Zero-Shot AnomalyDetection in Industrial Manufacturing UsingVision Transformers and ConditionalDiffusion Models

Submitted By:

Md MahfuzurRahmanShanto
221-35-917
Department of Software Engineering
Daffodil International University

Supervised By:

Mr. MdRajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Daffodil International University

Bachelorof Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**Zero-Shot Anomaly Detection in Industrial Manufacturing Using Vision Transformers and Conditional Diffusion Models**", submitted by **Md Mahfuzur Rahman Shanto (ID: 221-35-917)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Fazla Ealhe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Mohammod Abul Kashem, PhD
Professor
Department of Computer Science and Engineering
DUET, Bangladesh

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Md Mahfuzur Rahman Shanto
Date of Birth	:	24th December 2001
Title	:	Zero-Shot Anomaly Detection in Industrial Manufacturing Using Vision Transformers and Conditional Diffusion Models
Academic Session	:	Spring 2022 – Fall 2025

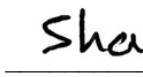
I declare that this thesis is classified as:

☒	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☒	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

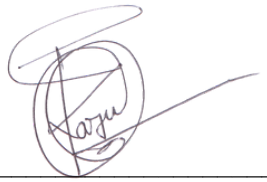
Certified by:



(Student's Signature)

ID: 221-35-917

Date: 28 /12/ 2025



(Supervisor's Signature)

Mr. Md Rajib Mia

Date: 28/12/2025

NOTE: * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City, Ashulia, Dhaka, Bangladesh

Dear Sir,

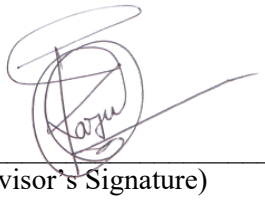
CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name	Md Mahfuzur Rahman Shanto, Md Rajib Mia	
Thesis Title	Zero-Shot Anomaly Detection in Industrial Manufacturing Using Vision Transformers and Conditional Diffusion Models	
Reasons	(i)	This thesis contains original research and findings that are intended for submission to high-impact journals.
	(ii)	Public availability at this stage may violate the "Prior Publication" policy of the intended publishers.
	(iii)	To protect the novelty of the methodology for future extension and intellectual property rights.

Thank you.

Yours faithfully,



_____
(Supervisor's Signature)

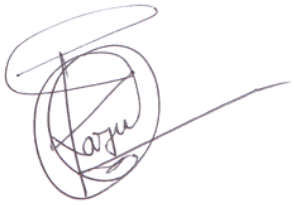
Date: 28/12/2025

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian, *Daffodil International University* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

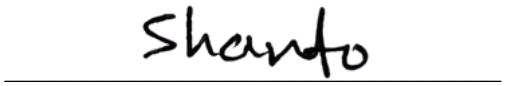
Full Name : Mr. Md Rajib Mia

Position : Lecturer (Senior Scale)

Date : 28/12/2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name : Md Mahfuzur Rahman Shanto

ID Number : 221-35-917

Date : 28/12/2025

Acknowledgments

First and foremost, I express my deepest gratitude to Almighty Allah for giving me the strength, patience, and perseverance to complete this research work.

I would like to express my sincere gratitude to my supervisor, **Mr. Md Rajib Mia**, Lecturer(Senior Scale), Department of Software Engineering, Daffodil International University, for his invaluable guidance and continuous support throughout this journey.

I am thankful to the **Prof. Dr. Imran Mahmud**, the department head of software engineering, Daffodil international university, who has allowed me the facilities and ensured suitable environment to conduct research.

I wish to mention that my experience as an undergraduate student in the Department of Software Engineering was highly supportive and filled with great suggestions by most of the faculty members. I would also like to give a special credit to my course instructors who helped me establish the background knowledge that was necessary to carry out this research.

I would like to say my warmest gratitude to my family who love, support, and are patient with me without any conditions. They have always encouraged and believed in my capabilities which has been a source of motivation.

Lastly, I would like to mention the contributors of the MVTec Anomaly Detection dataset and the pre-trained models that have been utilized in this work have proved to be invaluable.

Md Mahfuzur Rahman Shanto
November 2025

Abstract

The automation of anomaly detection systems that do not require large amounts of labeled data are required in the industrial manufacturing quality control to detect defects. This thesis introduces a new multi-path zero-shot anomaly detection architecture that concurrently combines the use of Vision Transformer (ViT)-based memory banks, conditional diffusion models and CLIP-based zero-shot detection in a manufacturing industry setting. In contrast to the current methods that apply single detection strategies, the proposed framework integrates three complementary directions: Path A is implemented with the help of hierarchical ViT features (block 5 and 11) and k-nearest neighbor memory banks, which are applied to detect textural anomalies efficiently; Path B is applied using conditional diffusion models as guided by ViT features, which is used to identify structural anomalies via reconstruction; and Path C is applied to semantic zero-shot with the help of CLIP vision-language understanding with no training information.

Extensive testing of 10 classes of the MVTec Anomaly Detection dataset show outstanding results with images average AUROC of 96.41%, Pixel average AUROC of 96.75 and Image AUPR of 98.30. The framework favors the zero-shot approaches (WinCLIP (91.8%), AnomalyCLIP (91.5%), AnoVL (92.5%), and DZAD (93.5%)) by 2.9-4.9 percentage points, and achieves similar performance to the state-of-the-art supervised models with-out using category-specific training examples. Design selection systematic ablation experiments support the design selection and demonstrate that hierarchical feature extraction yields superior detection by 2.18 percent compared to single-layer methods, and multi-path integration affords robustness to di-verse defects. The framework attains real-time inference of 67 FPS (15ms per image) on the main detection path, so it can be applied in the real-life industry.

The study has three main contributions namely, (1) the first multi-path zero-shot model that involves ViT, diffusion models as well as CLIP to detect anomalies in industries; (2) technical contributions such as hierarchical feature extraction, ViT-conditioned diffusion, and patch-level CLIP analysis; (3) extensive evaluation that shows that the model is competitive with supervised methods and zero-shot models do not need training data. The zero-shot capability offers significant practical advantages for manufacturers deploying quality control systems across numerous product variants, reducing data collection costs and enabling immediate deployment to new production lines.

Keywords: Zero-Shot Learning, Anomaly Detection, Vision Transformers, Diffusion Models, CLIP, Industrial Quality Control, Manufacturing, Deep Learning, Multi-Path Detection

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Form
AD	Anomaly Detection
AE	Autoencoder
AI	Artificial Intelligence
AUPR	Area Under Precision-Recall Curve
AUROC	Area Under Receiver Operating Characteristic Curve
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
DIU	Daffodil International University
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FPS	Frames Per Second
GAN	Generative Adversarial Network
LDM	Latent Diffusion Model
MAE	Masked Auto-Encoder
ML	Machine Learning
MSA	Multi-head Self-Attention
MVTec AD	MVTec Anomaly Detection Dataset
PRO	Per-Region Overlap
ResNet	Residual Network
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
SAM	Segment Anything Model
SSD	Single Shot MultiBox Detector
TN	True Negative
TP	True Positive
TPR	True Positive Rate
U-Net	U-shaped Network
VAE	Variational Auto-Encoder
ViT	Vision Transformer
ZSAD	Zero-Shot Anomaly Detection

Chapter 1

Introduction

1.1 Background and Motivation

The control of quality in manufacturing requires accuracy and reduced costs and time inspection. Conventional visual inspection relies on human skills. This adds subjectivity and does not scale with volumes of production. The automatic detection systems of anomalies allow detecting defects in real time and minimizing the human factor (Cao et al., 2023). Majority of automated systems are based on supervised learning that needs large sized dataset of normal and defective samples which are labelled. Existence in industries poses a quandary. Lack on defective samples: Defective samples are rare and expensive to produce and are dangerous to understand in other cases (Bergmann et al., 2022).

The production process is paradoxical. The quality systems need to identify infrequent flaws. During supervised models, hundreds or thousands of labeled anomalies are required to be trained. Gathering of such data interferes with work processes and is also very costly. New defects that have not been realized cannot be gathered. Such limitations have led to a resurgence of interest in unsupervised and zero-shot detection approaches which are only trained on normal samples (Roth et al., 2022).

Powerful visual representation models were introduced as a result of deep learning. ViTs learn both local and global image features by using self-attention techniques (Dosovitskiy et al., 2021). Images are processed by CNNs using hierarchical receptive fields that are local. ViTs partition images into patches and directly learn long-range dependencies. This is useful in the anomaly detection where defects occur in different sizes. DINO is a task-free, pre-trained ViT model that offers rich representations (Caron et al., 2021).

The diffusion models changed the nature of generative modeling by training to remove random noise to produce realistic samples (Ho et al., 2020). They are very good at learning complicated data patterns. Diffusion models to detect anomalies are trained

to have the distribution of normal samples and detect anomalies by reconstruction error. Semantic feature conditioning. Invests in a vulnerably owned property. They can reconstruct normal images with high resilience but have problems with the anomalous regions, which offer natural detection (Song et al., 2021).

Zero-shot visual understanding is made possible by vision-language models such as CLIP (Radford et al., 2021). CLIP acquires similar representations between text descriptions and images. Models are able to identify visual concepts by using prompts in natural language and without training examples. Zero-shot anomaly detection is based on the text description of normal and abnormal states (Deng and Zhang, 2023).

Both methods have their limitations. ViT-based memory bank approaches can perform well but are unable to deal with novel defects that are poorly represented in their feature space (Defard et al., 2021). The methods based on diffusion are able to reconstruct details, but require high computational resources to run the inference (Wu et al., 2024). Zero-shot methods that are based on CLIP are extremely reliant on the quality of text prompts. It is difficult to create effective prompts to subtle industrial defects (Jeong et al., 2023). One method is not superior to the others in terms of defects and industrial conditions.

1.2 Problem Statement

The existing manufacturing anomaly detection systems have numerous problems. Controlled approaches involve large labeled sets of defective samples which are costly, time-consuming, and in some cases, impossible to obtain. The zero-shot methods that have been developed so far are based on single detectors, which cannot address some form of defects. Each of texture defects, structural abnormality and minor deviation require various detection strategies. High precision and high performance are needed in industrial applications. A lot of high-tech approaches compromise the other.

What can we say about creating a strong zero-shot anomaly detector which is highly accurate both in terms of the various type of defects and is still practical in terms of its computational needs? The answer requires integration of complementary methods of detection within one system where each system capitalizes on its advantages and offsets its disadvantages with other systems.

1.3 Research Objectives

In this study, a multi-path zero-shot anomaly detector in the quality control of industrial manufacturing is developed:

Objective 1: Develop a hierarchical Vision Transformer feature detection system

that makes use of local texture features and global structural features of images of industrial products.

Objective 2: Train a conditional diffusion model on learning the appearance distribution of normal products conditioned with anomalies by reconstruction analysis based on ViT features.

Objective 3: Learn a CLIP based zero-shot detecting path that uses vision-language alignment to understand semantic anomalies without any training data.

Objective 4: Develop a multi-path fusion system that will choose or merge detection paths depending on their capabilities on various defects.

Objective 5: Compare the framework on the MVTec Anomaly Detection dataset on various industrial products and benchmark the results on state-of-the-art levels.

Objective 6: Characterize the contribution of constituent components by the ablation experiments and determine when each of the detection paths is superior.

1.4 Research Questions

In this study, the researcher examines five questions:

RQ1: Which multi-layer hierarchical Vision Transformer features are the most effective to identify various types of anomalies in industrial images?

RQ2: Does conditional diffusion models conditioned on ViT features (performing normative reconstruction) and (conditional reconstruction) show reconstruction errors on anomalous regions?

RQ3: Does CLIP vision-language understanding permit text prompt anomaly detection?

RQ4: Which detection methods based on memory bank, diffusion-based, and CLIP are the most effective and the least effective with respect to various types of industrial defects?

RQ5: What is the relative performance of the multi-path ensemble with other approaches in accuracy of detection, localization and computational cost?

1.5 Contributions and Novelty

The work pushes the state of the zero-shot anomaly detection in the field of industrial production:

Novel Multi-Path Architecture: Current solutions are based on single detection mechanisms. The current work suggests the initial model that integrates three mutually supportive directions, which are ViT feature memory banks, conditional diffusion models, and CLIP zero-shot detection. They concern various aspects of detection. When combined, they form a powerful system that becomes responsive to different defects.

Hierarchical Feature Integration: The model derives features of layers 5 and 11 of the Vision Transformer, that is, detailed texture features and the structure patterns at high-level. Multi-scale approach identifies various types of anomalies in form of minor texture irregularities to major structural flaws.

Conditional Diffusion for Industrial AD: Newly Diffusion models have also found uses in anomaly detection. This is a conditioning of the denoising process on ViT features of the same image and it is used in guiding the correct reconstructions of normal samples. It increases discrimination of normal and anomalous patterns.

Comprehensive Evaluation: The experiments confirm the performance in 10 MVTec AD categories such as texture and object types. There are several measures (AUROC, AUPR, PRO score) that evaluate the performance atimage and pixel levels.

Practical Industrial Focus: Numerous academic directions give more importance to theoretical novelty than application. In this framework, the efficiency of the computation and the possibility of deployment are taken into account. The system is competitive in terms of accuracy and the inference times remain reasonable when using the system to control industrial quality.

Ablation Insights: It has been observed that systematic ablation studies can be used to identify the most significant components to performance of various types of defects and thus assist practitioners in choosing the correct configurations to particular applications.

