Medical Anomaly Detection Using Generative Adversarial Network With Self Attention Mechanism

Fahim Abrar Fuad
200041109
Saqif Kaisar
200041121
Tanvir Hassan Ananta
200041153

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Fahim Abrar Fuad**, **Saqif Kaisar**, and **Tanvir Hassan Ananta** under the supervision of **Dr. Md. Hasanul Kabir**, Professor, Department of Computer Science and Engineering and co-supervision of **Md. Bakhtiar Hasan**, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Md. Hasanul Kabir
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Md. Bakhtiar Hasan
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Fahim Abrar Fuad
Student ID: 200041109
Date: September 29, 2025

Saqif Kaisar
Student ID: 200041121
Date: September 29, 2025

Tanvir Hassan Ananta
Student ID: 200041153
Date: September 29, 2025

Dedicated to our supervisor, co supervisor and the faculty members of Computer Vision lab group with deepest respect and gratitude, whose guidance and encouragement were invaluable to the completion of this thesis.

Contents

Acknowledgement

We are profoundly grateful to our supervisor, Dr. Md. Hasanul Kabir, and our co-supervisor, Md. Bakhtiar Hasan, for their invaluable guidance, patience, and encouragement throughout the course of this research. Their insightful feedback, constructive critiques, and unwavering support have been instrumental in shaping both this thesis and our academic journey.

We would also like to express our heartfelt thanks to our parents for their constant love, encouragement, and understanding. Their support has been a source of strength and motivation at every stage of this work.

To all of you, we extend our deepest appreciation.

Abstract

Anomaly detection in medical imaging plays a vital role in assisting early diagnosis and treatment. Traditional supervised methods rely heavily on annotated abnormal samples, which are often scarce and diverse, making them impractical for real-world deployment. This work explores unsupervised anomaly detection using a range of GAN-based frameworks, where models learn to capture the distribution of normal data and identify deviations without the need for labeled outliers. Various architectures, including reconstruction-based and feature-matching approaches, are evaluated and extended with enhancements such as self-attention mechanisms and positional encodings to improve spatial feature learning. Extensive experiments on medical imaging datasets demonstrate that the proposed techniques significantly improve detection performance, highlighting the effectiveness of adversarial learning for unsupervised medical anomaly detection.

Chapter 1

Introduction

Recent advancements in deep learning have significantly transformed medical image analysis, particularly in the detection and classification of diseases. X-ray imaging, a cost-effective and routine diagnostic tool, plays a critical role in identifying thoracic conditions such as pneumonia, tuberculosis, lung cancer, pulmonary edema, and other pathologies. However, the interpretation ofx-rays is often hindered by subjective variability, radiologist fatigue, and the subtle nature of early pathological signs. These limitations underscore the need for automated, robust tools to assist clinicians by flagging potentially abnormal scans for further review.

Unsupervised anomaly detection offers a promising approach to address these challenges. Unlike supervised methods that rely on extensive labeled datasets—which are often costly or difficult to obtain in medical contexts—unsupervised models learn the distribution of normal images and identify deviations, or anomalies, that may indicate disease. These models provide binary or continuous scores to indicate whether a scan appears normal or abnormal, making them valuable as pre-screening or triaging tools to prioritize cases for expert evaluation and reduce diagnostic delays.

(a) **(b)**

Figure 1.1: Examples of X-ray scans showing abnormalities.

1.1 Motivations and Scope

Unsupervised anomaly detection has demonstrated considerable promise in medical imaging, particularly for identifying rare or subtle pathologies without relying on large, annotated datasets. However, many existing approaches face significant limitations in effectively capturing the rich spatial and structural information inherent in X-ray images. Subtle, localized abnormalities in lung structures, such as minor opacities, nodules, or irregular tissue patterns, can easily be overlooked when models focus predominantly on global patterns or low-level texture features. These shortcomings reduce the clinical reliability of automated detection systems and underscore the need for models capable of preserving both local and global anatomical context. Motivated by these challenges, our work aims to develop anomaly detection models that are more spatially aware, enabling them to identify clinically meaningful deviations with higher precision and produce results that are interpretable and actionable for healthcare professionals.

The scope of this study encompasses the design, implementation, and evaluation of a generalized anomaly detection framework tailored for X-ray imaging. The primary focus is on improving spatial sensitivity and ensuring that the framework can generalize across diverse pathological conditions while maintaining robustness against noise and variability present in real-world clinical datasets. By integrating mechanisms to enhance spatial feature learning, this work seeks to bridge the gap between conventional unsupervised models and the nuanced requirements of practical medical image analysis.

1.2 Problem Statement

The central problem addressed in this thesis is the limited ability of unsupervised anomaly detection models to effectively utilize spatial information in X-ray images. Conventional convolutional networks can extract local features but often fail to capture global anatomical context or long-range dependencies. This limitation reduces sensitivity to subtle, distributed anomalies and increases the risk of overfitting to low-level textures that are not clinically meaningful.

To address this problem, we pursued the following objectives:

1. **Design and evaluate** a spatially aware GAN-based anomaly detection framework forx-rays that moves beyond conventional CNN-based backbones.
2. **Investigate architectural enhancements** aimed at improving spatial percep-

tion, including attention mechanisms, positional encodings, and Swin Transformer-based encoder–decoder structures.

3. **Experimentally assess** the effectiveness of these enhancements against strong baseline models to determine their impact on anomaly detection accuracy and interpretability.
4. **Analyze limitations and trade-offs** of spatial mechanisms in unsupervised anomaly detection, providing insights to guide future improvements in medical image analysis.

1.3 Research Challenges

Developing a spatially aware unsupervised anomaly detection model forx-rays presents several domain-specific challenges:

- **Limited Supervision and Scarce Labeled Data**: In medical imaging, large, high-quality labeled datasets are rare. Unsupervised models must learn normative patterns from healthy images alone, making it difficult to reliably detect rare or subtle pathologies.
- **Capturing Complex Spatial Dependencies**:x-rays contain intricate anatomical structures with long-range relationships. Standard convolutional models often miss subtle distributed anomalies, and incorporating mechanisms to model global context without losing local detail is a key challenge.
- **Generalization Across Diverse Pathologies**: Medical anomalies vary widely in shape, size, and location. Designing a model that can detect a broad spectrum of anomalies without overfitting to specific patterns requires flexible and robust representations.
- **Architectural Trade-offs**: Incorporating advanced spatial modeling into GAN frameworks introduces challenges related to model stability, training complexity, and computational efficiency, requiring careful design and optimization.

