

Received 20 January 2026, accepted 30 January 2026, date of publication 9 February 2026, date of current version 19 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3662352

RESEARCH ARTICLE

Multi-Scale Feature and Prompt Learning for Few-Shot Medical Image Anomaly Detection

ASIM NIAZ^{1,5}, MUHAMMAD UMRAIZ¹, SYED FARHAN ALAM ZAIDI², HYUN CHUL SONG³, FARHAN AKRAM⁴, AND KWANG NAM CHOI¹

¹Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea

²Premium Research Institute for Human Metaverse Medicine (WPI-PRIME), The University of Osaka, Suita, Osaka 565-0871, Japan

³Department of Virtual Reality and Convergence, Namseoul University, Cheonan 31020, South Korea

⁴Department of Pathology and Clinical Bioinformatics, Erasmus University Medical Center, 3015 GD Rotterdam, The Netherlands

⁵Institute for Assyriology and Hittitology, Ludwig-Maximilians-Universität München, 80539 Munich, Germany

Corresponding authors: Kwang Nam Choi (knchoi@cau.ac.kr) and Farhan Akram (f.akram@erasmusmc.nl)

This work was supported by the Ministry of Science and Information and Communication Technology (ICT) and National IT Industry Promotion Agency (NIPA) through the High Performance Computing (HPC) Support Project.

ABSTRACT Few-shot anomaly detection in medical imaging remains challenging due to the scarcity of labeled abnormal samples and the need for robust feature representations. We propose AnomaMed, a novel multi-scale feature and prompt learning framework that extends Contrastive Language-Image Pretraining (CLIP) for medical anomaly detection. We introduce Prompt Feature Fusion Modules, which integrate multi-scale image and prompt feature fusion at different layers of a pretrained CLIP model, enabling hierarchical feature refinement for improved anomaly localization. Additionally, we incorporate self-attention-based prompt learning, dynamically refining textual representations to enhance text-image alignment in few-shot settings. To further improve generalization, we introduce the Artificial Anomaly Blending Module (AABM), which synthetically generates diverse medical anomalies, enabling the model to learn a broader range of pathological patterns. We evaluate AnomaMed on three benchmark medical imaging datasets, demonstrating state-of-the-art performance over existing vision-language models for few-shot anomaly detection. Results confirm that AnomaMed significantly improves anomaly detection and localization, validating the effectiveness of multi-scale feature fusion, self-attentive prompt learning, and synthetic anomaly augmentation.

INDEX TERMS Medical image, prompt learning, vision language model, anomaly detection, few-shot learning.

I. INTRODUCTION

Advancements in image processing and computer vision have revolutionized the use of images and videos across agricultural [1], construction [2], industrial [3], [4], [5], and medical fields [6], [7]. In the medical domain, doctors increasingly rely on automated procedures for disease diagnostics; however, the accuracy of these algorithms varies. These methods can be broadly categorized into two groups: those following conventional segmentation and classification techniques, and those employing anomaly detection approaches [8]. Medical

Image Anomaly Detection (MIAD) uses AI to automatically identify abnormalities in medical images (e.g., X-rays, MRIs, CT scans) that may indicate diseases. It reduces human errors by consistently analyzing scans, detecting subtle anomalies, and minimizing misdiagnosis caused by fatigue, oversight, or judgment variability. However, obtaining medical image data is challenging due to privacy laws, the need for expert annotation, data imbalance, and hospital restrictions. Consequently, most AI models [9], [10], [11] are trained on normal patterns to detect and flag anomalies in medical images. Recently, many research studies [12], [13], [14], [15] have proven to be more accurate than professional clinicians in detecting anomalies; however, such models

The associate editor coordinating the review of this manuscript and approving it for publication was Adrian Stern¹.

require extensive datasets that must be densely annotated. A typical MIAD process is visually depicted in Fig. 1.

To overcome the time and labor costs of developing such models, we focus on AnomaMed, which can operate in few-shot settings without compromising detection performance. For each medical task, we provide a few normal images with artificially introduced anomalies during training. Although [16], [17] demonstrates the use of pre-trained CLIP knowledge for downstream classification tasks, fine-tuning it for specific tasks such as anomaly detection is impractical due to the extensive paired image-text data required for training. Additionally, fine-tuning CLIP for anomaly detection requires a deep understanding of prompt engineering to craft accurate text inputs, as even slight variations in word choice can compromise performance.

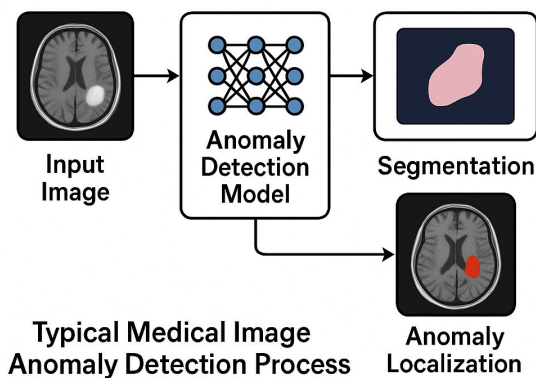


FIGURE 1. Typical medical image anomaly detection process. The workflow includes input image acquisition, processing by an anomaly detection model, segmentation of suspected regions, and final anomaly localization.

Inspired by [18], we introduce context optimization to automatically learn optimal prompt embeddings using self-attention-refined context vectors, enhancing text-image alignment for few-shot medical anomaly detection. Additionally, our Multi-Scale Image and Prompt Feature Fusion Modules integrate vision-language knowledge, improving hierarchical feature adaptation for anomaly detection and localization in medical imaging.

While MediCLIP [19] also leverages CLIP for anomaly detection, AnomaMed introduces multi-scale feature fusion and self-attention-based prompt learning, enabling dynamic prompt refinement instead of relying on static token representations. Moreover, AnomaMed integrates the AABM in the training phase to synthetically generate anomalies, improving model generalization in low-data settings. Previous models [4], [20], [21], [22] copy normal regions and paste them onto other normal regions to create disease-like patterns. However, this leads to model overfitting, and the anomalies do not appear realistic. Therefore, we utilized a diffusion model to blend artificial anomalies, mimicking real disease patterns more effectively. By incorporating self-attention and multi-scale feature fusion, AnomaMed

achieves better text-image alignment, anomaly localization, and interpretability in few-shot scenarios.

We validate AnomaMed on diverse datasets covering various medical imaging types, demonstrating superior performance over existing anomaly detection models. Results highlight its potential to assist medical professionals in real-world diagnostics. The main contributions of our work are:

- **AnomaMed**, a few-shot medical anomaly detection framework extending CLIP with multi-scale feature fusion and self-attentive prompt learning;
- **AABM** employing diffusion models to synthesize realistic anomalies and improve generalization;
- **Self-Attention-Based Prompt Learning** to dynamically refine textual prompts and enhance text-image alignment in few-shot settings; and
- **Prompt Feature Fusion Modules** to hierarchically integrate multi-scale image-text features for improved anomaly localization.