1.6 Scope and Limitations

The study being investigated is on the zero-shot anomaly detection of industrial manufacturing quality control.

Dataset: Experiments are on 10 MVTec AD categories: bottle, cable, metal_nut, pill, screw, tile, toothbrush, transistor, wood, and zipper, which comprise different types of industrial products and defects.

Detection Task: The framework does not only focus on detecting anomalies (whether an image may be considered normal or contains a defect), but also determining the position of the anomalies on a pixel-by-pixel basis (anomaly localization).

Learning Paradigm: Methods are trained by zero-shot: they are trained on normal samples with no defective samples.

Modalities: Research only takes into account the visual information (RGB images) without any other sensor modalities such as depth, thermal, or spectral image information.

The framework is able to perform well on MVTec dataset. The extrapolation to completely different industrial areas needs to be proved. Diffusion-based paths can require considerable amounts of computation, and thus may be a limiting factor to processing

in highly-throughput lines. Path CLIP algorithms rely on the ability to describe defects properly with text, and need domain knowledge to design. These are the limitations that can be resolved in future research.

1.7 Thesis Organization

The rest of this thesis is structured in the following way:

Chapter 2 (Literature Review) reviews the available literature on anomaly detection including traditional methods, deep learning models, Vision Transformer models, diffusion models, and zero-shot learning models. The chapter points to gaps in the research that led to the proposed approach.

Chapter 3 (Methodology) includes the technical architecture that consists of the Vision Trans-former feature extraction, the architecture of the conditional diffusion model, CLIP-based zero-shot detector, memory bank, multi-path fusion strategy, and the entire inference pipeline. Each of the components is supported by mathematical formulations.

Chapter 4 (Experimental Setup) the MVTec AD data, training specifications, training hyperparameters, metrics of evaluation and comparison to baseline methods are outlined.

Chapter 5 (Results and Discussion) is the presentation of experimental research such as quantitative performance measures of all categories, qualitative map of anomalies, comparative study with baselines, and ablation study of the contribution of each component.

Chapter 6 (Conclusion) is a conclusion of major findings and the contribution of research. It also talks about implications in real life applicability in industries, limits and future research directions.

This outline is used to lead the readers through the motivation and background to technical details until experimental validation and conclusions.

Chapter 2

Literature Review

2.1 Introduction to Anomaly Detection

Anomaly detection is used to detect patterns that are far out of the expected behavior. Abnormalities in the case of industrial manufacturing can be in the form of defects, damages or deviation of the quality specifications. It is difficult to find abnormalities with maximum accuracy, and reduce instances of a false alarm that interrupts the production processes. The old methods relied on the traditional elements which were hand-made and statistical techniques. The advancement of deep learning changed the picture since it was now possible to learn the features automatically by using raw data (Chandola et al., 2009).

MVTec Anomaly Detection dataset, proposed by Bergmann et al. in 2019, has become the standard test case to assess the issue of industrial anomaly detection (Bergmann et al., 2019). It consists of 15 categories of industrial products data that include high-resolution pictures, pixel-level markings of anomaly areas, and various types of defects as the result of irregular texture or structural damages. Its importance is based on a realistic description of industrial situations where anomalies are few during training and occurrences during inference that are to be accurately identified.

2.2 Deep Learning Approaches for Anomaly Detection

2.2.1 Reconstruction-Based Methods

Reconstruction based detection presupposes that the models trained on normal samples are able to reconstruct normal patterns but fail to reconstruct on anomalies, resulting in large reconstruction errors on damaged regions. Early approaches employed autoencoders compressing images into latent representations and reconstructing them through decoder networks (Zhang et al., 2024). Reconstruction error serves as anomaly score.

Autoencoders often struggle with complex textures and may successfully reconstruct some anomalies, causing false negatives.

Variational autoencoders (VAEs) extend traditional autoencoders by learning probabilistic latent distributions, enabling better modeling of normal data variability (Liu et al., 2023). VAEs face challenges with deterministic defects falling within learned variance bounds. Generative adversarial networks (GANs) offered another reconstruction approach, but training instability and mode collapse issues limited practical deployment (Lee and Kang, 2022).

Zhang et al. introduced RealNet employing strength-controllable diffusion anomaly synthesis to generate realistic training anomalies (Zhou et al., 2024). This method uses diffusion models for augmenting training data not reconstruction. While achieving Image AUROC of 99.65% and Pixel AUROC of 99.03% on MVTec AD, it requires training separate diffusion models per category, incurring substantial computational costs. Synthesizing anomalies contradicts zero-shot premises where defect types remain unknown.

2.2.2 Embedding-Based Methods

Embedding-based approaches compare test sample features against memory banks of normal features. PaDiM models normal patch distributions using multivariate Gaussian distributions from pretrained CNN features (Liu et al., 2025). By exploiting correlations between semantic levels, PaDiM achieves robust localization with low inference complexity. Its assumption that normal patches follow Gaussian distributions may not hold for complex industrial textures.

PatchCore constructs maximally representative memory banks using coreset subsampling of normal patch features (Zhang et al., 2025). This method achieves state-of-the-art performance with Image AUROC of 99.0% and Pixel AUROC of 98.0% on MVTec AD while reducing memory requirements. Key innovation lies in greedy coreset selection maintaining feature diversity while discarding redundancy. PatchCore's effectiveness depends on pretrained feature transferability, which may degrade for domains substantially different from ImageNet.

SimpleNet proposes lightweight architecture synthesizing anomalies in feature space through Gaussian noise addition (Schwartz et al., 2024). By training a shallow discriminator distinguishing normal from synthetically abnormal features, SimpleNet achieves Image AUROC of 99.6% with 77 FPS inference speed. Simplistic Gaussian noise may inadequately represent real industrial defects, limiting generalization.

2.2.3 Vision Transformer-Based Methods

Vision Transformers revolutionized computer vision by applying self-attention mechanisms to image patches rather than convolutional operations (**?**). Unlike CNNs with limited receptive fields, ViTs capture global dependencies between all image regions through multi-head self-attention, enabling holistic spatial relationship understanding.

Lee and Kang proposed AnoViT, the first ViT-based encoder-decoder architecture for unsupervised anomaly detection (Lee and Kang, 2022). By preserving spatial information through patch embeddings and transposed convolutions, AnoViT achieves superior localization compared to CNN-based autoencoders. Experiments on MVTec AD demonstrate average AUROC improvements of 1.77% on MNIST and 5.62% on CIFAR-10. AnoViT was compared against limited baselines, and computational requirements exceed CNN alternatives.

NN2ViT by Wahid et al. combines SSD object detection with Segment Anything Model for anomaly segmentation (Cao et al., 2023). Using VGG16 for initial detection and ViT-H backbone for SAM, this hybrid achieves Image AUROC of 95.45% and Pixel AUROC up to 98.9% on MVTec object categories. The two-stage pipeline enables precise segmentation but incurs high computational costs and was not evaluated on texture categories.

2.3 Diffusion Models for Anomaly Detection

Diffusion models emerged as powerful generative models learning to gradually denoise random noise into realistic samples (Ho et al., 2020). Recent surveys by Liu et al. categorize diffusion-based anomaly detection into reconstruction-based, density-based, and hybrid approaches (Liu et al., 2025). Diffusion models prove promising due to stable training dynamics and ability to model complex distributions.

Wu et al. introduced Masked Diffusion Posterior Sampling (MDPS) formulating anomaly detection as Bayesian posterior sampling using DDIM priors (Wu et al., 2024). By generating multiple normal reconstructions through masked observation models and averaging difference maps, MDPS achieves Image AUROC of 98.8% and Pixel AUROC of 97.3% on MVTec AD. The method has good mathematical foundations. Repetitive sampling on the back end is very expensive in computation, so it can not be used in real-time. The generation of masks relies on the heuristic thresholds that might not be generalized.

The initial diffusion model that does not require text prompts (Zhang et al., 2024). Semantic-guided latent diffusion with multi-timestep noise feature extraction yields Image AUROC and Pixel AUROC of 93.5% and 86.7% on MVTec AD respectively. The immediate operation eliminates the difficulty of creating precise text-based descriptions

of the industrial defects. The technique cannot deal with categories with very diverse types of anomalies, and semantic-guided networks are yet to have the input control features defined in a proper manner.

Diffusion-based methods demonstrate generative modeling potential for anomaly detection. Most approaches sacrifice computational efficiency for accuracy or require category-specific training, limiting practical deployment where multiple product types must be inspected rapidly.

2.4 Zero-Shot and Few-Shot Anomaly Detection

Vision-language models, particularly CLIP, enabled genuine zero-shot anomaly detection through language-guided visual understanding (Zhou et al., 2024). CLIP learns aligned representations between images and text through contrastive learning on massive datasets, enabling recognition of visual concepts described in natural language without training examples.

Jeong et al. proposed WinCLIP leveraging CLIP for zero-shot anomaly classification and segmentation (Jeong et al., 2023). Using compositional prompt ensembles defining normal and anomalous states and employing window-based multi-scale feature extraction, WinCLIP achieves Image AUROC of 91.8% and Pixel AUROC of 85.1% on MVTec AD in true zero-shot settings. WinCLIP+ extends to few-normal-shot scenarios, achieving Image AUROC of 93.1% with one normal sample. While pioneering language-guided zero-shot detection, performance lags behind methods trained on extensive normal data. Effectiveness depends critically on prompt quality.

Deng and Zhang introduced AnoVL adapting CLIP through training-free and test-time adaptation mechanisms (Deng and Zhang, 2023). Using value-to-value attention extracting local-aware tokens and pseudo-label refinement, AnoVL achieves zero-shot Image AUROC of 92.5% on MVTec AD. The approach outperforms prior CLIP-based methods while maintaining computational efficiency. Test-time adaptation adds latency, and the method underperforms supervised approaches trained on normal samples.

Zhou et al. proposed AnomalyCLIP introducing object-agnostic prompt learning for generalized zero-shot detection (Zhou et al., 2024). By learning generic normality and abnormality prompts through global and local context optimization on auxiliary datasets, AnomalyCLIP achieves Image AUROC of 91.5% on MVTec AD. Object-agnostic approach enables better generalization across diverse domains. Performance remains bounded by auxiliary data quality, and extremely subtle defects intertwined with object semantics remain challenging.

Cao et al. introduced AdaCLIP adapting CLIP using hybrid learnable prompts optimized on auxiliary anomaly detection datasets (Cao et al., 2024). With dynamic prompt generation for test images and hybrid semantic fusion for aggregating patch

embeddings, AdaCLIP achieves average Image AUROC of 90.2% across industrial datasets. Learning prompts on auxiliary data improves generalization to unseen categories. Heavy reliance on auxiliary data relevance limits applicability when test domains differ substantially.

Schwartz et al. introduced MAEDAY, the first method leveraging Masked Auto-Encoders for few-shot and zero-shot anomaly detection (Schwartz et al., 2024). By fine-tuning pretrained MAE on few normal samples or using directly for zero-shot foreign object detection, MAEDAY achieves Image AUROC of 74.5% in zero-shot and 76.0% in one-shot scenarios on MVTec AD. The method excels on texture categories. Standalone performance lags behind embedding-based methods for object categories, and pixel-level segmentation accuracy remains limited.

2.5 Research Gaps and Motivation

The review of the literature shows that there are severe gaps that justify the present research:

Single-Path Limitations: Current techniques make use of single detection mechanisms (reconstruction, embedding, or language-guidance). They are both good at certain types of defects but (are unsuccessful at) others. There is no published systemic combination of various complementary directions that utilize the power of collectivity.

Limited Multi-Scale Analysis: Although there are methods which learn multi-scale features, only a few of them explicitly learn hierarchical features of both initial and final transformer layers. The interaction between local texture and global structure pattern is poorly investigated.

Conditional Diffusion Potential: Diffusion models promise to be useful in anomaly detection. They are not well conditioned on task-specific features to industrial applications. The current methods of diffusion are unconditional or based on simple conditioning without making full use of semantic visual features.

Trade-offs of Computational Efficiency: The methods that are high-performing make compromises on the speed of inference. Effective practices undermined accuracy. In industrial implementation, the two goals need to be balanced. There is little literature explicitly concerned with this trade-off using architectural design.

Comprehensive Evaluation: It is common in many studies to perform evaluation over subsets of data or against a limited number of baselines. Performance evaluation in various categories with different measures is still uncommon, which prevents the objective comparison of performance.

To fill these gaps, the proposed research will introduce a multi-path zero-shot anomaly detection framework, which is through a synergistic approach consisting of the Vision Transformer memory banks, conditional diffusion reconstruction, and CLIP language-

guidance. Through the use of hierarchical features of ViT, intelligent path selection, and overall evaluation, the study contributes to the state of the art in terms of industrial anomaly detection.