1.4 Contributions

This thesis makes several contributions to the field of unsupervised anomaly detection in medical imaging, specifically forx-rays:

- **Development of a spatially aware anomaly detection framework**: A generalized model was designed and implemented to improve sensitivity to anatomical context and better detect subtle abnormalities, moving beyond conventional CNN-based approaches.
- **Systematic evaluation of spatial modeling strategies**: The research explored the potential and limitations of incorporating spatial awareness mechanisms in GAN-based models. This evaluation provides valuable insights into how different approaches impact detection performance and interpretability.
- **Comprehensive experimental analysis**: The framework was rigorously tested across a variety of X-ray images and compared against strong baseline models. Although performance improvements were not consistently observed, the analysis highlights key challenges and trade-offs in embedding spatial mechanisms in unsupervised models.
- **Guidance for future research**: By identifying practical limitations and challenges in current GAN-based anomaly detection frameworks, this work establishes a foundation for subsequent studies aiming to enhance model generalization, stability, and interpretability in clinical applications.

These contributions advance the understanding of spatial awareness in unsupervised anomaly detection and provide a clearer perspective on the practical challenges of deploying such models in medical imaging. The insights gained through this work are valuable for both the development of more robust anomaly detection frameworks and the broader adoption of automated X-ray analysis in clinical settings.

Chapter 2

Related Works

The development of unsupervised anomaly detection for medical imaging, particularlyx-rays, has gained significant attention due to its potential to address the scarcity of annotated data and the need for rapid, accurate screening in clinical settings. This section reviews key contributions in the field, organized into thematic categories that reflect major research directions: GAN-based approaches for anomaly detection, spatial awareness in deep learning, and enhancements in GAN stability and performance. These themes are explored in relation to *Brainomaly* [10], the base paper for this thesis, which focuses on unsupervised neurologic disease detection in brain MRIs and serves as a foundation for extending similar techniques to X-ray analysis with an emphasis on spatial awareness. The following papers—*GANomaly* [1], *HealthyGAN* [12], *Brainomaly* [10], *SAGAN[15]* [15], *Wasserstein GAN* [2], and *pix2pix* [4]—are introduced within these categories to provide a cohesive overview before detailing their methodologies.

2.1 GAN-based Approaches for Anomaly Detection

Generative Adversarial Networks (GANs) have emerged as a powerful framework for unsupervised anomaly detection by learning normative data distributions and identifying deviations as anomalies. This theme is critical for medical imaging, where annotated datasets are often limited.

GANomaly [1] introduces a semi-supervised approach using an encoder-decoder-encoder architecture to detect anomalies in general imaging tasks, laying the groundwork for applying GANs to medical contexts. This model improved upon earlier GAN-based methods like AnoGAN by incorporating an encoder, which significantly speeds up

inference and avoids iterative latent optimization, addressing the computational bottleneck of previous approaches.

Similarly, *HealthyGAN* [12] extends GAN-based methods to medical imaging by leveraging unannotated datasets containing both healthy and diseased samples. It employs one-directional image-to-image translation to map mixed data to healthy representations. HealthyGAN builds on GANomaly by demonstrating how unannotated clinical data can be effectively incorporated, improving generalization and enabling models to handle real-world medical datasets with mixed conditions. These works align with *Brainomaly* [10] by addressing the challenge of detecting anomalies without extensive labels, offering insights into adapting GAN-based methods for X-ray analysis in resource-constrained clinical environments.

2.2 Spatial Awareness in Deep Learning for Medical Imaging

The importance of spatial information in medical imaging cannot be overstated, as many pathologies manifest in specific anatomical regions or exhibit characteristic geometric patterns. This theme focuses on methods that enhance spatial modeling to improve anomaly detection.

SAGAN[15] [15] introduces self-attention mechanisms to capture long-range dependencies in images, enabling models to focus on semantically relevant regions. Compared to Brainomaly, SAGAN enhances spatial awareness by integrating attention modules that allow the model to more effectively capture fine-grained contextual and anatomical details, improving detection performance in distributed or subtle anomalies.

Brainomaly [10], the base paper, exemplifies the integration of spatial awareness in unsupervised anomaly detection, using unannotated brain MRIs to identify neurologic diseases by leveraging spatial relationships. It progresses on GANomaly and HealthyGAN by combining spatial attention and memory mechanisms, showing how anatomical consistency and contextual information can improve interpretability and localization of anomalies. Together, these works underscore the potential of spatially aware models to enhance the sensitivity and interpretability of anomaly detection systems, directly supporting the thesis objective of developing such models for thoracic imaging.

2.3 Enhancing GAN Stability and Performance

The reliability of GAN-based anomaly detection models depends on stable training and robust performance, particularly in the complex domain of medical imaging. This theme explores methods to improve GAN training and data handling.

Wasserstein GAN [2] proposes the use of Wasserstein distance to enhance training stability and mitigate mode collapse, a critical advancement for ensuring consistent performance in medical applications. This improvement provides a more reliable foundation for subsequent models like Brainomaly and SAGAN, enabling them to benefit from more stable training dynamics and improved convergence properties.

Additionally, *pix2pix* [4] offers a framework for image-to-image translation, which, while primarily designed for tasks like photo synthesis, can support data augmentation to address data scarcity in medical imaging. By facilitating better data representation and transformation, pix2pix complements earlier GAN-based anomaly detection methods and contributes to more robust and generalizable models in clinical imaging contexts.

2.4 Paper Details

2.4.1 AnoGAN[9]

Anomaly detection is a fundamental task in machine learning and computer vision, especially for applications such as medical diagnosis, industrial inspection, and cybersecurity. Traditional supervised learning methods rely on labeled examples of both normal and anomalous instances. However, in many real-world scenarios, anomalies are rare and diverse, making them difficult to represent comprehensively during training. To address this limitation, AnoGAN (Anomaly detection with Generative Adversarial Networks)[9] was introduced as a novel unsupervised method for detecting anomalies without requiring examples of anomalies during training.

Why AnoGAN Was Developed

Anomaly Detection Challenge: Anomaly detection involves identifying rare or unusual instances in a dataset that deviate significantly from the norm. This is critical in domains like medical imaging (e.g., detecting tumors in MRI scans), cybersecurity (e.g., identifying network intrusions), and industrial monitoring (e.g., detecting defects in manufacturing).