The subsequent sections present the related work and proposed methodology in detail, followed by the experimental evaluation, and conclude with the final remarks.

II. RELATED WORK

MIAD has gained significant attention due to its potential to assist clinicians in identifying abnormalities in medical images such as X-rays, MRIs, and ultrasound scans. The challenge of limited labeled data, particularly for abnormal cases, has driven research toward few-shot learning and vision-language models. This section reviews key works in MIAD, focusing on anomaly detection, vision-language pretraining, and synthetic data augmentation, while highlighting the gaps addressed by our proposed AnomaMed framework.

Early approaches to MIAD, such as those by Baugh et al. [9] and Naval Marimont and Tarroni [11], utilized unsupervised and self-supervised learning to detect anomalies without requiring extensive labeled datasets. Baugh et al. [9] proposed a multi-task learning approach to localize medical anomalies by synthesizing tasks, achieving robust performance on diverse imaging modalities. However, their method struggles with subtle anomalies due to limited feature granularity. Similarly, Naval Marimont and Tarroni [11] employed implicit field learning for unsupervised anomaly detection, but their approach lacks adaptability to few-shot settings, limiting its applicability in data-scarce scenarios.

Recent advancements have leveraged vision-language models, such as Contrastive Language-Image Pretraining (CLIP), to improve anomaly detection by aligning image and text representations. Tiu et al. [16] demonstrated expert-level pathology detection in chest X-rays using self-supervised CLIP-based learning, achieving high accuracy without annotated data. However, their method requires large-scale pretraining and struggles with fine-grained anomaly localization. Similarly, You et al. [17] introduced CXR-CLIP, a CLIP-based model for chest X-ray analysis, which excels in

classification but lacks mechanisms for precise localization in few-shot settings. MediCLIP [19], a CLIP adaptation for medical anomaly detection, improves performance by fine-tuning on medical datasets but relies on static prompt embeddings, limiting its flexibility for dynamic text-image alignment.

To address the scarcity of abnormal medical data, synthetic anomaly generation has emerged as a critical technique. Methods like CutPaste [4] and PatchSVDD [5] generate anomalies by copying and pasting normal regions to create disease-like patterns. While effective in some contexts, these approaches often produce unrealistic anomalies, leading to overfitting and poor generalization, as noted in our manuscript. AnoDDPM [13] employs Denoising Diffusion Probabilistic Models (DDPM) to generate anomalies, offering more realistic variations. However, AnoDDPM's anomalies lack medical specificity, reducing their effectiveness for MIAD. Cai et al. [10] proposed a dual-distribution discrepancy approach for chest X-ray anomaly detection, combining synthetic anomalies with self-supervised refinement, but their method is computationally intensive and less effective in low-data regimes.

Prompt learning has become a cornerstone for adapting vision-language models to specific tasks. Zhou et al. [18] introduced context optimization for vision-language models, enabling learnable prompt embeddings to improve downstream tasks. While effective for natural image classification, their approach does not account for the complex patterns in medical images. Our self-attention-based prompt learning builds on this by dynamically refining prompts with multi-head self-attention, enhancing text-image alignment for few-shot MIAD. Similarly, Xiang et al. [12] proposed SQUID, a deep feature in-painting method for anomaly detection, but it lacks the multi-scale feature integration critical for precise localization in medical imaging.

Despite these advancements, existing methods face challenges in few-shot MIAD, including limited adaptability to diverse medical imaging modalities, reliance on static prompts, and unrealistic synthetic anomalies. Our AnomaMed framework addresses these gaps by integrating multi-scale feature fusion, self-attention-based prompt learning, and the AABM. By leveraging DDPM with Perlin noise-based masks, AABM generates realistic anomalies, while multi-scale feature fusion and dynamic prompt learning enhance localization and generalization, outperforming prior methods like MediCLIP [19], AnoDDPM [13], and CutPaste [4] across diverse datasets.

III. PROPOSED METHOD

We propose AnomaMed, a multi-scale feature and prompt learning framework for few-shot anomaly detection. It integrates self-attention-based prompt learning, hierarchical feature fusion, and synthetic anomaly generation to enhance CLIP for improved anomaly localization. The AnomaMed pipeline is shown in Fig. 2.

A. PROBLEM STATEMENT

Detecting medical anomalies from images when only a few labeled examples are available is a difficult challenge. Medical images often contain subtle and diverse abnormalities, yet collecting large, well-annotated datasets is time-consuming and constrained by privacy and expert availability. Existing vision-language models like CLIP are powerful but not naturally tuned to capture these fine details or to adapt when training data is scarce. Our work addresses this gap by creating a framework that can learn from very limited abnormal samples, generate realistic synthetic anomalies to enrich training, and combine image and text information at multiple scales to accurately identify and localize anomalies across different types of medical images.

B. ARTIFICIAL ANOMALY BLENDING MODULE (AABM)

The AABM generates synthetic medical anomalies to mimic fractures, proliferative abnormalities, tumors, and cysts, enhancing anomaly detection in few-shot settings. AABM employs a Denoising Diffusion Probabilistic Model (DDPM) [23] to perturb normal images, introducing structured pathological variations. The forward diffusion process adds noise following $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1 - \alpha_t)I)$, while the reverse process reconstructs images via $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, where $\mu_\theta(x_t, t)$ is estimated by a neural network. Anomalies are introduced by modifying the variance using an anomaly strength factor s , leading to $p_\theta(x'_{t-1}|x_t) = \mathcal{N}(x'_{t-1}; \mu_\theta(x_t, t), (1 + s)\Sigma_\theta(x_t, t))$

where increasing s generates more distinct abnormalities resembling fractures, abnormal proliferations, and mass formations. The proposed method used pre-trained weights of the DDPM model. To ensure realistic anomaly placement, AABM applies Perlin noise-based masks to blend anomalies into normal images:

$$A = M \odot I + (1 - \delta)(M \odot I) + \delta(M \odot P) \quad (1)$$

where I is the normal image, P is the generated anomaly, M is the mask, $\odot$ is the element-wise multiplication operation and δ controls blending opacity. Adaptive thresholding ensures anomalies align with medically relevant regions. AABM image examples are presented by Fig. 3.

C. SELF-ATTENTION-BASED PROMPT LEARNING

CLIP relies on static prompt embeddings, which are manually designed or learned in isolation without contextual interactions between tokens. While effective for natural image classification, such prompts struggle to capture subtle and fine-grained variations in medical images, where textual alignment with complex pathological features is crucial. To address this limitation, we introduce Self-Attention-Based Prompt Learning, which enables dynamic refinement of prompt representations, improving text-image alignment for few-shot medical anomaly detection.