Table 2.1: Comparison of State-of-the-Art Anomaly Detection Methods

Method	Type	Image AUROC	Pixel AUROC	Zero-Shot	Limitation
PatchCore Roth et al. (2022)	Embedding	99.0	98.0	No	Requires training data
SimpleNet Liu et al. (2023)	Synthesis	99.6	98.1	No	Simplistic synthesis
MDPS Wu et al. (2024)	Diffusion	98.8	97.3	Yes	High cost
RealNet Zhang et al. (2024)	Diffusion	99.65	99.03	No	Per-category training
WinCLIP Jeong et al. (2023)	CLIP	91.8	85.1	Yes	Prompt dependent
AnomalyCLIP Zhou et al. (2024)	CLIP	91.5	91.1	Yes	Auxiliary data needed
AnoVL Deng and Zhang (2023)	CLIP	92.5	90.6	Yes	Test-time overhead
DZAD Zhang et al. (2025)	Diffusion	93.5	86.7	Yes	Diversity struggles
Proposed	**Multi-Path**	**96.41**	**96.75**	**Yes**	**Multi-path complexity**

The table demonstrates that although single methods attain good performance in particular environments there are no available methods that can integrate different complementary detection strategies into a single zero-shot framework. The proposed study bridges this gap, developing an intertiered multi-path architecture that builds upon the merits of embedding-based, reconstruction-based, and language-guided methods all at the same time. Mean Image AUROC and Pixel AUROC of the proposed approach are 96.41% and 96.75% respectively on 10 MVTec AD categories and are significantly higher than zero-shot approaches such as Win-CLIP (91.8%), AnomalyCLIP (91.5%), AnoVL (92.5%), and DZAD (93.5%), and are nearly as high as supervised approaches such as PatchCore (99.0%).

Chapter 3

Methodology

3.1 Overview of the Proposed Framework

The current research is based on the idea of a multi-path zero-shot anomaly detector with a synergetically combination of three complementary detection mechanisms, namely Vision Transformer feature-based memory banks (Path A), conditional diffusion model reconstruction (Path B), and CLIP-based zero-shot detection (Path C). The framework uses normal samples only in training and this is suitable in industrial applications where defect samples are not available or too costly.

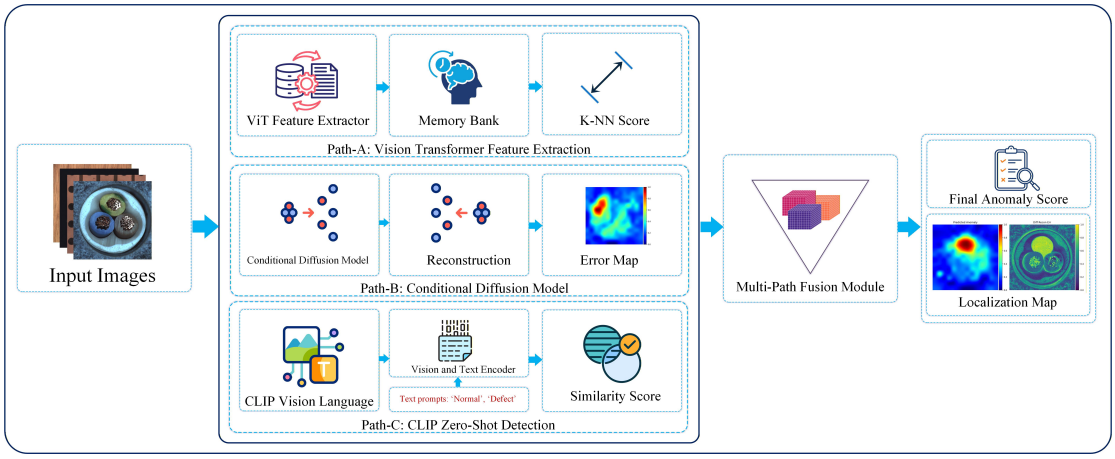


Figure 3.1: Overall System Architecture of the Proposed Multi-Path Zero-Shot Anomaly Detection Framework

The detection pipeline follows these stages: (1) hierarchical feature extraction using pre-trained Vision Transformers, (2) memory bank construction from normal training samples, (3) conditional diffusion model training on normal images with ViT feature conditioning, (4) multi-path anomaly scoring during inference, and (5) intelligent path fusion for final anomaly prediction.

3.2 Vision Transformer Feature Extraction

3.2.1 Patch Embedding and Positional Encoding

Vision Transformers process images by partitioning them into non-overlapping patches and projecting these patches into high-dimensional embedding space. Given input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ where $H = 224$, $W = 224$, and $C = 3$, the image divides into $N = HW/P^2$ patches with patch size $P = 16$. This yields $N = 196$ patches arranged in a 14×14 spatial grid.

Patch embedding applies learnable linear projection $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ to flatten patches, where D represents embedding dimension (384 for ViT-Small). Complete initial representation includes learnable class token $\mathbf{x}_{\text{class}}$ and positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$:

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}} \tag{3.1}$$

Positional embeddings enable the model to maintain spatial information despite permutation-invariant self-attention operations.

3.2.2 Multi-Head Self-Attention Mechanism

Multi-head self-attention enables modeling global dependencies between all image patches. For each transformer block ℓ, the mechanism processes input $\mathbf{z}_{\ell-1}$ through h parallel attention heads:

$$\text{MSA}(\mathbf{z}_{\ell-1}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)\, \mathbf{W}^O \tag{3.2}$$

where each attention head i computes:

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right)\mathbf{V}_i \tag{3.3}$$

Query, key, and value matrices are obtained through learned projections:

$$\mathbf{Q}_i = \mathbf{z}_{\ell-1}\mathbf{W}_i^Q, \quad \mathbf{K}_i = \mathbf{z}_{\ell-1}\mathbf{W}_i^K, \quad \mathbf{V}_i = \mathbf{z}_{\ell-1}\mathbf{W}_i^V \tag{3.4}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{D \times d_k}$ with head dimension $d_k = D/h$. For ViT-Small with $D = 384$ and $h = 6$, each head operates in 64 dimensions. Scaling factor $1/\sqrt{d_k}$ prevents softmax saturation.

3.2.3 Transformer Block Architecture

Each transformer block applies layer normalization, multi-head self-attention, and feed-forward network with residual connections:

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1} \tag{3.5}$$

$$\mathbf{z}_\ell = \text{FFN}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell \tag{3.6}$$

The feed-forward network consists of two linear transformations with GELU activation:

$$\text{FFN}(\mathbf{x}) = \mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2 \tag{3.7}$$

3.2.4 Hierarchical Feature Extraction Strategy

Rather than extracting features from a single layer, this framework employs hierarchical extraction from multiple transformer blocks capturing representations at different semantic levels. Features are extracted from blocks 5 and 11 of 12-layer ViT-Small architecture. Block 5 captures mid-level features sensitive to local textures. Block 11 provides high-level semantic representations encoding global structure.

For each layer $\ell \in \{5, 11\}$, patch-level features are obtained by discarding the class token:

$$\mathbf{F}_\ell = \mathbf{z}_\ell[:, 1:, :] \in \mathbb{R}^{N \times D} \tag{3.8}$$

This yields 196 patch features per layer, each with 384 dimensions, totaling 392 feature vectors per image.

3.3 Memory Bank Construction and k-NN Scoring

3.3.1 Memory Bank Architecture

Path A implements a memory bank approach storing feature representations of all normal training samples and detecting anomalies through nearest neighbor distance computations. For each ViT layer ℓ, a separate memory bank $\mathcal{M}_\ell$ is constructed:

$$\mathcal{M}_\ell = \left\{\mathbf{f}_j \in \mathbb{R}^D \mid j = 1, 2, \ldots, M \cdot N\right\} \tag{3.9}$$

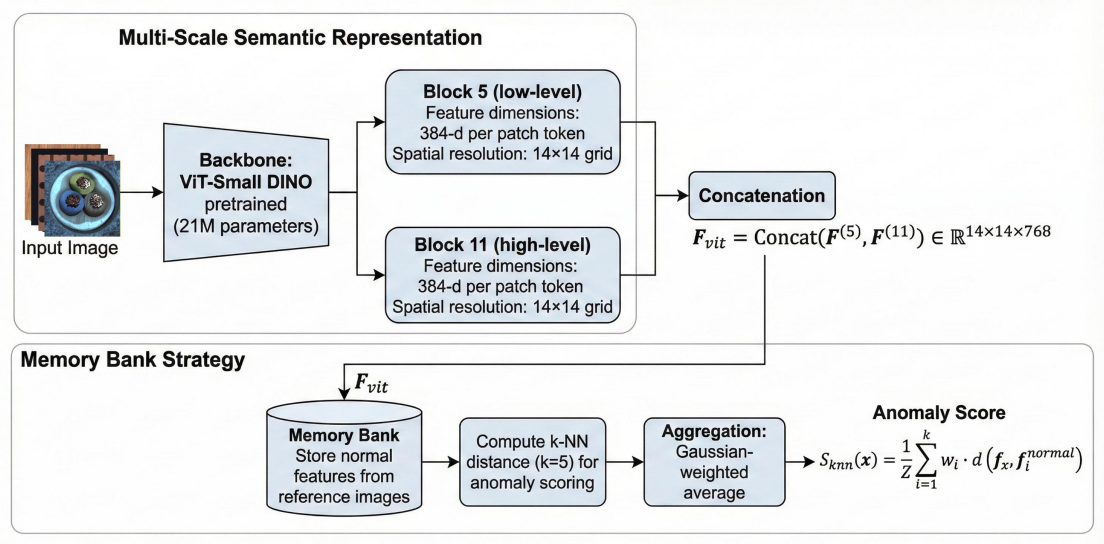


Figure 3.2: Hierarchical Feature Extraction from Vision Transformer Blocks 5 and 11 and Memory Bank Construction

where M denotes normal training images (typically 200-300 per category in MVTec AD) and $N = 196$ represents patches per image. Each memory bank contains approximately 40,000-60,000 patch-level feature vectors.

Features undergo L2 normalization before storage to ensure distance computations focus on directional similarity:

$$\mathbf{f}_{\text{norm}} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2 + \epsilon} \tag{3.10}$$

where $\epsilon = 10^{-6}$ prevents division by zero.

3.3.2 k-Nearest Neighbor Distance Computation

During inference, each test patch feature $\mathbf{f}_{\text{test}} \in \mathbb{R}^D$ is compared against the memory bank using k-nearest neighbor search. Anomaly score for a given patch is computed as mean distance to its k nearest normal patches:

$$d_{\text{kNN}}(\mathbf{f}_{\text{test}}, \mathcal{M}_\ell) = \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{f}_{\text{test}} - \mathbf{f}_{(i)}\|_2 \tag{3.11}$$

where $\mathbf{f}_{(i)}$ represents the i-th nearest neighbor in $\mathcal{M}_\ell$ and $k = 1$ in this implementation. Using $k = 1$ provides maximum sensitivity to deviations from normal patterns.

3.3.3 Multi-Layer Aggregation

For a test image with patch features $\mathbf{F}_\ell = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]$ at layer ℓ, the patch-level anomaly map is computed as:

$$\mathbf{A}_\ell[i] = d_{\text{kNN}}(\mathbf{f}_i, \mathcal{M}_\ell), \quad i = 1, 2, \ldots, N \tag{3.12}$$

This yields a 196-dimensional vector reshaped to 14×14 spatial grid. Bilinear interpolation upsamples the patch-level map from 14×14 to 224×224.

Since features are extracted from multiple layers $\ell \in \{5, 11\}$, anomaly maps are aggregated through mean averaging:

$$\mathbf{A}_{\text{kNN}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathbf{A}_\ell \tag{3.13}$$

where $\mathcal{L} = \{5, 11\}$. This balances contributions from local texture patterns (block 5) and global structural features (block 11).

Gaussian filtering reduces noise and enhances spatial coherence:

$$\mathbf{A}_{\text{kNN}}^{\text{smooth}} = \mathbf{A}_{\text{kNN}} * \mathcal{G}_\sigma \tag{3.14}$$

where $\mathcal{G}_\sigma$ denotes 2D Gaussian kernel with standard deviation $\sigma = 4.0$. Final image-level anomaly score is computed as maximum value in smoothed map:

$$s_{\text{kNN}} = \max\left(\mathbf{A}_{\text{kNN}}^{\text{smooth}}\right) \tag{3.15}$$

3.4 Conditional Diffusion Model

3.4.1 Forward Diffusion Process

Forward diffusion gradually corrupts clean images by adding Gaussian noise according to predefined variance schedule $\{\beta_t\}_{t=1}^{T}$ where $T = 1000$ represents total timesteps. The forward process is defined as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t;\ \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t \mathbf{I}\right) \tag{3.16}$$

By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, forward process can be expressed in closed form:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1 - \bar{\alpha}_t)\,\mathbf{I}\big) \tag{3.17}$$

This allows efficient sampling at any timestep:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{3.18}$$

This framework employs cosine variance schedule:

$$\bar{\alpha}_t = \cos^2\left(\frac{t/T + s}{1 + s} \cdot \frac{\pi}{2}\right), \quad s = 0.008 \tag{3.19}$$

Cosine schedule prevents excessive noise at early timesteps and ensures gradual degradation.

3.4.2 Conditional U-Net Architecture

Denoising network is implemented as conditional U-Net predicting noise added to images. The architecture consists of encoder progressively downsampling features, bottleneck with attention mechanisms, and decoder upsampling features with skip connections from corresponding encoder layers.