High-Dimensional Data: Traditional anomaly detection methods (e.g., one-class SVM, isolation forests) often struggle with high-dimensional data like images because they fail to capture complex, non-linear patterns effectively. For example, in medical imaging, normal images (healthy scans) have intricate patterns that are hard to model using simple statistical methods.

Motivation for Deep Learning: Deep learning, particularly convolutional neural networks (CNNs), excels at learning hierarchical features from high-dimensional data like images. However, most deep learning methods are supervised and require labeled anomalous data, which is often scarce or unavailable in anomaly detection tasks where only normal data is typically available.

Leveraging GANs for Unsupervised Anomaly Detection

Generative Adversarial Networks (GANs): Introduced by Goodfellow et al. (2014), GANs are powerful generative models that learn to approximate a data distribution $p_X(x)$ by training a generator G to produce realistic samples from random noise z, while a discriminator D distinguishes real data from generated samples.

Key Components and Methodology of AnoGAN

Training Phase

- AnoGAN uses a standard GAN architecture comprising a generator (G) and a discriminator (D).
- The generator learns to map latent variables z from a prior distribution (e.g., Gaussian) to realistic data samples $G(z)$.
- The discriminator is trained to distinguish between real data samples and synthetic ones generated by G.
- The GAN is trained only on normal data, enabling it to capture the distribution of normal patterns.

Anomaly Detection Phase

- Given a test data point x, the goal is to determine how well it fits within the learned distribution of normal data.
- GANs do not provide an explicit likelihood for a data point, nor do they directly support mapping data x back to a corresponding latent vector z.

- To address this, AnoGAN introduces an optimization-based approach: it performs iterative gradient descent to find a latent vector z^* such that the generated sample $G(z^*)$ is as close as possible to the input data x.
- The similarity between x and $G(z^*)$ is computed using:
  - **Residual Loss:** Measures pixel-level difference between x and $G(z^*)$.
  - **Discrimination Loss:** Uses features from the discriminator to compare the high-level representations of x and $G(z^*)$.
- A combined anomaly score is calculated from these losses—higher scores indicate that x deviates significantly from the normal distribution and is likely an anomaly.

Limitations of AnoGAN

AnoGAN's design has several drawbacks, which motivated the development of ALAD:

- **Computational Inefficiency:** The generator inversion step (finding z^*) requires iterative optimization for each test sample, involving hundreds of gradient descent steps. Each iteration requires backpropagation through both G and D, making the process computationally expensive. For example, the ALAD paper notes that AnoGAN takes 10,496 ms per sample on the SVHN dataset, compared to ALAD's 10.5 ms (Table IV).
- **Lack of Encoder:** AnoGAN does not train an encoder during the GAN training phase, so it cannot directly map x to z. This necessitates the costly optimization process at test time.
- **Sensitivity to Hyperparameters:** The balance parameter λ and the number of optimization steps can significantly affect performance, requiring careful tuning.
- **Reconstruction Quality:** The generator may not perfectly reconstruct normal samples, especially if the GAN training does not fully converge or if the normal data distribution is highly complex.
- **Feature Space Limitations:** AnoGAN uses pixel-wise reconstruction error $||x - G(z)||_1$, which can be sensitive to small shifts or noise in images and may not capture semantic differences effectively.

2.4.2 Adversarially Learned Anomaly Detection[14]

Adversarially Learned Anomaly Detection (ALAD) [14] is a framework that extends the capabilities of GAN-based models for unsupervised anomaly detection. ALAD builds upon the limitations observed in earlier models like AnoGAN by introducing an encoder-decoder structure within an adversarial training framework. This approach not only enhances computational efficiency but also improves the accuracy of anomaly detection by learning meaningful representations and reconstruction mappings jointly.

Motivation

The primary motivation for ALAD stems from the need for efficient and effective anomaly detection in high-dimensional datasets. Traditional GAN-based methods like AnoGAN, while innovative, suffer from computational inefficiencies due to their reliance on iterative optimization at test time. For instance, AnoGAN [9] requires hundreds of gradient descent steps to map a test sample back to the latent space, making it impractical for real-time applications or large-scale datasets.

Additionally, high-dimensional data such as images or network traffic logs often exhibit complex, non-linear patterns that require sophisticated modeling techniques beyond simple statistical methods.

ALAD addresses these issues by introducing an encoder network that maps data to the latent space in a single feed-forward pass, drastically reducing inference time. The authors were also motivated by the need for stability in GAN training, which is notoriously difficult due to issues like mode collapse and vanishing gradients. By incorporating spectral normalization and cycle-consistency constraints, ALAD ensures stable training and better modeling of the normal data distribution.

The ultimate goal is to provide a scalable, unsupervised anomaly detection method that can handle diverse datasets, from tabular data (e.g., KDD99 network intrusion dataset) to images (e.g., CIFAR-10, SVHN), with competitive performance.

Methodology

The ALAD (Adversarially Learned Anomaly Detection) framework enhances traditional GAN-based anomaly detection by incorporating bidirectional mapping and cycle-consistency, enabling efficient inference and robust reconstructions. It is inspired by the Bidirectional GAN (BiGAN) and Adversarially Learned Inference (ALI) architectures, with critical modifications tailored for the anomaly detection task.

The architecture of ALAD comprises three core components:

- **Encoder Network** $E(x)$: Maps input data $x \in \mathcal{X}$ to a latent representation $z \in \mathcal{Z}$.
- **Generator Network** $G(z)$: Maps latent vectors z back to the data space to produce reconstructions $\hat{x} = G(z)$.
- **Multiple Discriminator Networks**: Ensure the joint consistency and realism of both forward and reverse mappings.

ALAD trains the encoder and generator jointly within an adversarial framework by matching various joint distributions using discriminators.

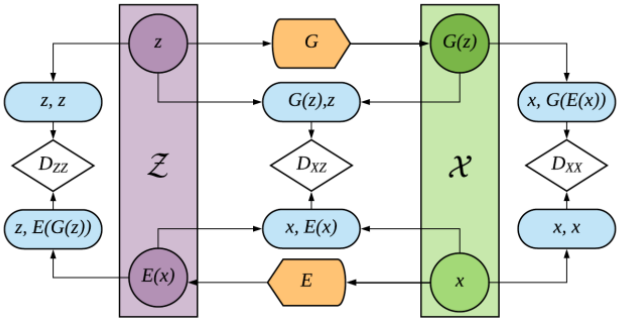


Figure 2.1: The GAN used in Adversarially Learned Anomaly Detection. Dzz, Dxz and Dxx denote discriminators (white), G the generator (orange), and E the encoder (orange); these networks are simultaneously learned during training.