In our approach, the prompt structure follows the formulation of [18], $p = [V_1][V_2] \dots [V_M][CLS]$, where $[V]$

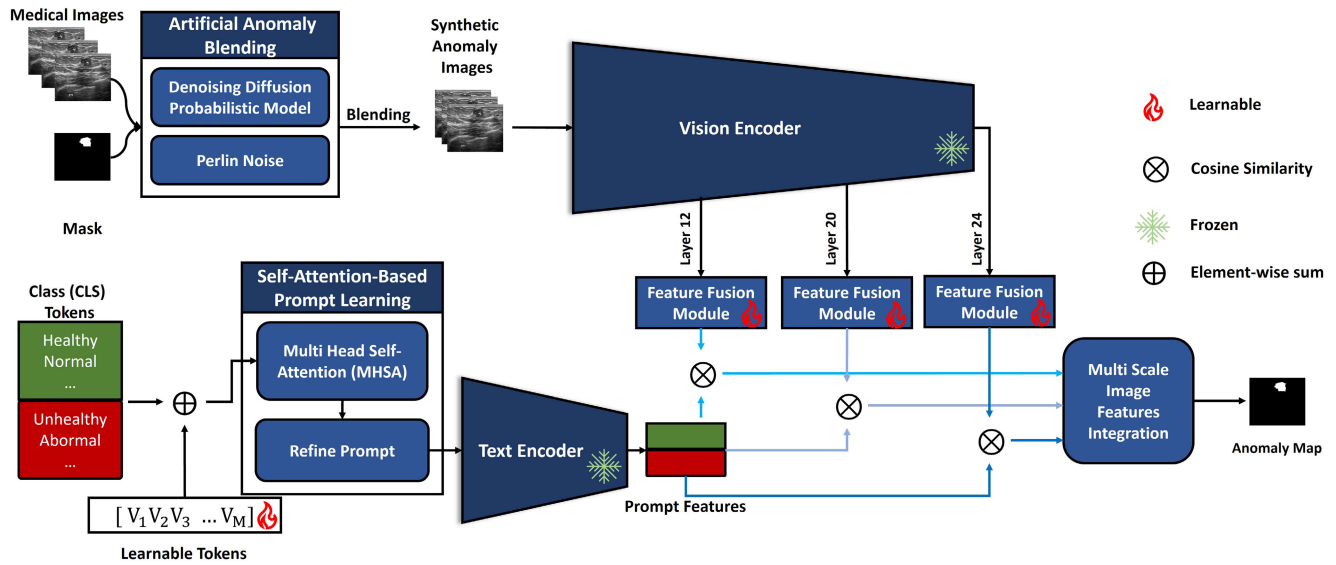


FIGURE 2. Complete workflow of the proposed AnomaMed framework for few-shot medical anomaly detection. The pipeline integrates (1) the Artificial Anomaly Blending Module (AABM) to synthesize diverse anomalies, (2) self-attention-based prompt learning for dynamic text-image alignment, and (3) multi-scale image-prompt feature fusion to refine hierarchical representations and produce accurate anomaly maps.

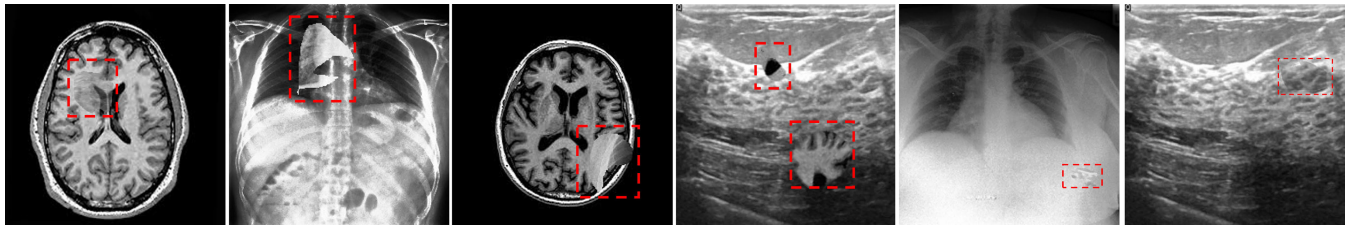


FIGURE 3. Visualization of medical anomaly localization results using AnomaMed.

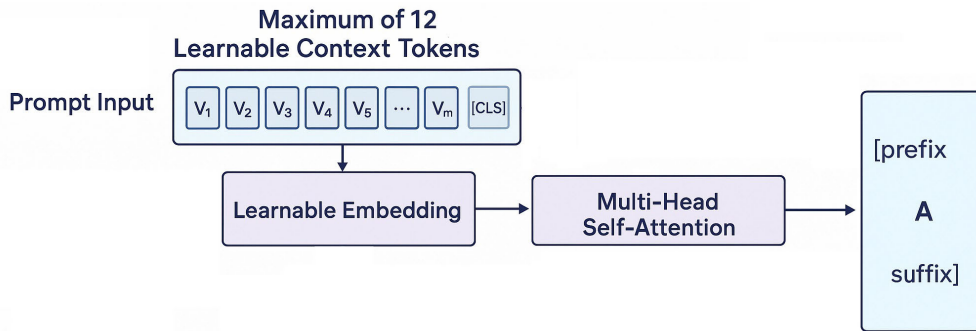


FIGURE 4. Overview of self-attention-based prompt learning, where learnable context tokens are refined via multi-head self-attention to produce enhanced prompts p' for improved text-image alignment in few-shot anomaly detection.

represents learnable context tokens, and $[CLS]$ is a fixed class token corresponding to medically relevant categories such as *healthy* and *disease*. Unlike static prompts, these learnable embeddings are optimized during training to capture complex medical patterns. To balance expressiveness and computational efficiency, we set the maximum number of learnable tokens to $M = 12$. This ensures sufficient

representation capacity while maintaining training stability and preventing overfitting.

To further enhance representation learning, we integrate a multi-head self-attention (MHSA) mechanism, allowing context-aware refinement of prompt embeddings. Given an input prompt embedding matrix $C \in \mathbb{R}^{n_{\text{ctx}} \times d}$, where n_{ctx} is the number of prompt tokens and d is the

embedding dimension, self-attention is applied as: $Q, K, V = CW_q, C W_k, C W_v$, and $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$, where Q, K, V represent the query, key, and value projections, and $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ are trainable weight matrices. This process enables the model to capture dependencies among prompt tokens, refining their contextual understanding. The final self-attention-enhanced prompt representation is formulated as:

$$p' = [\text{prefix}, A, \text{suffix}] \quad (2)$$

where A represents the attention-refined prompt embeddings, ensuring a more adaptive and semantically rich representation of medical features.