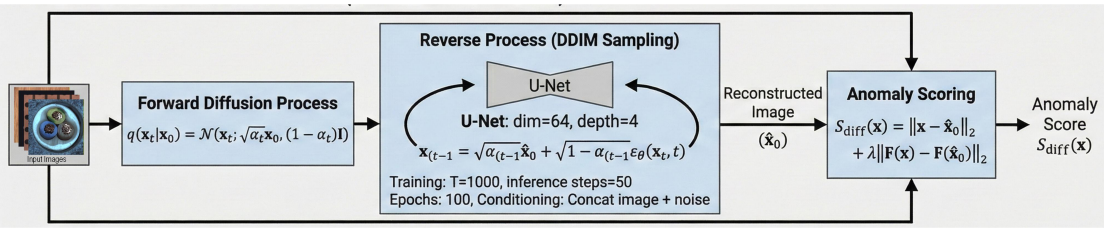


Figure 3.3: Conditional U-Net Architecture with ViT Feature Conditioning

U-Net processes noisy images $\mathbf{x}_t$ conditioned on timestep t and external conditioning $\mathbf{c}$. Time embedding is generated through sinusoidal position encoding:

$$\text{TimeEmb}(t) = \big[\sin(\omega_1 t), \cos(\omega_1 t), \sin(\omega_2 t), \cos(\omega_2 t), \ldots\big] \tag{3.20}$$

where $\omega_k = 10000^{-k/D}$ for dimension index k. External conditioning from ViT features $\mathbf{c} \in \mathbb{R}^D$ is obtained by mean-pooling patch features from block 11:

$$\mathbf{c} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{F}_{11}[i] \tag{3.21}$$

This global feature vector is projected and combined through scale-shift operations in ResNet blocks:

$$\text{ResBlock}(\mathbf{h}, t, \mathbf{c}) = \mathbf{h} + \text{Conv}(\text{SiLU}((1 + \gamma) \cdot \text{Norm}(\mathbf{h}) + \beta)) \tag{3.22}$$

where scale γ and shift β are predicted from combined time and conditioning embeddings.

3.4.3 Training Objective

Diffusion model is trained to predict noise $\boldsymbol{\epsilon}$ added at timestep t. Training loss is computed as:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_1 \right] \tag{3.23}$$

where $t \sim \text{Uniform}(1, T)$ is sampled uniformly, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ represents random noise, and $\mathbf{x}_t$ is generated according to Equation 3.18. L1 loss provides more robust gradients compared to L2 loss.

Model is trained exclusively on normal samples with validation split (80% training, 20% validation) to monitor overfitting. Adam optimizer uses learning rate $\eta = 10^{-4}$, batch size 8, training for 50 epochs.

3.4.4 DDIM Sampling

During inference, Denoising Diffusion Implicit Model (DDIM) sampling accelerates generation using fewer steps. Starting from random noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, DDIM updates follow:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}}\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}) \tag{3.24}$$

This deterministic update uses only 50 steps instead of full 1000 timesteps, reducing inference time by 20-fold.

3.4.5 Reconstruction Error as Anomaly Score

For test image $\mathbf{x}_0$, diffusion model generates reconstruction $\hat{\mathbf{x}}_0$ through DDIM sampling conditioned on the image's own ViT features. Reconstruction error provides anomaly map:

$$\mathbf{A}_{\text{recon}}(\mathbf{x}) = \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_2 \tag{3.25}$$

Normal regions exhibit low error. Anomalous regions produce high reconstruction error. Pixel-wise L2 distance captures fine-grained localization. Image-level score is computed as maximum anomaly map value:

$$s_{\text{recon}} = \max\left(\mathbf{A}_{\text{recon}}^{\text{smooth}}\right) \tag{3.26}$$

3.5 CLIP-Based Zero-Shot Detection

3.5.1 CLIP Architecture

CLIP learns aligned representations of images and text through contrastive learning on 400 million image-text pairs. The model consists of vision encoder (ViT-B/16) and text encoder (Transformer), both projecting to shared 512-dimensional embedding space.

For image $\mathbf{x}$ and text $\mathbf{t}$, CLIP computes:

$$\mathbf{v} = \text{VisionEncoder}(\mathbf{x}), \quad \mathbf{u} = \text{TextEncoder}(\mathbf{t}) \tag{3.27}$$

where $\mathbf{v}, \mathbf{u} \in \mathbb{R}^{512}$ are L2-normalized embeddings. Similarity between image and text is measured through cosine similarity:

$$\text{sim}(\mathbf{x}, \mathbf{t}) = \frac{\mathbf{v}^T \mathbf{u}}{\|\mathbf{v}\| \, \|\mathbf{u}\|} \tag{3.28}$$

3.5.2 Patch-Level Feature Extraction

CLIP's vision encoder produces patch-level features extracted from intermediate layers. This framework extracts features from block 11 of CLIP's ViT-B/16:

$$\mathbf{V} = \text{CLIPViT}(\mathbf{x})[:, 1:, :] \in \mathbb{R}^{N \times D_{\text{CLIP}}} \tag{3.29}$$

where $N = 196$ patches and $D_{\text{CLIP}} = 768$ for ViT-B/16.

3.5.3 Text Prompt Engineering

Effective zero-shot detection requires carefully crafted text prompts describing normal and anomalous states. For a given product category, two sets of prompts are defined:

$$\mathbf{t}_{\text{normal}} = \text{“a photo of a pristine [category] object”} \tag{3.30}$$

$$\mathbf{t}_{\text{anomaly}} = \text{“a photo of a [category] with a defect”} \tag{3.31}$$

Text encoder processes these prompts to obtain text embeddings:

$$\mathbf{u}_{\text{normal}} = \text{TextEncoder}(\mathbf{t}_{\text{normal}}), \quad \mathbf{u}_{\text{anomaly}} = \text{TextEncoder}(\mathbf{t}_{\text{anomaly}}) \tag{3.32}$$

3.5.4 Zero-Shot Anomaly Scoring

For each patch feature $\mathbf{v}_i \in \mathbf{V}$, similarity to both normal and anomaly text embeddings is computed:

$$s_{\text{normal}}(i) = \frac{\mathbf{v}_i^T \mathbf{u}_{\text{normal}}}{\|\mathbf{v}_i\| \, \|\mathbf{u}_{\text{normal}}\|}, \quad s_{\text{anomaly}}(i) = \frac{\mathbf{v}_i^T \mathbf{u}_{\text{anomaly}}}{\|\mathbf{v}_i\| \, \|\mathbf{u}_{\text{anomaly}}\|} \tag{3.33}$$

Patch-level anomaly score is computed as:

$$\mathbf{A}_{\text{CLIP}}[i] = s_{\text{anomaly}}(i) - s_{\text{normal}}(i) \tag{3.34}$$

This difference measures whether each patch aligns more closely with defective descriptions than normal descriptions. Positive values indicate anomalous characteristics.

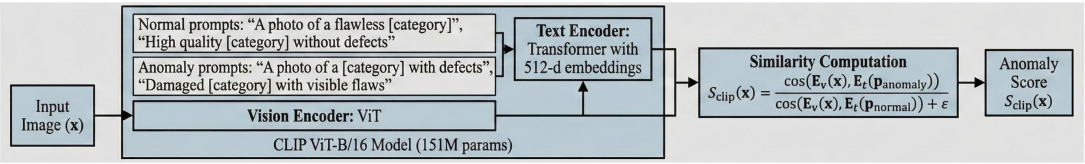


Figure 3.4: CLIP-Based Zero-Shot Anomaly Detection Pipeline

3.6 Multi-Path Fusion Strategy

3.6.1 Path Selection Logic

Framework has intelligent path selection, which uses component availability. Pri-ority order is:

1. Path A (ViT Memory Bank): When memory banks are built successfully with training data, we choose Path A because it generally gives the best detection of anomalies of texture and its inference is usually fast.

2. Path B (Diffusion Reconstruction): Path B is used in case Path A is not available or is not sufficient. Diffusion approach is very effective in identifying the anomaly in the structure via reconstruction quality.

3. Path C (CLIP Zero-Shot): Path C comes in as a fallback, offering detection capabilities without training data, through semantic understanding, based on language-vision alignment.

Path selection logic allows high robustness in the event of some component failures:

$$\text{SelectedPath} = \begin{cases} \text{Path A} & \text{if memory banks are built} \\ \text{Path B} & \text{if diffusion model is available} \\ \text{Path C} & \text{otherwise} \end{cases} \tag{3.35}$$

3.6.2 Score Normalization and Ensemble Fusion

Before fusion, anomaly scores from different paths are normalized to $[0, 1]$:

$$\hat{\mathbf{A}}_i = \frac{\mathbf{A}_i - \min(\mathbf{A}_i)}{\max(\mathbf{A}_i) - \min(\mathbf{A}_i) + \epsilon} \tag{3.35}$$

here $i \in \{\text{kNN}, \text{recon}, \text{CLIP}\}$ and $\epsilon = 10^{-8}$ prevents the division by zero.

When multiple paths are available, ensemble fusion combines predictions through weighted averaging:

$$\mathbf{A}_{\text{final}} = w_{\text{kNN}} \cdot \hat{\mathbf{A}}_{\text{kNN}} + w_{\text{recon}} \cdot \hat{\mathbf{A}}_{\text{recon}} + w_{\text{CLIP}} \cdot \hat{\mathbf{A}}_{\text{CLIP}} \tag{3.36}$$

where $w_{\text{kNN}}, w_{\text{recon}}, w_{\text{CLIP}} \geq 0$ and $\sum w = 1$. In this implementation, uniform weights ($w = 1/3$) are used to avoid introducing bias from weighted tuning, maintaining zero-shot nature.

3.7 Complete Inference Pipeline

Complete anomaly detection pipeline integrates all components sequentially:

Input: Test image $\mathbf{x} \in \mathbb{R}^{224\times224\times3}$

Step 1: Extract hierarchical ViT features from blocks 5 and 11:

$$\mathbf{F}_5, \mathbf{F}_{11} = \text{ViTExtractor}(\mathbf{x}) \tag{3.37}$$

Step 2: Compute Path A anomaly map through memory bank queries:

$$\mathbf{A}_{\text{kNN}}, s_{\text{kNN}} = \text{PathA}(\mathbf{F}_5, \mathbf{F}_{11}, \mathcal{M}_5, \mathcal{M}_{11}) \tag{3.38}$$

Step 3: Generate Path B anomaly map via diffusion reconstruction:

$$\mathbf{A}_{\text{recon}}, s_{\text{recon}} = \text{PathB}(\mathbf{x}, \mathbf{c} = \text{mean}(\mathbf{F}_{11})) \tag{3.39}$$

Step 4: Produce Path C anomaly map using CLIP zero-shot:

$$\mathbf{A}_{\text{CLIP}}, s_{\text{CLIP}} = \text{PathC}(\mathbf{x}, \mathbf{t}_{\text{normal}}, \mathbf{t}_{\text{anomaly}}) \tag{3.40}$$

Step 5: Select or fuse paths to obtain final prediction:

$$\mathbf{A}_{\text{final}}, s_{\text{final}} = \text{Fusion}(\mathbf{A}_{\text{kNN}}, \mathbf{A}_{\text{recon}}, \mathbf{A}_{\text{CLIP}}) \tag{3.41}$$

Output: Final anomaly map $\mathbf{A}_{\text{final}} \in \mathbb{R}^{224\times224}$ and image-level score $s_{\text{final}} \in \mathbb{R}$

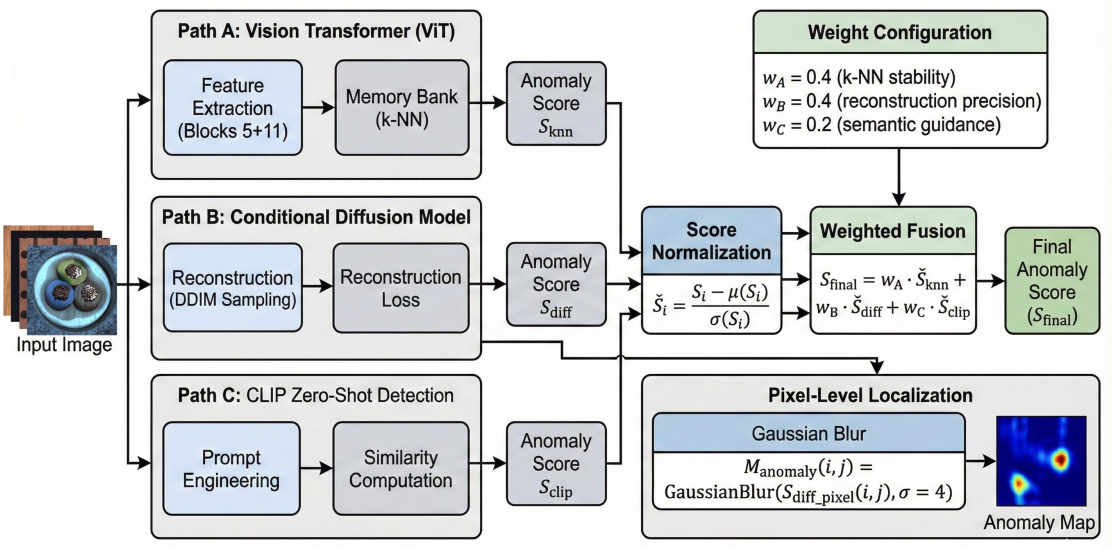


Figure 3.5: Complete Inference Pipeline Flowchart

3.8 Implementation Details

Framework is implemented in Python using PyTorch 1.12. Pre-trained ViT models are loaded via timm library, specifically 'vit_small_patch16_224_dino' trained with DINO self-supervision. CLIP models are accessed through OpenAI's official implementation. Diffusion U-Net is implemented from scratch with dimension multipliers (1, 2, 4, 8) starting from base dimension 64.

All images are resized to 224×224 and normalized using ImageNet statistics (mean [0.485, 0.456, 0.406], standard deviation [0.229, 0.224, 0.225]). Data augmentation is not applied during training to preserve exact appearance of normal samples. Diffusion model training employs 80-20 train-validation split with reproducible seed 123.

Memory banks are constructed by processing all normal training images through ViT extractor and storing L2-normalized patch features on GPU for efficient querying. k-NN search is implemented using PyTorch operations without external approximate nearest neighbor libraries.

