.CV]`.

[86] X. Zhang, W. Xie, C. Huang, Y. Zhang, and Y. Wang, "Self-supervised tumor segmentation through layer decomposition," *arXiv preprint arXiv:2109.03230*, 2021.

[87] X. Zhang, S. Yang, X. Zhang, W. Zhang, and J. Zhang, *Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning*, 2018. arXiv: `1805.10620 [cs.CV]`.

[88] Y. Zhao, "Omnial: A unified cnn framework for unsupervised anomaly localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3924–3933.

[89] C. Zhou, C. C. Loy, and B. Dai, *Extract free dense labels from clip*, 2022. arXiv: `2112.01071 [cs.CV]`.

[90] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, *Conditional prompt learning for vision-language models*, 2022. arXiv: `2203.05557 [cs.CV]`.

[91] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, Jul. 2022, ISSN: 1573-1405. DOI: `10.1007/s11263-022-01653-1`. [Online]. Available: `http://dx.doi.org/10.1007/s11263-022-01653-1`.

[92] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, *Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection*, 2023. arXiv: `2310.18961 [cs.CV]`.

[93] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 289–297.

[94] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*, Springer, 2022, pp. 392–408.

Appendix

A.1 Supplementary Materials for Chatper 3

A.1.1 Addtional Results on AUROC

Table A.1: Lesion detection performance as evaluated by the sample-level AUROC score, and Dice upper bound ⌈Dice⌉. The best results are highlighted in **bold**. And results marked with an * indicate the Dice value instead of the Dice upper bound. Results marked with ** are directly copied from the original paper.

Method	**Sample AUROC**	**⌈Dice⌉**
Padim [15]	60.87 ± 0.60	7.07 ± 0.25
Cut&Paste [42]	58.22 ± 1.92	–
Ganomaly [1]	68.93 ± 0.48	–
Patchcore [57]	70.66 ± 0.45	18.54 ± 1.15
Reverse Distillation [17]	70.08 ± 1.61	17.58 ± 2.51
Li et al [43]	**86.1****	-
Ours	75.51 ± 1.40	$\mathbf{53.03 \pm 1.78}$*

A.2 Supplementary Materials for Chatper 4

A.2.1 Prompt Templates

We follows AnoCLIP [20] to produce prompts descriptions. It's composed of base templates, descriptive state words, and domain-aware prompts, denoted as following: "[c]" represents each class category; "[s]" denotes the state prompts; "[d]" is the domain-aware prompts. By systematically substituting "[s]", "[d]", and "[c]" into the base templates, we generate a diverse array of prompts. These prompts effectively encompass both normal and anomalous scenarios within their respective domains.

- **Base Templates**
  - “a [d] cropped photo of the [s]”
  - “a [d] cropped photo of a [s]”
  - “a [d] close-up photo of a [s]”
  - “a [d] close-up photo of the [s]”
  - “a bright [d] photo of a [s]”
  - “a bright [d] photo of the [s]”
  - “a dark [d] photo of the [s]”
  - “a dark [d] photo of a [s]”
  - “a jpeg corrupted [d] photo of a [s]”
  - “a jpeg corrupted [d] photo of the [s]”
  - “a blurry [d] photo of the [s]”
  - “a blurry [d] photo of a [s]”
  - “a [d] photo of a [s]”
  - “a [d] photo of the [s]”
  - “a [d] photo of a small [s]”
  - “a [d] photo of the small [s]”
  - “a [d] photo of a large [s]”
  - “a [d] photo of the large [s]”
  - “a [d] photo of the [s] for visual inspection”
  - “a [d] photo of a [s] for visual inspection”
  - “a [d] photo of the [s] for anomaly detection”
  - “a [d] photo of a [s] for anomaly detection”
- **Descriptive State Words**

  normal states:

  - s := “normal [c]”

  - s := “unblemished [c]”
  - s := “flawless [c]”
  - s := “perfect [c]”
  - s := “[c] without flaw”
  - s := “[c] without damage”
  - s := “[c] without defect”

  abnormal states:

  - s := “damaged [c]”
  - s := “abnormal [c]”
  - s := “imperfect [c]”
  - s := “blemished [c]”
  - s := “[c] with flaw”
  - s := “[c] with damage”
  - s := “[c] with defect”

- **Domain Prompts**
  - For all categories:
    - d := “industrial”
  - For surface categories (carpet, leather, grid, tile, wood):
    - d := “textural”
    - d := “surface”
  - For all other categories:
    - d := “manufacturing”

A.2.2 Visualizations of Anomaly Localization Results

We visualize the zero-shot anomaly detection results on MVTecAD in Fig. A.1 and Fig. A.2 and VisA in Fig. A.3 and Fig. A.4. Notably, the segmentation threshold is selected based on the max value of the F1 score. Also, our