Incorporating self-attention into prompt learning enhances context-aware feature refinement, text-image alignment, and few-shot generalization, improving semantic consistency, accurate anomaly mapping, and reducing reliance on manual prompt design. Fig. 4 illustrates how learnable context tokens are refined through multi-head self-attention to produce enhanced prompt embeddings for improved text-image alignment in medical anomaly detection.

D. MULTI-SCALE IMAGE AND PROMPT FEATURE FUSION MODULES

The standard CLIP model is optimized for image-text classification but lacks the capability for precise anomaly detection and localization. To address this, we introduce Multi-Scale Image and Prompt Feature Fusion Modules, which enhance hierarchical feature extraction by incorporating multi-resolution spatial and contextual information from both image and text modalities. These modules are applied at layers 12, 20, and 24 of the CLIP image encoder, capturing both low-level structural variations and high-level semantic features to improve anomaly localization.

Given a medical image $I \in \mathbb{R}^{H \times W \times 3}$ from the support set D , we extract multi-scale feature representations using the CLIP vision encoder. The feature maps from the 12th, 20th, and 24th layers are denoted as:

$$\{F_{12}(I), F_{20}(I), F_{24}(I)\}, \quad F_j(I) \in \mathbb{R}^{H_j \times W_j \times C_j} \quad (3)$$

where H_j, W_j, C_j represent the height, width, and number of channels at layer j .

To integrate these multi-scale image features with textual prompt embeddings, we introduce a multi-scale feature fusion module $\Psi_j(\cdot)$, which maps each extracted feature into a shared representation space, aligning it with the self-attention-enhanced prompt embeddings A_n and A_a . The transformed feature representation is given by: $Z_j = \Psi_j(F_j(I))$, $Z_j \in \mathbb{R}^{H_j \times W_j \times C}$, where Z_j represents the adapted feature at layer j . The fusion process consists of a 1×1 convolution for dimensionality transformation, followed by a 3×3 convolution to incorporate spatial dependencies. A ReLU activation introduces non-linearity to enhance feature expressiveness.

To associate the transformed visual features with self-attention-refined prompt embeddings, we compute cosine

Algorithm 1 AnomaMed: Few-Shot Medical Anomaly Detection

Require: Normal image set I , query image I_q , text descriptions T_n, T_a , CLIP encoders f_v, f_t , learnable prompts $[V_1, \dots, V_M]$, training steps T , learning rate η

Ensure: Anomaly map S_a , anomaly score A_q

1: **Step 1: Artificial Anomaly Blending (AABM)**

2: **for** each $I \in D$ **do**

3: Generate synthetic anomaly via DDPM:

$$x'_{t-1} \sim \mathcal{N}(\mu_\theta(x_t, t), (1 + s)\Sigma_\theta(x_t, t))$$

4: Blend anomaly P into I :

$$A = M \odot I + \delta(M \odot P)$$

5: **end for**

6: **Step 2: Self-Attention-Based Prompt Learning**

7: Initialize prompts:

$$p = [V_1, \dots, V_M][\text{CLS}]$$

8: Compute embeddings C .

9: Apply self-attention:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

10: Refine prompt:

$$p' = [\text{prefix}, A, \text{suffix}]$$

11: **Step 3: Multi-Scale Feature Fusion**

12: **for** each $j \in \{12, 20, 24\}$ **do**

13: Extract CLIP features $F_j(I)$.

14: Transform via Ψ_j :

$$Z_j = \Psi_j(F_j(I))$$

15: Compute similarity with prompts:

$$S_{nj}, S_{aj} = \frac{\exp((Z_j, A)/\tau)}{\sum_A \exp((Z_j, A)/\tau)}$$

16: **end for**

17: Compute anomaly maps:

$$S_n, S_a = \frac{1}{3} \sum_j \text{Upsample}(S_{nj}, S_{aj})$$

18: **Step 4: Training and Optimization**

19: Compute loss:

$$\mathcal{L} = \mathcal{L}_{\text{focal}}([S_n, S_a], Y) + \mathcal{L}_{\text{dice}}(S_a, Y)$$

20: Update model:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$$

21: **Step 5: Inference**

22: Extract features $F_j(I_q)$, compute S_n, S_a .

23: Compute anomaly score:

$$A_q = \max(S_a)$$

24: **return** S_a, A_q .

similarity between Z_j and the textual representations:

$$S_{nj}^{(h,w)} = \frac{\exp\left((Z_j^{(h,w)} \otimes A_n)/\tau\right)}{\sum_{A \in \{A_n, A_a\}} \exp\left((Z_j^{(h,w)} \otimes A)/\tau\right)};$$

$$S_{aj}^{(h,w)} = \frac{\exp\left((Z_j^{(h,w)} \otimes A_a)/\tau\right)}{\sum_{A \in \{A_n, A_a\}} \exp\left((Z_j^{(h,w)} \otimes A)/\tau\right)} \quad (4)$$

where $\otimes$ represents cosine similarity, and τ is a temperature scaling parameter. The similarity matrices S_{nj} and S_{aj} are computed for each adapted multi-scale feature. To integrate multi-scale information, the similarity matrices from layers 12, 20, and 24 are upsampled to match the original input resolution and averaged as $S_n, S_a = \frac{1}{3} \sum_{j \in \{12, 20, 24\}} \text{Upsample}(S_{nj}, S_{aj})$, where the resulting multi-scale anomaly maps S_n and S_a highlight normal and anomalous regions, respectively.

To optimize the learnable self-attention-enhanced prompts and fusion modules, we define the objective function: $\mathcal{L} = \mathcal{L}_{\text{focal}}([S_n, S_a], Y) + \mathcal{L}_{\text{dice}}(S_a, Y)$, where $\mathcal{L}_{\text{focal}}$ represents Focal Loss to handle class imbalance, and $\mathcal{L}_{\text{dice}}$ represents Dice Loss for precise segmentation of anomaly regions. The notation $[S_n, S_a]$ denotes channel-wise concatenation.

During inference, the query image is processed through the Multi-Scale Image and Prompt Feature Fusion Modules at layers 12, 20, and 24, and the pixel-wise anomaly map is obtained from S_a , where the maximum value in S_a is used as the image-level anomaly score.

The overall training and inference procedure of the proposed AnomaMed framework is summarized in Algorithm 1.

IV. EXPERIMENTS

This section describes the datasets used to train and evaluate our model. We then present both qualitative and quantitative results, along with graphical patterns observed in the data. The section concludes with ablation studies.