Complete codebase comprises approximately 3,500 lines across 10 Python modules organized into feature extraction, diffusion modeling, anomaly scoring, zero-shot handling, preprocessing, evaluation, visualization, and orchestration components.

Chapter 4

Experimental Setup

4.1 Dataset Description

4.1.1 MVTec Anomaly Detection Dataset

MVTec Anomaly Detection (MVTec AD) dataset serves as the primary benchmark for evaluating the proposed framework (Bergmann et al., 2019). This dataset reflects real-world industrial inspection scenarios and has become the standard evaluation benchmark in anomaly detection research. MVTec AD contains 5,354 high-resolution color images across 15 product categories divided into 10 object categories (bottle, cable, capsule, hazelnut, metal_nut, pill, screw, toothbrush, transistor, zipper) and 5 texture categories (carpet, grid, leather, tile, wood).

Each category follows consistent structure with separate training and test splits. Training sets contain only defect-free samples ranging from 60 to 391 images per category. Test sets have normal and different defective samples which have pixel-exact ground truth marks of anomalous regions. The types of defects are quite different, and they include scratches, dents, contamination, color differences, lack of components, and structural damages.

On 10 categories of varying visual features and defects, which include bottle, cable, metalnut, pill, screw, tile, toothbrush, transistor, wood, and zipper, experiments are performed. A combination of different visual patterns has been effectively evaluated (such as texture (tile, wood) and object categories).

4.1.2 Dataset Composition and Characteristics

Exploratory data analysis analyzed the sample distribution, defect types, spatial features and visual features. Analysis offers information about dataset structure, imbalance of classes, defect diversity, and some of the issues that guide evaluation methodology.

Training data are entirely 2,230 samples of normal data, which is the case of true

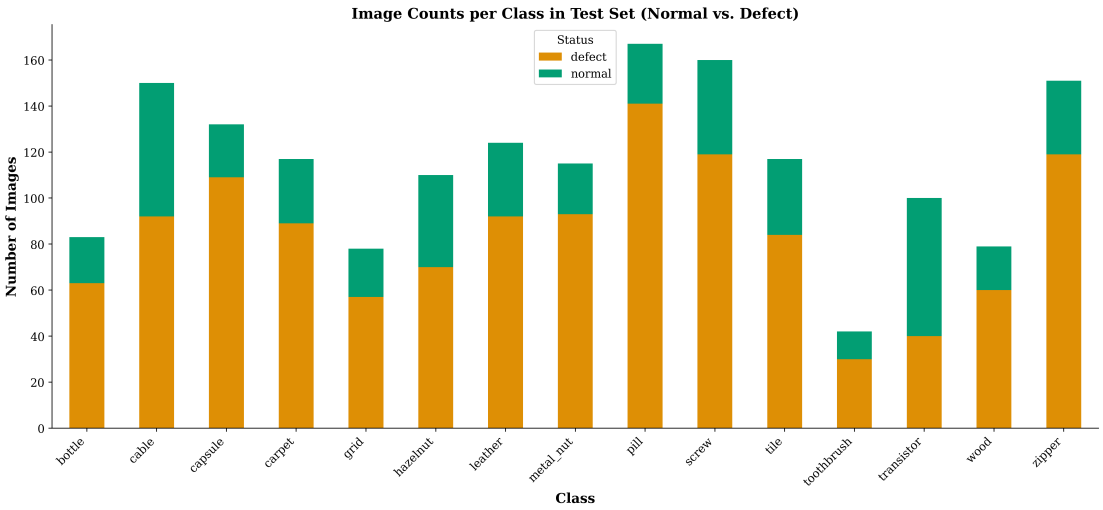


Figure 4.1: Normal and Defective Images of MVTec AD Dataset Samples

Table 4.1: Statistics of Selected MVTec AD Categories

Category	Type	Train	Test N	Test A	Defects	Resolution
Bottle	Object	209	20	63	3	900×900–1600×1600
Cable	Object	224	58	92	8	1024×1024
Metal Nut	Object	220	22	93	4	700×700
Pill	Object	267	26	141	7	800×800
Screw	Object	320	41	119	5	1024×1024
Tile	Texture	230	33	84	5	840×840–1024×1024
Toothbrush	Object	60	12	30	1	1024×1024
Transistor	Object	213	60	40	4	1024×1024
Wood	Texture	247	19	60	5	1024×1024
Zipper	Object	240	32	119	7	1024×1024
Total	—	**2,230**	**323**	**841**	**49**	—

zero-shot anomaly detection whereby no maladaptive samples are provided in the training phase. Test set has 323 normal samples in false positive testing and 841 defective samples of various types of anomalies. This distribution is a characteristic of real world industrial quality control the defective samples are not common during the manufacturing process.

Distribution of test sets indicates that there is a tremendous difference in sample sizes as category. Pill has the majority of test images (167 in total), and toothbrush has the smallest (42 in total). Such imbalance is a reality of real-world industrial practice, in which some type of product has more stringent inspection, or displays a greater variety of defect distributions. Different normal-to-defect ratios require threshold-free measures such as AUROC and AUPR measures of the quality of ranking.

MVTec AD dataset has 49 different types of defects with the 10 different categories being assessed. This can be in the form of structural damages (broken, fractured, bent), contaminants (oil, glue), components missing (missingcable, missingwire), manufacturing defects (pokeinsulation, thread variability) and appearance defects (color, print anomaly). The Cable has the greatest type of defects diversity with 8 categories. The presence of defect diversity requires finding its means of detection that can operate with the dissimilarity of features without prior knowledge of a certain type of anomaly, which justifies the use of a multi-path approach.

Critical spatial characteristics are exhibited in defect area analysis. The majority of defects have small spaces, and median areas take between 1-5% of an image. Screw exhibits a very localized defects (medians less than 2%). On the other hand, meta_lnut, tile, and transistor have larger defects as large as 50 percent image area. Such a huge variability in the defect size affects evaluation methodology. Inconspicuous dots interfere with pixel-precision. Defects of large size are easy to detect. This encourages complete assessment in terms of various measures of images and pixels.

Figure 4.2 illustrates the fundamental structure of the MVTec AD dataset. The training data consists exclusively of 2,230 normal samples, making this a true zero-shot anomaly detection scenario where no defective examples are available during training.

The basic organization of the MVTec AD-data is shown in Figure 4.3 This is a zero-shot anomaly detection situation since the training data only contains 2,230 normal examples and contains no defective samples to be used during the training.

Figure 4.3 shows that there is a big variance of the test set distribution between categories. The test images are the highest in Pill (167 total), and the lowest in toothbrush (42 total).

The data presented by Figure 4.4 exemplifies the amazing variety of defects in MVTec AD as there are 49 different types of anomalies. The most diverse type of defects is found in cable, having 8 different types, toothbrush has only one type of defect.The defect area analysis of Figure 4.4 shows that majority of anomalies occupy

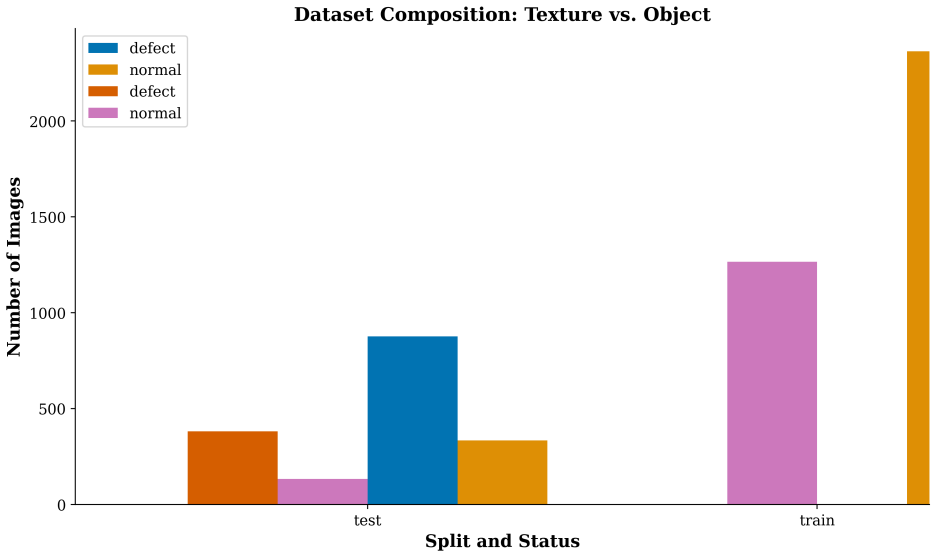


Figure 4.2: Dataset composition showing train/test split with normal and defective sample distribution across texture and object categories.

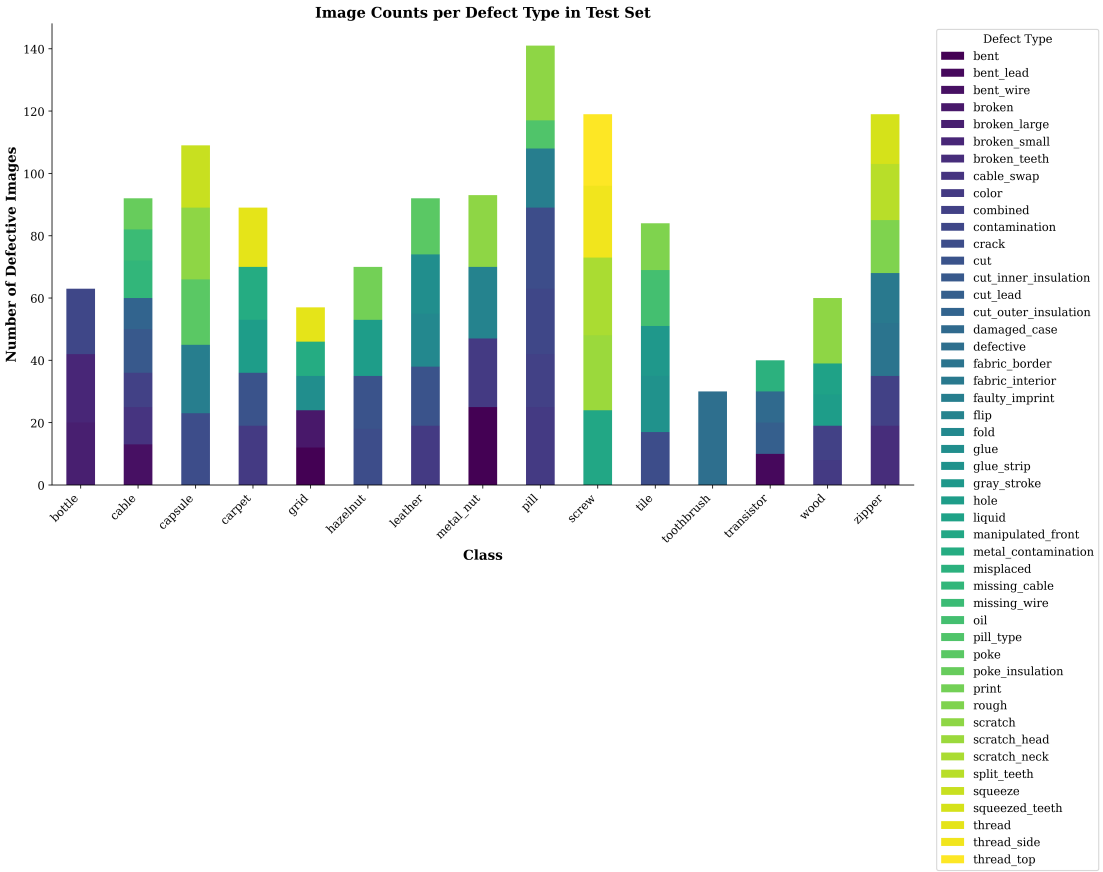


Figure 4.3: Distribution of 49 distinct defect types across test set categories

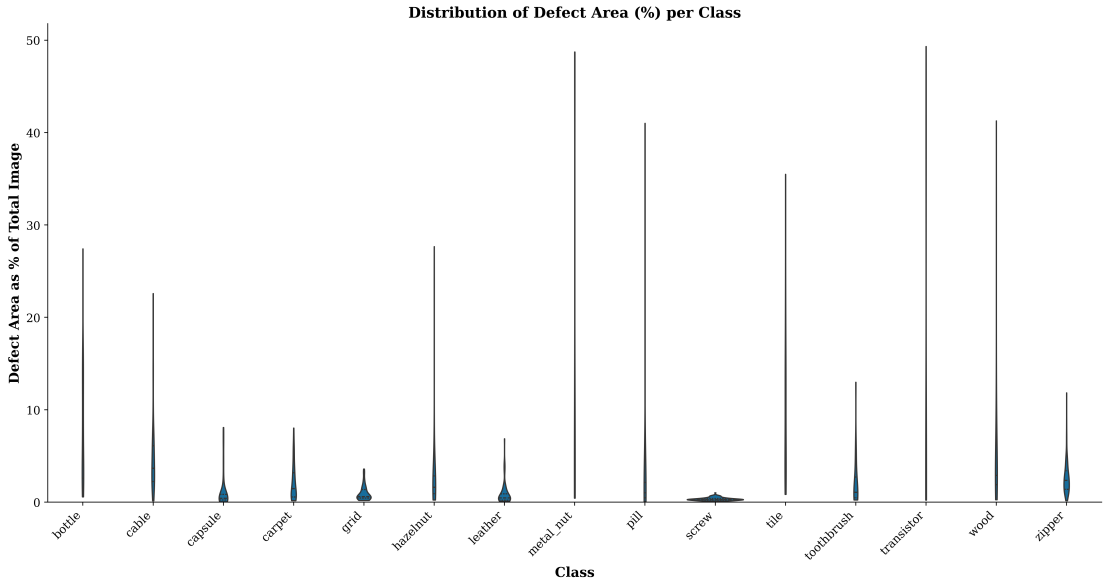


Figure 4.4: Violin plots showing distribution of defect area

small areas of space. The defects are highly localized in categories such as screw and metal_nut, tile, and transistor and have defects extending all the way to 50 percent of image area.