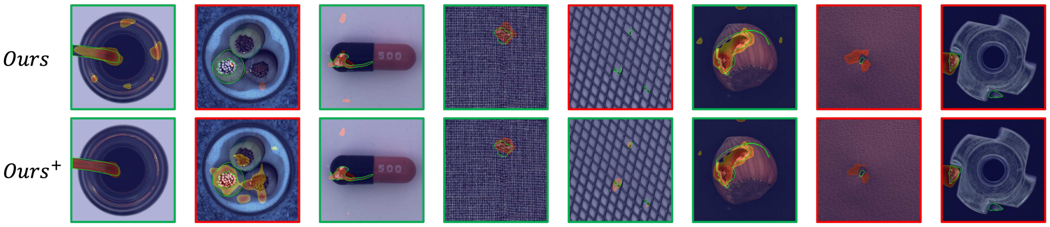


Figure A.1: Visualization of prediction examples from “bottle”, “cable”, “capsule”, “carpet”, “grid”, “hazelnut”, “leather”, and “metalnut” categories. Green line in the images denotes the ground truth of the anomaly. The success and failure cases are bordered with red and green, respectively.

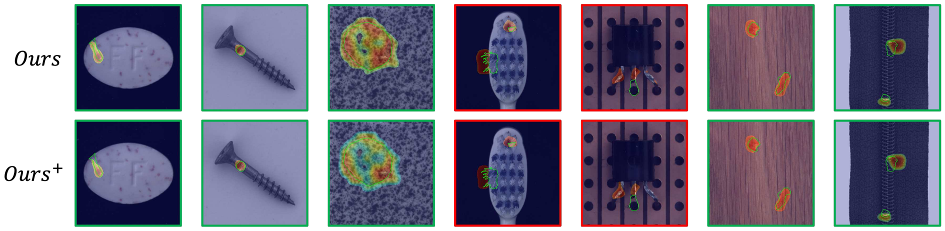


Figure A.2: Visualization of prediction examples from “pill”, “screw”, “toothbrush”, “transistor”, and “zipper” categories. Green line in the images denotes the ground truth of the anomaly. The success and failure cases are bordered with red and green, respectively.

TTA enhanced method which is denoted as $ours^+$, yields more accurate and comprehensive results, even in difficult scenarios and failure cases.

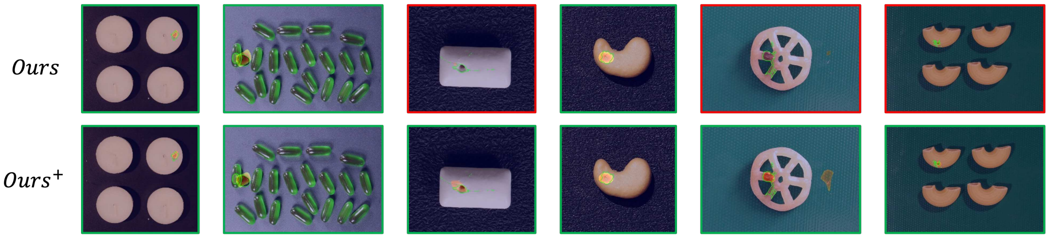


Figure A.3: Visualization of prediction examples from “candle”, “capsules”, “chewing gum”, “cashew”, “fryum”, and “macaroni1” categories. Green line in the images denotes the ground truth of the anomaly. The success and failure cases are bordered with red and green, respectively.

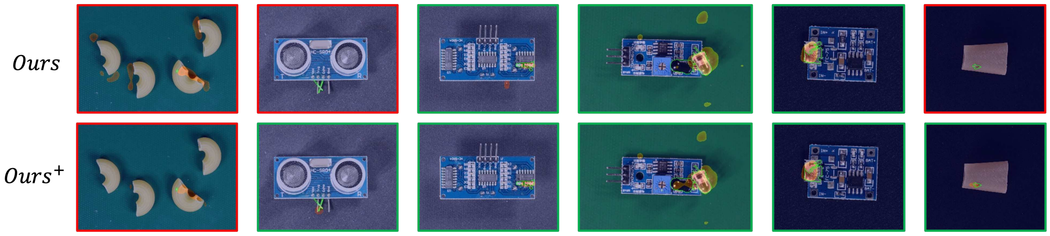


Figure A.4: Visualization of prediction examples from “macaroni2”, “pcb1”, “pcb2”, “pcb3”, “pcb4”, and “pipe fryum” categories. Green line in the images denotes the ground truth of the anomaly. The success and failure cases are bordered with red and green, respectively.