A. DATASETS AND IMPLEMENTATION DETAILS

We evaluated AnomaMed on three datasets: BrainMRI [24], CheXpert [25], and BUSI [26]. BrainMRI contains 2D brain MRI images of normal and tumor cases. CheXpert includes 224,316 chest radiographs from 65,240 patients, labeled with 14 diagnostic categories as positive, negative, or uncertain. BUSI consists of 780 ultrasound images from 600 female patients (ages 25–75), collected in 2018, categorized as normal, benign, or malignant, with ground truth annotations for breast cancer detection and segmentation.

We implemented AnomaMed using PyTorch with a fixed random seed of 100, training and evaluating on images resized to 224×224 . The CLIP model, based on ViT-L-14 with features extracted from layers 12, 20, and 24, incorporates AABM for synthetic anomaly generation and self-attention-based context optimization prompt learning

with a maximum of 12 learnable tokens. The key hyperparameters for the proposed method are summarized in Table 1.

TABLE 1. Key hyperparameters for the proposed approach.

Parameter	Value / Setting
AABM (Artificial Anomaly Blending Module)	
Diffusion Model	Pre-trained DDPM
Noise Schedule	$\bar{\alpha}_t$ as in [24]
Anomaly Strength Factor (s)	Controls severity of synthetic anomalies
Blending Mask	Perlin noise-based mask
Blending Opacity (δ)	Controls visibility of blended anomalies
AnomaMed Configuration	
Vision Backbone	ViT-L-14
Multi-scale Layers	12, 20, 24
Image Size	224×224
Learnable Prompt Tokens (M)	12
Training Setup	
Epochs	200
Batch Size	8
Random Seed	100

B. RESULTS

Fig. 5 illustrates the medical anomaly localization results obtained using AnomaMed, organized in three columns. Each column displays representative images from a specific dataset along with the corresponding localization results. The first column shows samples from the BUSI dataset, the second from CheXpert, and the third from BrainMRI.

Table 2 reports the few-shot anomaly detection performance (Image-AUROC %) of AnomaMed and baseline methods across three medical imaging datasets: BrainMRI, CheXpert, and BUSI, under varying shot settings ($k = 4, 8, 16, 32$).

AnomaMed consistently achieves the best results across all datasets and settings. On BrainMRI, it reaches up to 98.0% AUROC with just 32 training samples, outperforming the next-best method (MediCLIP) by a notable margin. Similar trends are observed on CheXpert and BUSI, where AnomaMed achieves 75.1% and 93.0% AUROC, respectively, at $k = 32$. Notably, AnomaMed shows strong performance even in low-shot settings ($k = 4$), highlighting its robustness in data-scarce scenarios.

These results confirm that the multi-scale feature fusion, self-attention-based prompt learning, and AABM module work together to provide a robust and effective framework for few-shot medical anomaly detection across diverse imaging modalities, delivering 2–3 % gains over MediCLIP on BrainMRI and BUSI and about 1–2 % on the harder CheXpert set, and demonstrating strong cross-modality generalization even in low-shot settings.

Fig. 6 visually complements the tabulated results by highlighting the consistent performance gains of AnomaMed across datasets and increasing shot counts, making overall trends and method comparisons easier to interpret.

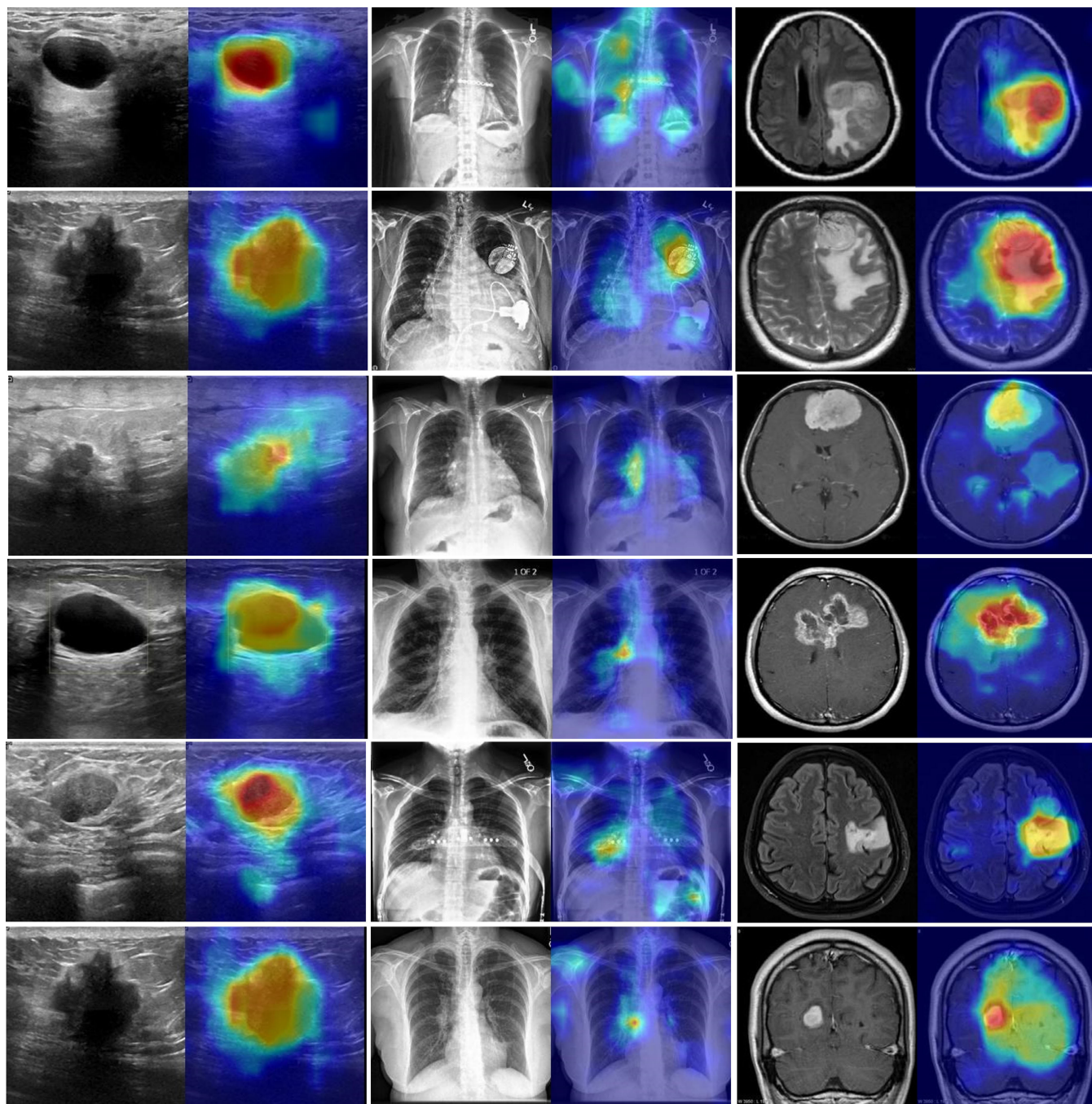


FIGURE 5. Visualization of medical anomaly localization results using AnomaMed.