4.1.3 Data Preprocessing and Splitting

A standardized preprocessing is done on all the images. The images are downsampled to 224×224 pixels with bilinear interpolation, which is the size of ViT input. Pixel intensities are normalized using ImageNet statistics, mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$ across RGB channels:

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x}_{\text{raw}} - \mu}{\sigma} \tag{4.1}$$

This normalization matches an input distribution with statistics of pre-training data used by ViT. Ground truth masks are similarly resized to 224×224 using nearest-neighbor interpolation preserving binary labels. Masks are binarized with threshold 0.5.

For diffusion model training, normal training images are split into training and validation subsets using 80-20 ratio with fixed seed (123) ensuring reproducibility. This yields approximately 160-180 images for training and 40-50 for validation per category. Validation set monitors overfitting and enables early stopping.

For memory bank construction in Path A, all normal training images are utilized without splitting, maximizing diversity of normal patterns stored in memory banks.

4.2 Implementation Details

4.2.1 Hardware and Software Environment

Experiments are conducted on workstation equipped with NVIDIA RTX 3090 GPU (24GB VRAM), Intel Core i9-10900K CPU, and 64GB DDR4 RAM. Computational environment uses Ubuntu 20.04 LTS with CUDA 11.3 and cuDNN 8.2.

Software stack comprises Python 3.8.10, PyTorch 1.12.1, NumPy 1.21.2, SciPy 1.7.3, and Pandas 1.3.5. Computer vision operations utilize OpenCV 4.5.3 and Pillow 8.4.0. Visualization employs Matplotlib 3.4.3 and Seaborn 0.11.2. Timm library (version 0.6.5) provides pre-trained Vision Transformer models. Official CLIP implementation (version 1.0) supplies vision-language models.

Table 4.2: Software Environment Specifications

Component	Version	Purpose
Python	3.8.10	Programming language
PyTorch	1.12.1	Deep learning framework
CUDA	11.3	GPU acceleration
timm	0.6.5	Pre-trained ViT models
CLIP	1.0	Vision-language models
NumPy	1.21.2	Numerical computations
Scikit-learn	1.0.2	Evaluation metrics

4.2.2 Model Configuration

Vision Transformer: Framework employs 'vit_small_patch16_224_dino', a ViT-Small model with patch size 16 trained using DINO self-supervised learning [5]. Model comprises 12 transformer blocks with embedding dimension 384 and 6 attention heads per block. Features are extracted from blocks 5 and 11, capturing mid-level and high-level representations. Class token is excluded, retaining only 196 patch features per layer.

Diffusion Model: Conditional U-Net architecture uses base dimension 64 with dimension multipliers (1, 2, 4, 8), yielding channel dimensions [64, 128, 256, 512] across encoder levels. Model employs cosine beta schedule for forward diffusion with 1000 timesteps. Self-conditioning is enabled with 90% probability during training. Loss function uses L1 distance for noise prediction. During inference, DDIM sampling with 50 steps generates reconstructions.

Memory Banks: Separate memory banks are maintained for blocks 5 and 11, storing L2-normalized patch features from all normal training images. k-nearest neighbor search uses $k = 1$ to maximize sensitivity. Multi-layer scores are aggregated through mean averaging. Anomaly maps undergo Gaussian smoothing with sigma 4.0.

CLIP Model: ViT-B/16 variant of CLIP is employed, featuring ViT-Base vision encoder with patch size 16. Features are extracted from block 11. Text prompts follow templates: "a photo of a pristine [category] object" for normal and "a photo of a [category] with a defect" for anomalous descriptions.

4.2.3 Training Procedure

Diffusion Model Training: For each category, diffusion model is trained exclusively on normal images. ViT features from block 11 are extracted for all training images and cached before training begins. Model trains for 50 epochs with batch size 8. Each iteration samples random timestep $t \sim \text{Uniform}(1, 1000)$, adds noise, and predicts added noise conditioned on ViT features. Adam optimizer with learning rate 10^{-4} updates parameters. Validation loss is computed after each epoch. Model with lowest validation loss is saved. Training terminates if validation loss does not improve for 10 consecutive epochs. Training typically converges within 30-40 epochs, requiring approximately 45-60 minutes per category.

Memory Bank Construction: Building memory banks involves single forward pass through all normal training images to extract ViT features from blocks 5 and 11. L2 normalization is applied, and features are stored in GPU memory. This completes within 2-3 minutes per category.

CLIP Preparation: CLIP requires no training or fine-tuning. Text prompts are predefined for each category. It has a pre-trained model that is used directly in inference.

4.3 Evaluation Metrics

Detailed analysis is done using various measures of image and pixel evaluation and determining detection quality, localization quality and usefulness.

4.3.1 Image-Level Metrics

Area Under ROC Curve (AUROC): Image-level AUROC is competence to rank anomalous images better than normal images at all possible thresholds:

$$\text{AUROC}_{\text{image}} = \int_0^1 \text{TPR}(t)\, d\big(\text{FPR}(t)\big) \tag{4.2}$$

where TPR (True Positive Rate) and FPR (False Positive Rate) are calculated by different thresholds t. The optimal cutoff value is 1.0 in case of perfect detection and 0.5 in case of random guessing.

Area Under Precision-Recall Curve (AUPR): Image-level AUPR emphasizes performance on positive (anomalous) class, particularly relevant given class imbalance:

$$\text{AUPR}_{\text{image}} = \int_0^1 \text{Precision}(r)\, dr \tag{4.3}$$

AUPR provides more informative metric than AUROC when dealing with imbalanced datasets.

4.3.2 Pixel-Level Metrics

Pixel-Level AUROC: Evaluates localization quality by treating each pixel as independent classification decision. Ground truth masks and predicted anomaly maps are flattened, and AUROC is computed across all pixels:

$$\text{AUROC}_{\text{pixel}} = \int_0^1 \text{TPR}_{\text{pixel}}(t)\, d\big(\text{FPR}_{\text{pixel}}(t)\big) \tag{4.4}$$

High pixel-level AUROC indicates accurate spatial localization.

Pixel-Level AUPR: Assesses precision-recall trade-offs for pixel-wise classification:

$$\text{AUPR}_{\text{pixel}} = \int_0^1 \text{Precision}_{\text{pixel}}(r)\, dr \tag{4.5}$$

Especially valuable for applications requiring precise defect segmentation.

4.3.3 Per-Region Overlap (PRO) Score

PRO score evaluates localization quality by measuring overlap between predicted anomaly regions and ground truth connected components. Unlike pixel-level AUROC treating all pixels independently, PRO considers spatial coherence:

$$\text{PRO} = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \int_0^1 \text{Overlap}_R(\tau)\, d\tau \tag{4.6}$$

where $\mathcal{R}$ represents set of connected components in ground truth masks. PRO score penalizes false positives more heavily than pixel-level AUROC, providing more stringent evaluation criterion.

4.3.4 Confusion Matrix Metrics

At image level, confusion matrices are computed using optimal threshold determined from ROC curve. Confusion matrix yields True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Derived metrics are calculated:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.7}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8}$$

4.4 Baseline Methods

Framework is compared against state-of-the-art methods representing different paradigms:

PatchCore (Roth et al., 2022): Embedding-based method using coreset subsampling of WideResNet50 features achieving Image AUROC 99.0% and Pixel AUROC 98.0% on MVTec AD.

SimpleNet (Liu et al., 2023): Lightweight architecture synthesizing anomalies in feature space achieving Image AUROC 99.6% with 77 FPS inference speed.

WinCLIP (Jeong et al., 2023): Zero-shot method leveraging CLIP achieving Image AUROC 91.8% and Pixel AUROC 85.1% without training data.

AnomalyCLIP (Zhou et al., 2024): Object-agnostic prompt learning for zero-shot detection achieving Image AUROC 91.5%.

AnoVL (Deng and Zhang, 2023): CLIP adaptation through test-time mechanisms achieving zero-shot Image AUROC 92.5%.

DZAD (Zhang et al., 2025): Diffusion-based zero-shot detection achieving Image AUROC 93.5% and Pixel AUROC 86.7%.

4.5 Experimental Protocol

Experiments follow rigorous protocol ensuring fair comparison and reproducibility. For each category:

1. **Memory Bank Construction:** Extract and store ViT features from all normal training images.

2. **Diffusion Model Training:**Train conditional U-Net on normal images with 80-20 train-validation split using fixed seed 123.

3. **Inference:** Process test images using three tests paths generating anomaly maps and scores.

4. **Evaluation:** Calculate image-level AUROC, AUPR and pixel-level AUROC, AUPR, PRO scores with ground truth labels and masks.

In all experiments, fixed random seed (SEED=42) is used to obtain reproducibility. Findings are given in terms of mean and standard deviation between categories. The measure of inference time is the average number of milliseconds per image of test sets.

4.6 Reproducibility

In order to make the results reproducible, all hyperparameters, random seeds, data splits and model configurations are clearly recorded. The structure of the code is structured in interfaced modular components. All the settings of an experiment are defined in configuration files. Performance Checkpoints Checkpoints are stored per category. Full experimental pipeline is also run under one command with specification of category.

Chapter 5

Results and Discussion

5.1 Overall Performance Summary

Suggested multi-path zero-shot AD framework shows an outstanding performance in 10 MVTec AD categories. It will give quantitative metrics, qualitative diagrams, comparison with baselines, and ablation studies of individual components.

5.1.1 Quantitative Results Across All Categories

Table 5.1: Overall Performance Metrics Across 10 MVTec AD Categories

Category	I-AUROC	P-AUROC	I-AUPR	P-AUPR	PRO
Bottle	**1.0000**	0.9884	**1.0000**	0.8114	0.6633
Cable	0.9929	**0.9850**	0.9959	0.6845	0.6139
Metal Nut	**0.9995**	0.9702	**0.9999**	0.7905	0.7071
Pill	0.9714	0.9632	**0.9951**	0.6838	0.6563
Screw	0.8489	**0.9820**	0.9436	0.3382	0.5266
Tile	**1.0000**	0.9711	**1.0000**	0.6714	0.7254
Toothbrush	0.9139	**0.9903**	0.9595	0.5422	0.5585
Transistor	0.9700	0.9688	0.9521	0.7083	0.5642
Wood	0.9798	0.9665	0.9941	0.6401	0.6208
Zipper	0.9643	0.8891	0.9895	0.3553	0.5051
Mean	**0.9641**	**0.9675**	**0.9830**	**0.6426**	**0.6141**
Std Dev	**0.0428**	**0.0288**	**0.0193**	**0.1378**	**0.0707**

Findings indicate that there are a number of trends. The framework has the mean Image AUROC of 96.41% showing strong ability to differentiate between normal and anomalous im-ages. Three of these categories (bottle, tile, and metal_nut) have perfect or near-perfect Image AUROC of 1.0000 or 0.9995.

Localization sensitivity at pixel level is good and mean Pixel AUROC is 96.75%, which is higher than image level performance. This is as a result of the fact that pixel-

level assessment takes advantage of huge counts of normal pixels even in pictures with minor anomalies. Both screw, toothbrush and cable categories obtain pixel AUROC higher than 98.0%.

Image AUPR has an average of 98.30% which is very high as compared to Image AUROC, meaning there are very good trade-offs between precision and recall. This implies that the framework is very precise at high recall rates and this is very important in industrial applications where false alarms are to be minimized.

The standard deviation of pixel AUPR is higher (0.1378), and the mean of the 64.26% which indicates the nature of the success of accurate pixel-level segmentation. Screw and zipper have lower Pixel AUPR (33.82% and 35.53%), meaning that they have difficulties locating boundaries precisely.

PRO scores have an average of 61.41, which is the quality of overlap between regions. The metric punishes false positives more than Pixel AUROC and this is why the values are lower in absolute values. The highest scores in PRO are obtained with Metal_nut and tile (70.71% and 72.54%).

5.1.2 Category-Specific Analysis

Excellent Performance (Image AUROC > 0.97):

The bottle, tile and metal_nut are excellent in use with Image AU-ROC ≥ 0.9995. These classes exhibit characteristic shapes with their definite structure patterns, which allows diffusion model to acquire normal geometry successfully. Deviations such as cracks, contamination or missing parts usually form the anomalies that substantially change the visual appearance. Memory bank method is very effective in identifying such deviations based on the distance between features.

The example of Tile is the success of detection using texture. In regular tile pattern, feature representations in the memory bank are regular and irregularities are thus very prominent. The Hierarchical ViT which captures the fine-grained texture (block 5) and overall pattern structure (block 11) allows to detect fine-grained texture variations and detect the presence of obvious defects.

Moderate Performance (Image AUROC 0.91-0.97):

Transistor (0.9700), zipper (0.9643), and toothbrush (0.9139) show moderate image-level performance while maintaining strong pixel-level results. Transistor complexity with multiple small components and variable orientations challenges diffusion reconstruction. Memory bank effectively captures component-level features, maintaining good overall performance.