Joint Distribution Matching The key idea is to match joint distributions between real and generated pairs using three discriminators:

- D_{xz}: Distinguishes between joint pairs $(x, E(x))$ and $(G(z), z)$, encouraging consistency between the encoder and generator.
- D_{xx}: Compares the original input x and its reconstruction $G(E(x))$, enforcing the generator to reconstruct realistic samples.
- D_{zz}: Ensures that latent vectors are cycle-consistent, i.e., z should match $E(G(z))$, promoting meaningful latent representations.

Each discriminator is trained to distinguish real versus fake pairs, and the encoder-generator pair is trained adversarially to fool the discriminators.

Cycle-Consistency Objectives ALAD introduces cycle-consistency as a regularization mechanism. This ensures that:

$$x \approx G(E(x)) \quad \text{and} \quad z \approx E(G(z))$$

This not only improves reconstruction fidelity but also promotes semantic consistency between the latent and data spaces, which is essential for anomaly detection.

Training Losses The overall training objective consists of multiple adversarial losses:

$$\mathcal{L}_{xz} = \mathbb{E}_{x\sim p_X(x)}[\log D_{xz}(x, E(x))] + \mathbb{E}_{z\sim p_Z(z)}[\log(1 - D_{xz}(G(z), z))]$$

$$\mathcal{L}_{xx} = \mathbb{E}_{x\sim p_X(x)}[\log D_{xx}(x, x)] + \mathbb{E}_{x\sim p_X(x)}[\log(1 - D_{xx}(x, G(E(x))))]$$

$$\mathcal{L}_{zz} = \mathbb{E}_{z\sim p_Z(z)}[\log D_{zz}(z, z)] + \mathbb{E}_{z\sim p_Z(z)}[\log(1 - D_{zz}(z, E(G(z))))]$$

The generator G and encoder E are trained to minimize the above losses, effectively fooling all three discriminators into believing that the generated and encoded pairs are real.

Inference and Anomaly Scoring During inference, anomaly detection is performed by evaluating how well a test sample x can be reconstructed via the encoder and generator:

- The test input is passed through the encoder to obtain the latent code: $z = E(x)$.
- The generator reconstructs the sample: $\hat{x} = G(z)$.

An anomaly score $A(x)$ is computed based on the reconstruction error:

$$A(x) = \|x - G(E(x))\|_1$$

Optionally, feature-level differences (e.g., discriminator-based perceptual distances) can also be included to capture semantic mismatches. A high reconstruction error indicates that the sample does not conform to the learned normal data distribution and is thus flagged as an anomaly.

How ALAD Solves Challenges of AnoGAN

AnoGAN, a pioneering GAN-based anomaly detection method, has several limitations that ALAD directly addresses:

- **Computational Inefficiency:** AnoGAN requires iterative optimization to find the latent vector z^* for a test sample x, taking 10,496 ms per sample on SVHN. ALAD introduces an encoder E that maps x to z in a single feed-forward pass, reducing inference time to 10.5 ms—a 1000x speedup.
- **Lack of Encoder:** AnoGAN does not train an encoder, while ALAD trains E jointly, enabling direct mapping and eliminating costly optimization at test time.
- **Poor Reconstruction Quality:** AnoGAN uses pixel-wise reconstruction error, which can be sensitive to noise. ALAD uses feature-based anomaly scores derived from D_{xx}, improving robustness.
- **Training Instability:** ALAD stabilizes training by incorporating spectral normalization and using multiple discriminators with cycle-consistency objectives.
- **Performance:** ALAD achieves better performance (e.g., AUROC of 0.5753 on SVHN compared to AnoGAN's 0.5410), leveraging cycle-consistency and a robust anomaly score.

In summary, ALAD overcomes AnoGAN's key challenges by introducing an encoder for efficiency, using feature-based scoring for robustness, and incorporating spectral normalization and cycle-consistency for stable training and better performance.

Comparative Table with AnoGAN

Table 2.1: Comparison between AnoGAN and ALAD across key dimensions.

Feature	**AnoGAN**	**ALAD**
Inference Time	Slow (requires optimization)	Fast (encoder enables one-shot inference)
Latent Inversion	Requires iterative search	Direct via encoder
Reconstruction Quality	Moderate	Improved via cycle-consistency
Discriminator Use	Single D	Multiple: D_{xz}, D_{xx}, D_{zz}
Anomaly Score	Residual + Discriminator feature loss	Reconstruction error (optionally with perceptual loss)
Training Complexity	Moderate	Higher (more networks), but stable

2.4.3 Adversarially Learned One-Class Classifier for Novelty Detection[7]

Novelty detection is a cornerstone of computer vision and machine learning, focusing on identifying observations that deviate significantly from a known target class, referred to as inliers, while labeling these deviations as outliers or anomalies. This task is vital in applications such as anomaly detection in surveillance videos, outlier removal in image datasets, and denoising in visual data processing. Unlike traditional classification, novelty detection often operates under the constraint of having only positive (target) class data during training, with negative (novelty) class data either

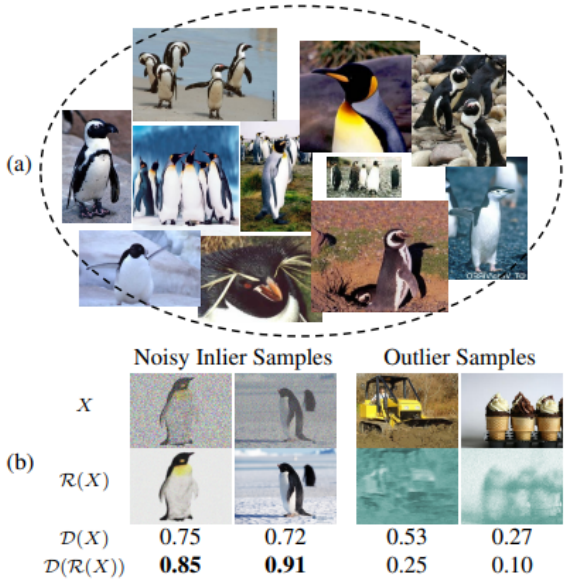


Figure 2.2: Result of using D(x) and D(R(x))

absent, poorly sampled, or undefined. This scenario aligns with one-class classification, where the objective is to model the target class's distribution and reject non-conforming samples. The paper introduces a pioneering end-to-end deep learning framework [7] to tackle this challenge, leveraging a dual-network architecture comprising $\mathcal{R}$ and $\mathcal{D}$, trained adversarially to learn the target class's intrinsic characteristics. The framework excels in detecting novelties, as demonstrated through experiments on MNIST, Caltech-256, and UCSD Ped2 datasets, offering a robust and generalizable solution for one-class classification tasks.