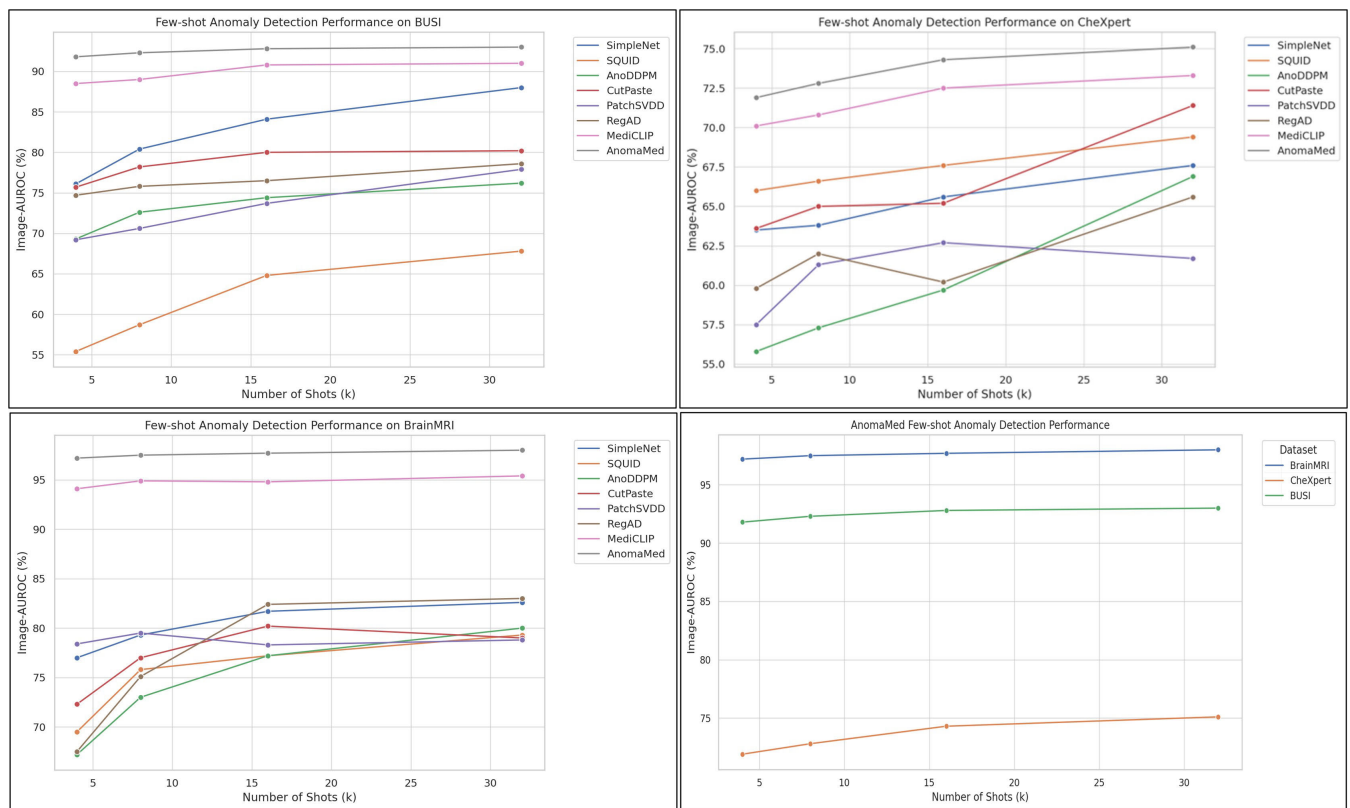
AnomaMed demonstrates superior performance in few-shot anomaly localization on the BUSI dataset, consistently surpassing existing methods in effectively localizing lesion regions, as presented in Table 3.

Compared with baseline anomaly detection approaches, which primarily rely on single-scale visual features or handcrafted anomaly patterns, AnomaMed demonstrates more stable and accurate performance across datasets and shot settings. This improvement can be plausibly

explained by the proposed multi-scale feature fusion, which captures both fine-grained structural variations and higher-level contextual cues essential for detecting subtle medical abnormalities. In addition, the self-attention-based prompt learning mechanism dynamically refines textual representations, enabling more effective alignment between image regions and anomaly-related semantics, particularly in low-shot scenarios. The Artificial Anomaly Blending Module further enhances generalization by exposing

TABLE 2. Few-shot anomaly detection performance (Image-AUROC %): Comparison of AnomaMed and other methods across three datasets.

Dataset	k	SimpleNet	SQUID	AnoDDPM	CutPaste	PatchSVDD	RegAD	MediCLIP	AnomaMed
BrainMRI	4	77.0±3.0	69.5±4.4	67.2±1.8	72.3±1.8	78.4±1.4	67.5±1.0	94.1±0.7	97.2±0.2
	8	79.3±2.9	75.8±1.9	73.0±3.6	77.0±4.9	79.5±0.8	75.1±2.8	94.9±0.4	97.5±0.7
	16	81.7±4.4	77.2±0.6	77.2±2.5	80.2±2.7	78.3±3.0	82.4±1.6	94.8±0.5	97.7±0.3
	32	82.6±0.9	79.3±1.7	80.0±2.1	79.0±4.3	78.8±1.4	83.0±2.4	95.4±0.4	98.0±0.4
CheXpert	4	63.5±2.4	66.0±1.9	55.8±1.3	63.6±2.8	57.5±1.1	59.8±1.4	70.1±1.6	71.9±1.2
	8	63.8±0.7	66.6±3.9	57.3±3.5	65.0±2.9	61.3±1.5	62.0±3.0	70.8±1.5	72.8±1.1
	16	65.6±3.7	67.6±4.7	59.7±2.8	65.2±3.2	62.7±2.3	60.2±1.7	72.5±1.7	74.3±0.8
	32	67.6±1.1	69.4±2.2	66.9±3.2	71.4±2.0	61.7±1.4	65.6±1.4	73.3±1.3	75.1±0.6
BUSI	4	76.1±2.5	55.4±2.2	69.3±1.3	75.7±1.4	69.2±1.1	74.7±5.0	88.5±0.6	91.8±0.4
	8	80.4±4.8	58.7±1.6	72.6±4.6	78.2±2.1	70.6±2.4	75.8±2.7	89.0±0.3	92.3±0.2
	16	84.1±2.7	64.8±4.5	74.4±2.9	80.0±0.9	73.7±6.0	76.5±1.2	90.8±0.4	92.8±0.3
	32	88.0±0.3	67.8±0.4	76.2±4.0	80.2±1.7	77.9±1.6	78.6±1.2	91.0±0.3	93.0±0.2