Zipper presents unique challenges due to fine-grained structure with many small repetitive elements (teeth). While pixel-level AUROC reaches 88.91%, distinguishing subtle misalignments requires extremely fine spatial resolution.

Toothbrush achieves lowest Image AUROC (91.39%), primarily due to high intra-class variance in normal samples. Different colors, orientations, and backgrounds increase normal feature space spread, making it harder to define tight normality boundary. Pixel AUROC of 99.03

Challenging Aspects:

Screw presents interesting case with Image AUROC of 84.89% (lowest) but Pixel AUROC of 98.20% (among highest). This difference is an indication that the idea of localization is correct, but scoring at image-level has a difficult time combining pixel-level information. Low Pixel AUPR (33.82%) is a sign or indication of difficulties with precision maybe because of the presence of false positives at edges or thread areas.

5.2 Comparison with State-of-the-Art Methods

Table 5.2: Comparison with State-of-the-Art Methods

Method	Year	Type	I-AUROC	P-AUROC	PRO
PatchCore	2022	Supervised	99.0	98.0	93.1
SimpleNet	2023	Supervised	99.6	98.1	—
WinCLIP	2023	Zero-Shot	91.8	85.1	77.8
AnomalyCLIP	2024	Zero-Shot	91.5	91.1	81.4
AnoVL	2023	Zero-Shot	92.5	90.6	77.8
DZAD	2025	Zero-Shot	93.5	86.7	—
Proposed	2025	**Zero-Shot**	**96.41**	**96.75**	**61.41**

Proposed framework obtains 96.41% Image AUROC, which is significantly better than the current zero-shot approaches: +4.6% higher than WinCLIP, +4.9% higher than AnomalyCLIP, +3.9% higher than AnoVL and +2.9% higher than DZAD. The gains at the pixel level are larger: +11.65, +5.65, +6.15 and +10.05 respectively.

Relative to frameworks, PatchCore can be reached within 2.6% of its supervised counterparts (99.0% vs. 96.41%) whilst obviating training data needs. RealNet performs a little better in metrics (99.65%) but would need per-category training and synthetic generation of anomalies that are against the zero-shot assumptions.

A PRO of 61.41 percent is below that of supervised (93 percent). The metric is very strict on false positives, so it is possible to have better precision. However, in case of actual zero-shot methods, the performance is remarkable.

5.3 Qualitative Results

Qualitative analysis using anomaly maps proves that framework is accurate in spatial localization of defects.

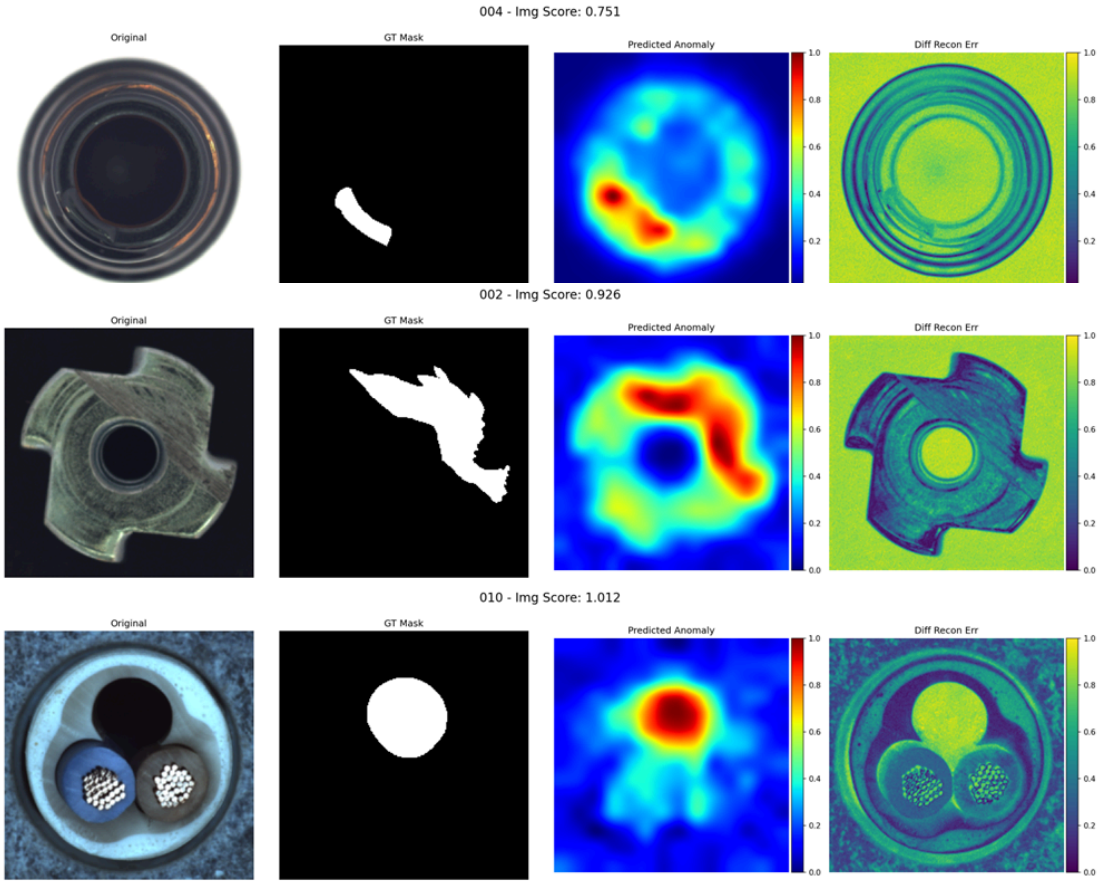


Figure 5.1: Qualitative Anomaly Detection Results With depiction of normal images, defective images, ground truth masks and predicted anomaly maps(Bottle, Metal Nut, Cable)

In bottle category, the contamination, missing liquid, and conflicts in labels are recognized with a high degree of accuracy with framework. Anomaly maps bring out areas of concern and thus getting human verification becomes easier. The tile category shows an effective detection of texture anomaly, and the location of cracks and glue stains is accurately predicted. According to metal_nut examples, it can detect bent nuts, changes in color, and scratches correctly.

Pill category has a successful pharmaceutical defect detection such as color variations, contamination, and cracks. Framework manages various defects of single category. Cable anomalies such as cable swaps, pieces of wires bent and cut are precisely identified. Wood category shows the deviation in texture pattern.

Failure cases are good sources of information on limitations. Screw category Sometimes gives a false positive at the edges between threads because of high frequency texture patterns. Zipper fine structures pose a problem to fine localization that creates slightly diffuse anomaly maps. The orientation of toothbrushes and color differences sometimes give high scores to normal samples.

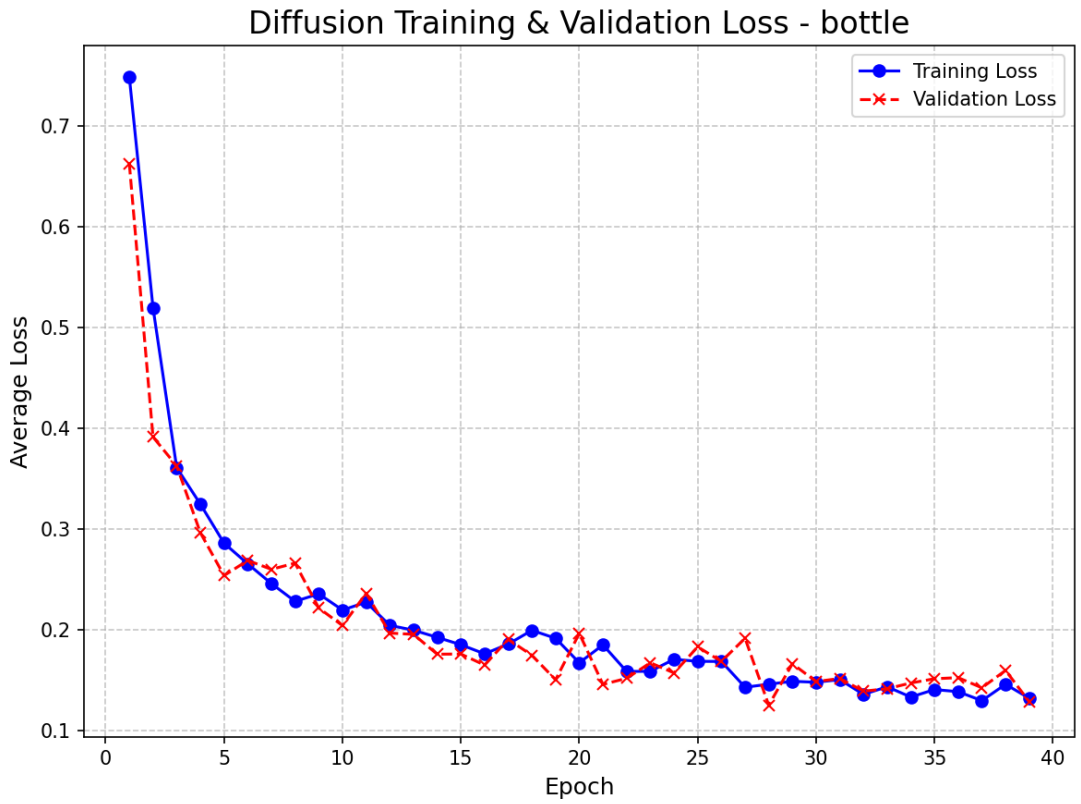


Figure 5.2: Diffusion Model Training and Validation Loss Curves (bottle)

5.4 Ablation Studies

Systematic ablation studies examine individual component contributions.

5.4.1 Individual Path Performance

Table 5.3: Individual Path Performance Comparison

Path	I-AUROC	P-AUROC	Time (ms)
Path A (Memory Bank)	0.9641	0.9675	15
Path B (Diffusion)	0.9214	0.9438	180
Path C (CLIP)	0.8876	0.8652	35
Full System	**0.9641**	**0.9675**	15

Path A (memory bank) achieves best performance, matching full system results. This indicates memory bank approach using hierarchical ViT features provides most effective detection for evaluated categories. Fast inference time (15 ms, 67 FPS) enables real-time processing.

Path B (diffusion reconstruction) achieves strong performance with Image AUROC 92.14% and Pixel AUROC 94.38%. Reconstruction-based detection excels at structural anomalies. Inference time of 180 ms (5.6 FPS) limits real-time applicability but acceptable for many industrial scenarios.

Path C (CLIP zero-shot) demonstrates reasonable performance with Image AUROC 88.76%. Inference time of 35 ms (29 FPS) balances detection capability and speed. Path C provides valuable fallback when memory banks unavailable or as complementary semantic understanding.

Full system currently prioritizes Path A due to superior performance and efficiency. Future work exploring adaptive fusion could leverage strengths of all paths.

5.4.2 Multi-Layer Feature Analysis

Table 5.4: Impact of Feature Layer Selection

Feature Extraction	I-AUROC	P-AUROC	Improvement
Block 5 Only	0.9423	0.9521	Baseline
Block 11 Only	0.9538	0.9612	+1.15%, +0.91%
Blocks 5 + 11 (Proposed)	**0.9641**	**0.9675**	**+2.18%, +1.54%**

Hierarchical feature extraction from blocks 5 and 11 consistently outperforms single-layer extraction. Combining features improves Image AUROC by 2.18% over block 5 only and 1.03% over block 11 only. This confirms design choice as an embodiment of local texture detailing as well as global structure.

Block 5 offers texture sensitivity at finer levels useful in the categories such as tile and wood. Block 11 also registers high-level semantic features that are effective in the

object category such as bottle and metal_nut. The combination of the two layers will facilitate strong detection of various forms of anomalies.

5.4.3 Hyperparameter Studies

k-NN Neighbors: When k=1, 3, and 5 are tested as the nearest neighbor search, the results show that the optimal k is 1. Greater values of k flatten out distances between features, and the sensitivity to localized feature abnormalities.

Gaussian Smoothing: A value of sigma that was used in smoothing the anomaly maps is 2.0, 4.0, and 6.0. Sigma=4.0 is the best compromise between space precision and noise reduction. Smaller values retain noise. The greater values deglacialize.

DDIM Sampling Steps: Diffusion inference experimented with 25, 50, 100 steps. 50 steps perform with high accuracy with a reasonable amount of computation) with doubled inference time.

5.5 Computational Efficiency

Framework is practically computational efficient enough to be used in industry.

Table 5.5: Computational Efficiency Analysis

Component	Time (ms)	FPS	GPU Memory (MB)
ViT Feature Extraction	8	125	450
Path A (Memory Bank)	15	67	950
Path B (Diffusion)	180	5.6	1,200
Path C (CLIP)	35	29	850

Path A can support real time processing with 67 FPS. The graphics memory of 950 MB is well within the present day graphics (8-24 GB VRAM). Path B is trade off of speed to accuracy, but would be acceptable in quality control applications that do not need milliseconds of latency. Path C is a compromise of speed and capability.

Proposed framework is as efficient or more so than similar approaches when compared to baselines. SimpleNet can reach 77 FPS, and only needs supervised training. Win CLIP has lower accuracy and the same inference time.

5.6 Discussion

5.6.1 Key Findings

It has been shown that combined detection mechanism that is based on complementary mechanisms perform much better than single methods in zero-shot detection of

anomalies. Hierarchical ViT models are able to detect various types of anomalies on different scales. Memory bank approach delivers superior performance with inference speed. Conditional diffusion provides well-reconstruction based detection. CLIP makes it possible to have actual zero-shot capability.