What Challenges It Solved

Novelty detection poses significant challenges due to the absence of negative class data during training, complicating the definition of boundaries between inliers and outliers. Traditional one-class classification methods, such as statistical modeling or self-representation techniques, rely on handcrafted features or shallow models that struggle to capture the complex, high-dimensional distributions of visual data like images and videos. These approaches falter in real-world scenarios where data is noisy, outliers are diverse, or the target class exhibits intricate variations. Deep learning holds promise, but training end-to-end deep networks for one-class classification is challenging without negative class samples. The proposed framework addresses these issues by introducing an end-to-end deep learning approach inspired by GANs. It employs $\mathcal{R}$ and $\mathcal{D}$ to model the target class without negative class data. The $\mathcal{R}$ network enhances inlier samples and distorts outliers, improving separability, while $\mathcal{D}$ detects novelties based on the target class distribution. The method achieves robust

performance across applications like outlier detection in MNIST and Caltech-256 and anomaly detection in UCSD Ped2, with added robustness to noise via Gaussian noise in training.

Proposed Methodology

The methodology introduces an end-to-end deep learning framework for one-class classification, tailored for novelty detection. It consists of two components:

- **Reconstructor ($\mathcal{R}$)**: A convolutional encoder-decoder CNN that preprocesses inputs by reconstructing target class samples with high fidelity while distorting outliers, enhancing separability.
- **Discriminator ($\mathcal{D}$)**: A sequence of convolutional layers that distinguishes original target class samples from reconstructed ones, modeling the target class distribution (p_t).

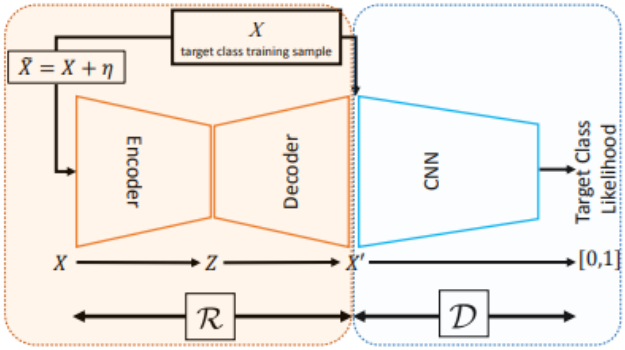


Figure 2.3: Overview of the proposed structure for one-class classification framework

During training, only target class data is used, with Gaussian noise added to inputs for robustness. The adversarial training involves $\mathcal{R}$ producing reconstructions to fool $\mathcal{D}$, while $\mathcal{D}$ rejects reconstructed samples. In testing, $\mathcal{D}$ evaluates either the original input ($\mathcal{D}(X)$) or the reconstructed input ($\mathcal{D}(\mathcal{R}(X))$), with the latter offering superior performance. The framework's generality enables it to address outlier detection in MNIST and Caltech-256 and anomaly detection in UCSD Ped2, leveraging both networks collaboratively.

Training Procedure

The training procedure optimizes $\mathcal{R}$ and $\mathcal{D}$ concurrently using only target class data, following a GAN-inspired adversarial paradigm. It is formulated as a two-player min-

imax game with the objective function:

$$\mathcal{L} = \mathcal{L}_{\mathcal{R}+\mathcal{D}} + \lambda \mathcal{L}_{\mathcal{R}}$$

The adversarial loss is:

$$\min_{\mathcal{R}} \max_{\mathcal{D}} \left(\mathbb{E}_{X \sim p_t}[\log(\mathcal{D}(X))] + \mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma}[\log(1 - \mathcal{D}(\mathcal{R}(\tilde{X})))] \right)$$

where $X \sim p_t$ are target class samples, $\tilde{X} = X + \eta$ includes Gaussian noise ($\eta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$), and $\mathcal{L}_{\mathcal{R}} = \|X - X'\|^2$ ensures $\mathcal{R}$'s outputs ($X' = \mathcal{R}(\tilde{X})$) resemble inliers. The hyperparameter $\lambda = 0.4$ balances terms. Training stops when $\mathcal{R}$ reconstructs inliers with minimal error ($\|X - X'\|^2 < \rho$). Implemented in TensorFlow on an NVIDIA TITAN X, this ensures mature network weights without overtraining.

R and D Architecture

The framework's effectiveness relies on the architectures of $\mathcal{R}$ and $\mathcal{D}$.

$\mathcal{R}$ Network Architecture

The $\mathcal{R}$ network is a convolutional encoder-decoder CNN with encoding (convolutional) and decoding (deconvolutional) layers. Each layer's properties include kernel dimensions, input channels, and output channels. Batch regardless of the specific page number. Batch normalization is applied after each layer, and pooling layers are omitted for spatial information preservation. This enables $\mathcal{R}$ to reconstruct inliers (e.g., enhancing noisy penguin images) and distort outliers (e.g., non-digit "1" samples). Gaussian noise during training ensures robustness, akin to a denoising autoencoder.

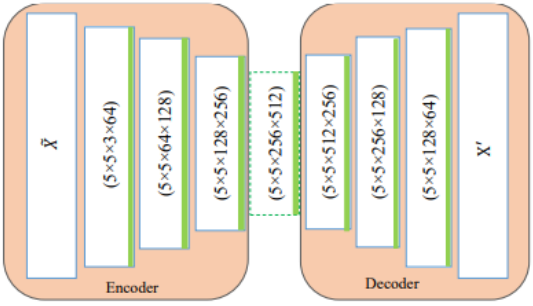


Figure 2.4: R network

$\mathcal{D}$ Network Architecture

The $\mathcal{D}$ network is a sequence of convolutional layers outputting a scalar likelihood score for the target class distribution (p_t). It distinguishes original samples from reconstructions, with the final layer producing a value in [0,1]. Convolutional layers capture complex visual patterns, effective for image and video data. Both architectures use batch normalization ($\epsilon = 10^{-6}$, decay = 0.9) for stability across MNIST, Caltech-256, and UCSD Ped2.