**FIGURE 6.** Few-shot anomaly detection performance across different datasets and methods. The plots illustrate Image-AUROC (%) versus the number of shots (k) for BrainMRI, CheXpert, and BUSI datasets. The bottom-right panel focuses exclusively on AnomaMed's performance.**TABLE 3.** Few-shot anomaly localization performance (Pixel-AUROC %): Comparison of AnomaMed and other methods on the BUSI dataset.

k	SimpleNet	SQUID	AnoDDPM	PatchSVDD	RegAD	MediCLIP	AnomaMed
4	73.5±4.9	56.8±5.1	68.3±1.9	74.4±0.9	71.9±3.1	88.0±0.3	90.94±0.2
8	77.6±3.4	56.8±1.2	74.7±4.3	73.6±3.4	76.1±2.6	88.6±1.0	91.18±0.8
16	78.2±1.5	65.5±2.8	79.5±1.4	78.3±2.2	74.6±1.0	89.9±0.9	91.73±0.7
32	79.8±0.4	68.4±2.2	79.1±1.3	77.6±3.4	76.3±0.4	90.1±1.0	92.11±0.7

the model to diverse and realistic synthetic anomalies during training, reducing overfitting commonly observed

in conventional augmentation strategies. Together, these components explain the consistent performance gains of

AnomaMed over baseline methods across different medical imaging modalities.

C. ABLATION STUDIES

To evaluate the efficacy of the key components in AnomaMed, we conduct ablation studies on the BrainMRI dataset using Image-AUROC (%) as the evaluation metric (Table 4). The analysis focuses on two core components: self-attention-based prompt learning and multi-scale image feature fusion.

We first examine the effect of prompt token length while fixing the image encoder to the final (24th) layer. Increasing the token length from 8 to 12 improves performance from 94.28% to 94.71%, indicating that more expressive prompt representations better capture anomaly-related semantics. This validates the necessity of sufficient learnable context tokens for effective prompt learning.

We then evaluate multi-scale feature fusion by varying encoder layer combinations with the token length fixed at 12. Multi-layer configurations consistently outperform single-layer or shallow-layer settings. The best result of 97.77% is achieved using layers 12, 20, and 24, confirming that combining mid-level structural and high-level semantic features is critical for accurate anomaly detection. The performance drop observed with simpler configurations demonstrates that multi-scale fusion is an essential component. In addition to component-level ablations, Table 5 reports zero-shot anomaly detection results across different train–test dataset combinations using a fixed support set of $k = 16$. These results provide further insight into the generalization ability of AnomaMed when applied to unseen medical imaging domains.

TABLE 4. Performance of various token lengths and image encoder layers on BrainMRI with Image-AUROC (%).

Token Length	Encoder Layer	Image-AUROC
8	24	94.28%
12	24	94.71%
12	16, 20, 24	96.11%
12	4, 12, 18, 24	96.45%
12	14, 17, 24	96.46%
12	6, 12, 18, 24	96.60%
12	12, 18, 24	96.97%
12	12, 20, 24	97.77%

TABLE 5. Zero-shot anomaly detection results of AnomaMed with $k = 16$ support set.

Train Task	Test Task	Image-AUROC (%)
BrainMRI	CheXpert	69.4±0.6
	BrainMRI	97.7±0.3
	BUSI	86.8±0.9
CheXpert	CheXpert	74.3±0.8
	BrainMRI	95.0±0.5
	BUSI	90.8±0.4
BUSI	CheXpert	72.4±0.5
	BrainMRI	94.3±0.8
	BUSI	92.9±0.6

Overall, the ablation results confirm that both self-attention-enhanced prompt representations and multi-scale feature fusion are necessary for achieving robust few-shot anomaly detection in AnomaMed.

V. DISCUSSION

While AnomaMed achieves strong performance in few-shot medical anomaly detection, several limitations merit discussion. First, the framework relies on a CLIP backbone pretrained on natural images, which may limit its ability to capture fine-grained medical semantics, especially for rare or subtle abnormalities. Incorporating medical-specific vision–language pretraining could further enhance domain alignment.

Second, although the AABM module improves generalization, the generated anomalies are not explicitly guided by anatomical or pathological priors, which may affect realism in complex clinical cases. Future work could integrate anatomy-aware or pathology-conditioned diffusion models to address this limitation.

Finally, AnomaMed currently operates on 2D images and assumes a fixed data distribution. Extending the framework to 3D and longitudinal imaging data, as well as incorporating domain-adaptive or continual learning strategies, would improve robustness and facilitate deployment in evolving clinical environments.

VI. CONCLUSION

This study explores a novel avenue in MIAD under few-shot settings, introducing AnomaMed, a multi-scale feature and prompt learning framework. By integrating Prompt Feature Fusion Modules, self-attention-based prompt learning, and the AABM, our approach enhances anomaly localization and generalization. AnomaMed outperforms existing models on BrainMRI, CheXpert, and BUSI, covering a broad range of medical imaging modalities, including MRI, X-ray, and ultrasound. Our findings validate its effectiveness, establishing AnomaMed as a promising clinical tool for medical diagnostics, paving the way for real-world applications in healthcare. Looking ahead, future work will focus on adopting a report-based vision–language framework that leverages descriptive radiological reports instead of class-level labels, enabling joint diagnosis and explanation generation. Further extensions include applying AnomaMed to 3D and longitudinal medical imaging, integrating domain-adaptive and continual learning for evolving clinical settings, and developing anatomy-aware anomaly synthesis to improve the realism and specificity of generated abnormalities.

ACKNOWLEDGMENT

Asim Niaz and Muhammad Umraiz are co-first authors.

REFERENCES

- [1] M. Umraiz, A. Khan, and H. Kim, “Real time fruit segmentation using multi-scale context aggregation,” *Control Robot Syst.*, pp. 34–36, 2020.
- [2] R. Hussain, A. Sabir, D.-Y. Lee, S. F. A. Zaidi, A. Pedro, M. S. Abbas, and C. Park, “Conversational AI-based VR system to improve construction safety training of migrant workers,” *Autom. Construct.*, vol. 160, Apr. 2024, Art. no. 105315.