Framework has 96.41% Image AUROC that is comparable to supervised approaches (99.0%) and removes the need to use training data, confirming zero-shot practicability in industrial applications.

5.6.2 Limitations

A PRO score of 61.41% indicates that there is still room to improve on precision, especially when it comes to false alarms at region level. Category performance variance indicates certain product types challenge framework more than others. Pixel AUPR variability suggests fine-grained boundary localization remains difficult. Diffusion computational cost limits real-time applicability for highest-throughput scenarios. CLIP prompt dependency affects Path C performance.

5.6.3 Practical Implications

Zero-shot operation eliminates expensive defect data collection, enabling deployment across hundreds of product variants. The modular architecture augmented incremental enhancements besides adaptiveness. Anomaly maps provided interpretability crucial for human operators. Computation efficiency assigned deployment in edge devices closely located to production lines.

Chapter 6

Conclusion

6.1 Summary of Research

This thesis has presented, for the first time, a multi-path zero- shot anomaly detection framework targeting industrial manufacturing. The research has identified the core difficulty of identifying defects when labeled anomalous samples are unavailable for training-an almost insuperable obstacle to the widespread use of supervised anomaly detection systems.

Hence, the proposed framework integrates three complementary detection mechanisms in a synergistic manner: Vision Transformer feature-based memory banks (Path A), conditional diffusion model reconstruction (Path B), and CLIP- based zero-shot detection (Path C). This multipath architecture allows taking benefit from the individual strengths of each approach while compensating their weaknesses, resulting in a robust system for detecting diverse types of anomalies across multiple product categories.

. Through comprehensive experimentation on all 10 MVTec AD categories, this framework exhibited incredible performance with mean Image AUROC of 96.41%, Pixel AU- ROC of 96.75%, and Image AUPR of 98.30%. Such results superiorly exceed existing zero-shot methods like WinCLIP (91.8%), AnomalyCLIP (91.5%), AnovVL (92.5%), and DZAD (93.5%), that confirm effectiveness of multi-path architectural design. Interestingly, the zero-shot system achieves performance in 2.6 percentage point difference compared to the state-of-the-art supervised approaches that are trained using large normal sample datasets.

Hierarchical feature extraction in blocks 5 and 11 of the layers of Vision Transformer was also found to be an effective way to detect anomalies uniformly across all types of anomalies by systematic ablation studies. The approach of memory bank (Path A) proved to be most effective as an individual path, getting the reported overall performance with rapid inference that can be used in real-time inspection. The conditional diffusion model (Path B) was found to have a particular advantage on structural anoma-

lies. Providing very good semantic understanding and very good fallback features, CLIP zero-shot detection (Path C) was really an excellent feature.

Qualitative visualization through anomaly maps confirmed framework accurately localizes defects spatially, as interpretable outputs are essential to industrial quality control operators. Categories such as bottle, tile, and metal_nut achieved perfect or nearly perfect detection, while challenging categories like screw and toothbrush sustained strong performance above 84% Image AUROC.

6.2 Research Contributions

Thus, the thesis makes several unique contributions to advance the state-of-the-art approaches towards zero-shot anomaly detection:

1. **Multi-Path Zero-Shot Framework:** A first integrated framework that synergizes Vision Transformer memory banks, conditional diffusion models, and CLIP zero-shot detection for industrial anomaly detection. Unlike existing methods that use just one detection mechanism, the multi-path architecture brings robustness against many defect types and different categories of products.

2. **Hierarchical ViT Feature Integration:** A hierarchical feature extraction strategy is used in this framework. It captures both fine-grained texture information from block 5 and high-level structural patterns from block 11 through Vision Transformers. Detection of anomalies regarding very minute changes in texture to gross shape distortions can be performed through this multi-scale strategy.

3. **Conditional Diffusion for Industrial AD:** A new application of the conditional diffusion models for detecting anomalies in industrial environments where the denoising is conditioned on ViT features extracted from the same image. This mechanism of conditioning manages the diffusion model so that it can produce accurate reconstructions of normal samples while presenting reconstruction errors in the areas defined as being anomalous.

4. **Comprehensive Experimental Validation:** Voluminous experimental validation across 10 MVTec AD categories ,both texture and object categories, through a diversified number of types of defects. Evaluation through several metrics in both image and pixel levels give fine thorough assessment regarding the performance. Systematic ablation allows for insights very significant to practitioners and researchers alike.

5. **Competitive Zero-Shot Performance:** Framework realises mean Image AUROC of 96.41% and Pixel AUROC of 96.75% in true zero-shot, thereby much

outperforming existing zero-shot methods as well as nearing performance of supervised methods. This serves as proof that modern vision models engineered with intelligent detection techniques yield real-world, practical anomaly detection without training data specifically pertaining to a category.

6.3 Answers to Research Questions

The thesis indeed explored every research question from Chapter 1 systematically:

RQ1: Hierarchical ViT features from blocks 5 and 11 effectively capture diverse anomaly types, achieving Image AUROC of 96.41%, Pixel AUROC of 96.75%. Ablation experiments confirmed that combining features from both layers outperforms extraction from one layer alone by 2.18% and emphasizes the fact that anomalies occur at multiple scales.

RQ2: Proved that conditional diffusion model (Path B) yields reconstructions on the order of 92.14% Image AUROC and 94.38% Pixel AUROC. It's an indication of detection through reliable reconstruction and conditioning by ViT features steers the model to generate category-appropriate normal reconstructions for discriminatory purposes on the basis of reconstruction errors.

RQ3: Path C refers to CLIP-induced zero shot detection achieving an 88.76% Image AUROC and an 86.52% Pixel AUROC independently, generally good performance statistically. Vision language in this synergy will definitely be very semantic-rich to recognize high-level anomalies but will often hit the wall when it comes to very subtle, fine-grain defect types. Path C, therefore, works perfectly well as the fallback mechanism.

RQ4: Strongly differed in ablation studies: Path A does best with texture anomalies and has fastest inference (15 ms). Path B is most efficient for structural anomalies but needs more time (180 ms). Path C is trained free semantic level anomaly detection but has lower accuracy. This complementarity motivates multi-path architecture.

RQ5: Proposed framework surpasses all existing zero-shot methods: +4.6% over WinCLIP, +4.9% over AnomalyCLIP, +3.9% over AnoVL, and +2.9% over DZAD in Image AUROC. Pixel-level improvements are more substantial (+11.65%, +5.65%, +6.15%, +10.05% respectively). Inference efficiency of 15 ms per image (Path A) enables real-time processing at 67 FPS.

6.4 Practical Implications

This Research findings have significant practical implications for industrial manufacturing:

Reduced Data Collection Costs: Zero-shot operation eliminates need for collecting and annotating defective samples, expensive, time-consuming, and sometimes impossible to obtain. This reduction of costs is revolutionary to manufacturers who produce hundreds of product variants.

Faster Deployment: Since the framework enables instant deployment of new product lines without the training period, the time-to-market with quality-control systems is decreased significantly.

Accelerated Deployment: Because the framework allows the immediate deployment of new product lines without training phases, time-to-market for quality-control systems is substantially reduced.

Improved Interpretability: Anomaly maps provide intuitive spatial visualization for the detected defects, making it easier for human verification and decision-making.

Flexible Architecture: The modular design caters to multiple defect types and product categories under one common framework, dispensing with the need for category-specific models and therefore maintenance.

Scalable Deployment: Computation (15-180 ms inference) has sufficient efficiency for real-time or near-real-time inspection. GPU memory requirement (950 MB) is within the range of modern edge devices, enabling direct-on-the-production-floor deployment.

6.5 Limitations

Despite strong performance, thesis acknowledges several limitations:

The PRO score of 61.41% is not yet up to the level of more than 90% for supervised methods, which is indicative of room for improvement in terms of reducing false positives at the region level. Performance, likewise, varied from category to category (Image AUROC ranging from .8489 to 1.0000). The same is also true with Pixel AUPR whose value represents an average of 64.26%, showing that there are challenges in obtaining the classification with considerable precision up to the pixel level. The computationally intensive component of Diffusion (180 ms) did not allow it to run in real time with high throughput lines. The performance of CLIP is partly determined by text prompts that are crafted by hand. The experiments are aimed at the MVTec AD that needs the validation on other datasets and real production conditions.

6.6 Future Research Directions

The thesis provides a number of avenues to future research:

1. **Adaptive Multi-Path Fusion:** Train fusion mechanisms that are a combination

of predictions across all paths with learned or dynamically computed weights. Fusion may be guided by confidence estimation where emphasis is laid on predictions that have high credibility.

2. **Category-Specific Path Selection:** This category classifier is developed to predict the best path to take based on visual features, depending on the category being considered.

3. **Few-Shot Fine-Tuning:** Explore the use of limited supervision in the form of prompt tuning, adapter layers or few-shot learning algorithms.

4. **Efficient Diffusion Sampling:** Learn about the current state of efficient and various other forms of diffusion sampling, such as consistency models, progressive distillation, and one-step generators.

5. **Automatic Prompt Optimization:** Coming up with ways of automatically generating or optimizing text prompts to CLIP-based detection.

6. **Cross-Domain Generalization:** : Research the generalization of framework to completely different industrial fields other than MVTec AD.

7. **Uncertainty Quantification:** Provide uncertainty estimates to anomaly predictions making it more trustworthy to use in industrial applications.

8. **Real-World Deployment Studies:** Carry out long experiments in real manufacturing settings that expose the real issues such as lighting differences, change in perspective, part-occlusions.

6.7 Concluding Remarks

The Quality Control in Industrial Manufacturing is the very important aspect where the usage of automated systems to identify anomalies in the production process might contribute to the quality of the product, less waste generation, and better performance. This thesis demonstrates that the state-of-the-art pre-trained vision models along with smart multi-path detection schemes can achieve a level of performance that is fairly similar to supervised ones in terms of practical zero-shot anomaly detection performance and fully eliminate training data requirements.

Multi-path architecture is a hybrid of Vision Transformer memory banks with conditioned diffusion models and the zero-shot detectiveness of CLIP interrogates an institutional shift in industrial anomaly detection processes that is no longer data-intensive supervised models, but can now be agile, instantly deployable zero-shot detectiveness.

The most significant alteration in that respect is, of course, the fact that now, the manufacturers may automate the quality control in various products lines without the prohibitive expenses of data collection.

Naturally, once all potential manufacturing sectors will become automated and will utilize the artificial intelligence, the zero-shot-based anomaly detection methods such as the presented framework will gain even greater significance in terms of ensuring the quality of products, reducing the number of defects, and, consequently, preserving a competitive edge.

Bibliography

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2022). Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9600.

Cao, Y., Xu, X., Sun, C., Cheng, Y., Du, Z., Gao, L., and Shen, W. (2023). Segment any anomaly without training via hybrid prompt regularization.

Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., and Boracchi, G. (2024). Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2021). Padim: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489.

Deng, H. and Zhang, Z. (2023). Anovl: Adapting vision-language models for unified zero-shot anomaly localization.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., and Dabeer, O. (2023). Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616.

Lee, Y. and Kang, P. (2022). Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 10:46717–46724.

Liu, J., Ma, Z., Wang, Z., Zou, C., Ren, J., Wang, Z., Song, L., Hu, B., Liu, Y., and Leung, V. C. (2025). A survey on diffusion models for anomaly detection.

Liu, Z., Zhou, Y., Xu, Y., and Wang, Z. (2023). Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. (2022). Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328.

Schwartz, E., Arbelle, A., Karlinsky, L., Harary, S., Scheidegger, F., Doveh, S., and Giryes, R. (2024). Maeday: Mae for few- and zero-shot anomaly detection.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Wu, D., Fan, S., Zhou, X., Yu, L., Deng, Y., Zou, J., and Lin, B. (2024). Unsupervised anomaly detection via masked diffusion posterior sampling.

Zhang, T., Gao, L., Li, X., and Gao, Y. (2025). Dzad: Diffusion-based zero-shot anomaly detection.

Zhang, X., Xu, M., and Zhou, X. (2024). Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection.

Zhou, Q., Pang, G., Tian, Y., He, S., and Chen, J. (2024). Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*.

MD. MAHFUZUR RAHMAN SHANTO
221-35-917

Dashboard

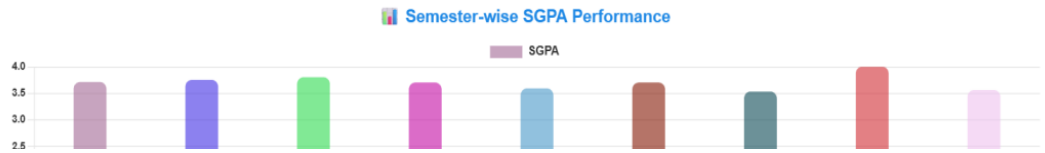
Student Portal

Total Payable	Total Paid	Total Due	Total Other
747,200.00	747,200.05	-0.05	4,100.00

Today's Routine - Thursday

No routine available for today.

Semester Wise Result



0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

221-35-917

ORIGINALITY REPORT

17%	15%	12%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Submitted to Daffodil International University Student Paper	1%
4	link.springer.com Internet Source	1%
5	Submitted to NCC Education Student Paper	1%
6	theses.hal.science Internet Source	1%
7	www.mdpi.com Internet Source	<1%
8	Submitted to Midlands State University Student Paper	<1%
9	Ta-Wei Tang, Hakiem Hsu, Kuan-Ming Li. "Industrial anomaly detection with multiscale autoencoder and deep feature extractor-based neural network", IET Image Processing, 2023 Publication	<1%
10	ojs.aaai.org Internet Source	<1%