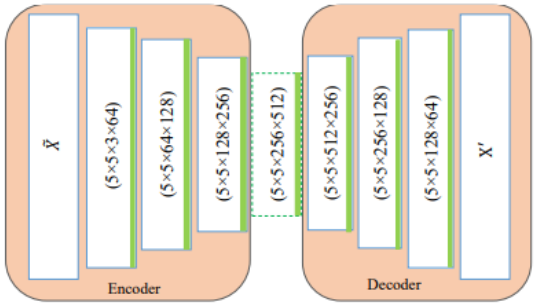


Figure 2.5: D network

OCC1 and OCC2

Two classification strategies are proposed:

OCC1: Using Only $\mathcal{D}$

OCC1 uses $\mathcal{D}$ to classify inputs:

$$\text{OCC}_1(X) = \begin{cases} \text{Target Class} & \text{if } \mathcal{D}(X) > \tau, \\ \text{Novelty (Outlier)} & \text{otherwise,} \end{cases}$$

where τ is a threshold. It outperforms methods like LOF and DRAE on MNIST and achieves high AUC/F_1-scores on Caltech-256, but relies on original inputs, potentially reducing separability.

OCC2: Using $\mathcal{R}$ and $\mathcal{D}$

OCC2 preprocesses inputs with $\mathcal{R}$:

$$\text{OCC}_2(X) = \begin{cases} \text{Target Class} & \text{if } \mathcal{D}(\mathcal{R}(X)) > \tau, \\ \text{Novelty (Outlier)} & \text{otherwise.} \end{cases}$$

$\mathcal{R}$ enhances inliers and distorts outliers, improving separability. OCC2 yields higher F_1-scores on MNIST, better AUC/F_1 on Caltech-256, and a lower EER (13%) on UCSD Ped2 vs. OCC1's 16%. Visualizations show $\mathcal{D}(\mathcal{R}(X))$ scores have a smaller reject region.

Anomaly Calculation

Anomaly calculation determines if an input is an inlier or outlier using $\mathcal{D}$'s output, with or without $\mathcal{R}$ preprocessing:

1. **Input Processing**: OCC1 feeds X to $\mathcal{D}$, outputting $\mathcal{D}(X) \in [0, 1]$. OCC2 processes X through $\mathcal{R}$, then $\mathcal{D}$ evaluates $\mathcal{R}(X)$.
2. **Likelihood Scoring**: Higher scores indicate inliers; lower scores suggest outliers (e.g., normal UCSD Ped2 patches score high).
3. **Threshold-Based Classification**: If $\mathcal{D}(X) > \tau$ (OCC1) or $\mathcal{D}(\mathcal{R}(X)) > \tau$ (OCC2), the input is an inlier; otherwise, an anomaly.
4. **Evaluation Metrics**: OCC2 outperforms OCC1 with higher F_1-scores on MNIST, better AUC/F_1 on Caltech-256, and lower EER (13%) on UCSD Ped2. Visualizations confirm better separability for $\mathcal{D}(\mathcal{R}(X))$.
5. **Practical Considerations**: Noise robustness and $\mathcal{R}$'s outlier distortion enhance real-world anomaly detection.

OCC2's collaborative approach demonstrates the framework's strength in novelty detection without negative class data.

2.4.4 GANomaly[1]

Anomaly detection in medical imaging, particularly in X-ray analysis, is a critical task for identifying conditions such as infections, lung diseases, and other abnormalities. The challenge lies in the scarcity of abnormal samples, making supervised learning approaches less feasible. The GANomaly framework, introduced by Akcay et al. in

their paper "GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training,"[1] addresses this challenge through a semi-supervised learning paradigm using a conditional Generative Adversarial Network (GAN). This report provides a comprehensive overview of the GANomaly model, its methodology, contributions, experimental results, and its relevance to anomaly detection in X-ray imaging as a preprocessing tool. The framework's ability to model normal data distributions and identify anomalies as outliers makes it a promising approach for enhancing clinical workflows by prioritizing anomalous images for further diagnostic evaluation.

Background and Problem Statement

Anomaly detection involves identifying instances that deviate from a learned norm, a task particularly challenging when datasets are heavily biased toward normal samples. In X-ray imaging, normal images are abundant, but abnormal cases—such as those indicating pneumonia, tumors, or fractures—are rare and diverse, complicating the modeling of abnormal classes. Traditional supervised methods require large, labeled datasets, which are often unavailable for anomalies. Semi-supervised approaches, which train only on normal samples and detect anomalies as outliers, offer a solution. The GANomaly framework leverages adversarial training to address this problem, focusing on learning the distribution of normal X-ray images and identifying anomalies based on deviations from this distribution. This aligns with the preprocessing goal of distinguishing healthy from anomalous images, enabling radiologists to focus on cases requiring further investigation.

Architecture

The GANomaly framework employs a novel encoder-decoder-encoder pipeline within a conditional Generative Adversarial Network (GAN) framework, designed for semi-supervised anomaly detection. The architecture, as illustrated in Figure 2, comprises three sub-networks: a generator, a second encoder, and a discriminator, each tailored to learn the distribution of normal data and detect anomalies as outliers.

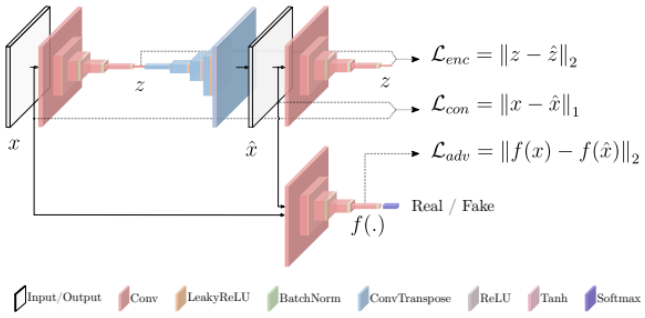


Figure 2.6: Ganomaly network pipeline

Generator (G) The generator operates as a bow-tie autoencoder, consisting of an encoder (G_E) and a decoder (G_D). It learns to represent and reconstruct input images. For an input image $x \in \mathbb{R}^{w \times h \times c}$, the encoder G_E uses convolutional layers, batch normalization, and Leaky ReLU activation to compress x into a latent vector $z \in \mathbb{R}^d$, known as bottleneck features, which capture the most compact representation of x. The decoder G_D, inspired by the Deep Convolutional GAN (DCGAN) generator [6], employs convolutional transpose layers, ReLU activation, batch normalization, and a final tanh layer to upscale z and reconstruct the image as $\hat{x} = G_D(z)$, where $z = G_E(x)$. Formally, the generator produces:

$$\hat{x} = G_D(G_E(x)). \tag{2.1}$$

This sub-network aims to replicate the input image, learning the manifold of normal data during training.