- [3] A. Niaz, M. Umraiz, S. Soomro, and K. N. Choi, "Vision transformer and mamba-attention fusion for high-precision PCB defect detection," *PLoS ONE*, vol. 20, no. 9, Sep. 2025, Art. no. e0331175.
- [4] C. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9659–9669.
- [5] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 375–390.
- [6] A. Niaz, A. A. Memon, K. Rana, A. Joshi, S. Soomro, J. S. Kang, and K. N. Choi, "Inhomogeneous image segmentation using hybrid active contours model with application to breast tumor detection," *IEEE Access*, vol. 8, pp. 186851–186861, 2020.
- [7] A. Niaz, E. Iqbal, F. Akram, J. Kim, and K. N. Choi, "Self-initialized active contours for microscopic cell image segmentation," *Sci. Rep.*, vol. 12, no. 1, p. 14947, Sep. 2022.
- [8] A. Niaz, S. Ul Amin, S. Soomro, H. Zia, and K. Nam Choi, "Spatially aware fusion in 3D convolutional autoencoders for video anomaly detection," *IEEE Access*, vol. 12, pp. 104770–104784, 2024.
- [9] M. Baugh, J. Tan, J. P. Müller, M. Dombrowski, J. Batten, and B. Kainz, "Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2023, pp. 162–172.
- [10] Y. Cai, H. Chen, X. Yang, Y. Zhou, and K.-T. Cheng, "Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102794.
- [11] S. N. Marimont and G. Tarroni, "Implicit field learning for unsupervised anomaly detection in medical images," in *Proc. MICCAI*, Strasbourg, France, Sep. 2021, pp. 189–198.
- [12] T. Xiang, Y. Zhang, Y. Lu, A. L. Yuille, C. Zhang, W. Cai, and Z. Zhou, "SQUID: Deep feature in-painting for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23890–23901.
- [13] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 650–656.
- [14] Y. Cai, H. Chen, X. Yang, Y. Zhou, and K. Cheng, "Dual-distribution discrepancy for anomaly detection in chest X-Rays," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2022, pp. 584–593.
- [15] K. Zhou, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, Z. Gu, J. Liu, and S. Gao, "Encoding structure-texture relation with P-Net for anomaly detection in retinal images," in *Proc. ECCV*, Aug. 2020, pp. 360–377.
- [16] E. Tiu, E. Talus, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1399–1406, Sep. 2022.
- [17] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "CXR-CLIP: Toward large scale chest X-ray language-image pre-training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2023, pp. 101–111.
- [18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [19] X. Zhang, M. Xu, D. Qiu, R. Yan, N. Lang, and X. Zhou, "MediCLIP: Adapting CLIP for few-shot medical image anomaly detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2024, pp. 458–468.
- [20] J. Sato, Y. Suzuki, T. Wataya, D. Nishigaki, K. Kita, K. Yamagata, N. Tomiyama, and S. Kido, "Anatomy-aware self-supervised learning for anomaly detection in chest radiographs," *iScience*, vol. 26, no. 7, Jul. 2023, Art. no. 107086.
- [21] J. Tan, B. Hou, J. Batten, H. Qiu, and B. Kainz, "Detecting outliers with foreign patch interpolation," 2020, *ArXiv:2011.04197*.
- [22] Tan, "Detecting outliers with Poisson image interpolation," in *Proc. MICCAI*, Sep. 2024, pp. 581–591.
- [23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.
- [24] N. Chakrabarty. (2019). *Brain Mri Images for Brain Tumor Detection*. [Online]. Available: <https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection>
- [25] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 590–597.
- [26] W. Al-Dhabyani, M. Goma, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.



ASIM NIAZ received the B.S. degree in electrical (computer) engineering from the COMSATS Institute of Information Technology, Wah Cantt., Pakistan, in 2016, and the M.S. and Ph.D. degrees in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2020 and 2025, respectively.

He was a Research Assistant with the Visual Image Media Laboratory and later visited INRIA Sophia Antipolis, France, followed by a research internship with Ericsson Canada, Montreal, Canada. He is currently a Research Assistant with the AI Vision Laboratory, Chung-Ang University, and has collaborated with the Urban Design and Analytics Laboratory, Department of Urban Design and Studies, Chung-Ang University. He is also working remotely for the Institute for Assyriology and Hittology, Ludwig-Maximilians-Universität München, Munich, Germany. His research interests include action recognition, video understanding, anomaly detection in images and videos, remote sensing, medical image analysis, and image segmentation.



MUHAMMAD UMRAIZ received the bachelor's degree in electrical engineering from the COMSATS Institute of Information Technology, Wah Cantt., in 2017, and the M.S. degree in electronics and information engineering from Jeonbuk National University, Jeonju, South Korea, in 2020. He is currently a Research Assistant with the AI Vision Laboratory, Chung-Ang University, Seoul, South Korea. His research interests include medical image segmentation, anomaly detection, and vision language models.



SYED FARHAN ALAM ZAIDI was born in Lahore, Pakistan, in 1993. He received the B.S. degree in information technology from the University of Education, Lahore, in 2014, the M.S. degree in computer science from COMSATS University Islamabad, Pakistan, in 2017, and the Ph.D. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea.

After completing the B.S. degree, he was a Web Developer with a software development company in Pakistan. After completing the M.S. degree, he was a Visiting Lecturer with various universities. During the Ph.D. studies, he was a Research Assistant with the Real-Time Software Engineering Laboratory, Chung-Ang University. He completed his assignment as a Postdoctoral Researcher with Chung-Ang University. He is currently an Assistant Professor with the Premium Research Institute for Human Metaverse Medicine (WPI-PRIME), The University of Osaka, Suita, Osaka, Japan. His research interests include machine learning, deep learning, computer vision, medical imaging, natural language processing, data mining, and image processing.



HYUN CHUL SONG received the B.S. and M.S. degrees in computer science and engineering from Chung-Ang University, South Korea, and the Ph.D. degree, focused on deep learning from Chung-Ang University, in 2019. He is currently with the Department of Virtual Reality, Namseoul University, South Korea. His research interests include object detection, deep learning, and their applications in intelligent systems.



KWANG NAM CHOI received the B.S. and M.S. degrees from the Department of Computer Science, Chung-Ang University, Seoul, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in computer science from the University of York, U.K., in 2002. He is currently a Professor with the School of Computer Science and Engineering, Chung-Ang University. His current research interests include motion tracking, object categorization, and 3-D image recognition.

...



FARHAN AKRAM received the Ph.D. degree in computer science from Rovira i Virgili University, in 2017. His postdoctoral work with A*STAR, Singapore, focused on AI-driven medical image analysis, including cancer recurrence prediction and prostate cancer grading.

At Khalifa University, United Arab Emirates, he applied deep learning for Alzheimer's disease prediction. Currently, he leads the PHANTOM Group at Erasmus University Medical Center, Rotterdam, where he develops AI models for analyzing H and E, IHC, and IF whole-slide images, with applications in cancer prognostication and molecular subtyping. He also works on AI-based kidney transplant rejection prediction and explores 3-D pathology to enhance tumor biology insights. He holds a university teaching qualification, reflecting his dedication to education and mentorship.