Second Encoder (E) The second encoder E is a unique component of GANomaly, compressing the reconstructed image $\hat{x}$ into a latent vector $\hat{z} = E(\hat{x})$, with the same dimensionality as z for consistent comparison. It shares the architectural details of G_E (convolutional layers, batch normalization, Leaky ReLU) but uses different parametrization. Unlike traditional autoencoders that rely on bottleneck features for latent space minimization, E explicitly learns to minimize the distance between z and $\hat{z}$, enhancing anomaly detection by comparing latent representations. This sub-network is critical during testing, where the anomaly score is derived from the dissimilarity between z and $\hat{z}$.

Discriminator (D) The discriminator D follows the standard DCGAN architecture [6], classifying input images x as real and reconstructed images $\hat{x}$ as fake. It uses convolutional layers to evaluate image realism, focusing on distinguishing normal data

from generated outputs. The discriminator employs feature matching, comparing intermediate layer representations of real and generated images, to stabilize training and guide the generator toward producing realistic images.

Training Method

GANomaly's training method optimizes the generator and discriminator to model the distribution of normal samples, enabling anomaly detection by identifying deviations in latent and image spaces. The training leverages a composite loss function combining adversarial, contextual, and encoder losses, each targeting specific aspects of the sub-networks. The hypothesis is that the generator, trained only on normal samples, fails to reconstruct abnormalities, leading to dissimilar latent representations that flag anomalous inputs.

Adversarial Loss To stabilize GAN training and reduce instability, GANomaly uses a feature matching loss, as proposed by Salimans et al. [8]. This loss minimizes the $\mathcal{L}_2$ distance between the intermediate feature representations of the discriminator for real and generated images. Let f be a function that outputs an intermediate layer of the discriminator D for an input x drawn from the data distribution p_X. The adversarial loss is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim p_X} \left| f(x) - \mathbb{E}_{x \sim p_X} f(G(x)) \right|_2 . \tag{2.2}$$

This loss updates the generator based on the discriminator's internal representation, encouraging the production of realistic images without relying solely on the discriminator's real/fake output.

Contextual Loss To ensure that the generator captures contextual information about the input data, a contextual loss penalizes the $\mathcal{L}_1$ distance between the input image x and the reconstructed image $\hat{x} = G(x)$. Following Isola et al. **Isola2017**, the $\mathcal{L}_1$ norm is used to produce less blurry results compared to $\mathcal{L}_2$. The contextual loss is defined as:

$$\mathcal{L}_{con} = \mathbb{E}_{x \sim p_X} \left| x - G(x) \right|_1 . \tag{2.3}$$

This loss optimizes the generator to maintain contextual fidelity, ensuring that reconstructed normal images closely resemble their inputs.

Encoder Loss To enforce consistency in the latent space, an encoder loss minimizes the $\mathcal{L}_2$ distance between the bottleneck features of the input ($z = G_E(x)$) and the

encoded features of the generated image ($\hat{z} = E(G(x))$). This loss is defined as:

$$\mathcal{L}_{enc} = \mathbb{E}_{x \sim p_X} \left| G_E(x) - E(G(x)) \right|_2 . \tag{2.4}$$

The encoder loss ensures that the generator and second encoder learn to produce similar latent representations for normal samples. For anomalous inputs, the generator's inability to reconstruct abnormalities results in dissimilar z and $\hat{z}$, which is leveraged for anomaly detection.

Overall Objective The generator's overall objective function combines the three losses with weighting parameters to balance their contributions:

$$\mathcal{L} = w_{adv}\mathcal{L}_{adv} + w_{con}\mathcal{L}_{con} + w_{enc}\mathcal{L}_{enc}, \tag{2.5}$$

where w_{adv}, w_{con}, and w_{enc} are empirically determined weights adjusting the impact of each loss. The paper suggests $w_{adv} = 1$, $w_{con} = 50$, and $w_{enc} = 1$ to prioritize contextual fidelity while maintaining adversarial and latent space consistency **Akcay2018**.

The training process alternates between optimizing the generator to minimize $\mathcal{L}$ and updating the discriminator to distinguish real images from generated ones. By training only on normal samples, GANomaly models the normal data manifold, enabling the detection of anomalies as outliers during testing based on the encoder loss.

Contributions

GANomaly makes several key contributions to anomaly detection:

- **Novel Architecture**: The encoder-decoder-encoder pipeline integrates adversarial autoencoders, capturing both image and latent space distributions. This contrasts with prior methods like AnoGAN, which require computationally expensive two-stage training, and EGBAD, which uses joint training but lacks the explicit second encoder.
- **Efficiency**: By avoiding two-stage training and leveraging joint learning, GANomaly achieves superior computational performance, with runtimes significantly lower than AnoGAN (e.g., 2.79ms vs. 7120ms on MNIST).
- **Generalizability**: The framework demonstrates robustness across diverse datasets, including X-ray security screening (UBA and FFOB datasets), indicating its applicability to medical imaging tasks like X-ray anomaly detection.

- **Reproducibility**: The publicly available code (github.com/samet-akcay/ganomaly) ensures transparency and facilitates further research.

Experimental Evaluation

Datasets GANomaly was evaluated on four datasets, two of which are relevant to X-ray imaging:

- **University Baggage Anomaly Dataset (UBA)**: Contains 230,275 X-ray image patches, with normal samples from sliding windows and abnormal samples (knives, guns, gun components) manually cropped. This dataset simulates X-ray scenarios where anomalies (e.g., foreign objects) must be detected.
- **Full Firearm vs. Operational Benign (FFOB)**: Includes 4,680 firearm images (abnormal) and 67,672 benign images (normal) from X-ray security screening, providing a real-world context for anomaly detection.

Additionally, MNIST and CIFAR10 were used to benchmark performance on simpler tasks.

Results Performance was measured using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). Key findings include:

- **MNIST**: GANomaly outperformed AnoGAN (AUC: 0.7120), EGBAD (AUC: 0.892), and variational autoencoders (VAE), achieving the highest AUC across all digit classes chosen as anomalies.
- **CIFAR10**: GANomaly achieved the best AUC for each class treated as anomalous, despite challenges from similar normal and abnormal classes (e.g., plane vs. bird).
- **UBA and FFOB**: GANomaly demonstrated superior performance in X-ray security screening, with AUC values of 0.266 and 0.253, respectively, compared to EGBAD (0.888, 0.887) and AnoGAN (0.7110, 0.7223). Qualitative results (Figure 7) showed the model's failure to reconstruct abnormal objects, confirming its ability to detect anomalies.
- **Computational Efficiency**: Table 2 highlights GANomaly's runtime advantage, with inference times of 2.21–2.79ms across datasets, compared to 7110–7223ms for AnoGAN and 8.71–8.92ms for EGBAD.

Analysis The results underscore GANomaly's ability to model normal data distributions effectively, with the encoder loss providing a robust anomaly score. The model's failure to reconstruct abnormal samples (e.g., guns in FFOB) validates its hypothesis that anomalies lead to dissimilar latent representations. The t-SNE visualization (Figure 6b) and histogram of scores (Figure 6a) further confirm the separation of normal and abnormal samples in latent space.

Limitations

Despite its strengths, GANomaly has limitations:

- **CNN Limitations**: The reliance on CNNs may limit the capture of long-range dependencies in x-rays, where diffuse patterns (e.g., in interstitial lung disease) are common.
- **Dataset Bias**: The model's performance depends on the quality and diversity of normal training data, which may not fully represent all healthy X-ray variations.
- **Threshold Sensitivity**: The anomaly score threshold (ϕ) requires careful tuning, which may vary across clinical contexts.

2.4.5 HealthyGAN[12]

In the ever-evolving landscape of medical image analysis, the automation of disease detection and diagnosis is increasingly reliant on machine learning and deep learning techniques. Despite these advances, one of the persistent challenges in building effective diagnostic models is the scarcity of large-scale annotated datasets. Obtaining annotations for medical images—particularly those associated with rare or complex diseases—requires expert radiological interpretation and is often prohibitively expensive, time-consuming, and sometimes even infeasible in resource-constrained clinical environments.

To circumvent these limitations, HealthyGAN[12] proposes a transformative approach to anomaly detection that does not rely on annotated datasets. Instead, it introduces a novel unsupervised image-to-image translation framework capable of learning disease representations from a mixture of healthy and diseased images, without requiring any explicit labels. In doing so, HealthyGAN opens up new possibilities for deploying AI-driven diagnostic tools in real-world clinical settings where only uncurated, unannotated medical image archives are available.

Unlike conventional models which operate under the assumption that only healthy

images are accessible during training and treat diseased samples as statistical outliers during testing, HealthyGAN is fundamentally different. It is specifically designed to learn from mixed datasets that mirror the diversity and complexity of real clinical data repositories. This approach not only enhances the generalizability of the model but also significantly broadens the range of application domains where anomaly detection can be reliably performed without manual labeling.

HealthyGAN is formulated as a one-directional unpaired image-to-image translation framework. The key insight behind this approach is to learn a mapping from any input image—regardless of its health condition—to a "healthy-like" version. This transformation allows the model to estimate what a healthy version of the input might look like. The difference between the input and the translated output is then used to detect and localize potential anomalies.

What sets HealthyGAN apart from other generative models like CycleGAN or DualGAN is its one-way translation design. While traditional unpaired translation methods enforce cycle-consistency by mapping both from normal to abnormal and back, this bi-directionality is impractical in clinical applications where abnormal images cannot be reliably reverse-mapped due to the vast diversity and unpredictability of disease presentation. HealthyGAN sidesteps this limitation by focusing solely on abnormal-to-healthy translation, thereby avoiding the need for paired data or bidirectional constraints.

Network Architecture

The HealthyGAN architecture is composed of two main components:

Generator The generator is the core engine responsible for transforming a given input image into a healthy-looking version. However, this transformation is not a naïve one. Instead of directly outputting the translated image, the generator first produces two intermediate outputs:

- An intermediate healthy image (B_{int})
- A spatial mask (M)

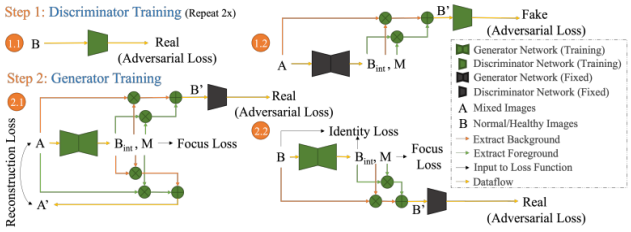


Figure 2.7: Pipeline of Healthygan

The spatial mask M, with values ranging from 0 to 1, acts as an attention map that determines which parts of the original image should be replaced with the generated healthy content. The final output image is computed by blending the original and generated images according to this mask:

$$B = B_{\text{int}} \cdot M + A \cdot (1 - M) \tag{2.6}$$

$$A = A \cdot M + B_{\text{int}} \cdot (1 - M) \tag{2.7}$$

For images from the healthy dataset, the mask ideally becomes empty, preserving the image. For diseased images, the mask highlights abnormal regions that need to be corrected. This masking mechanism enables partial modification of images, ensuring that the generator only alters regions associated with potential abnormalities while preserving the rest of the anatomical structure. This is crucial for maintaining biological plausibility and interpretability.

Discriminator The discriminator employs the PatchGAN design, which operates at the level of local image patches rather than full images. This allows it to focus on fine-grained spatial features and texture irregularities, helping it better distinguish between real and synthesized healthy images. The PatchGAN structure encourages the generator to produce images with high-frequency realism and structural fidelity.

Training Methodology

HealthyGAN adopts an adversarial training scheme similar to other GAN-based methods but with several nuanced differences tailored for the medical imaging context. The model is trained using two datasets:

- **Set A**: An unannotated mixed dataset containing both healthy and diseased samples.
- **Set B**: A known healthy dataset containing only verified healthy images.

Discriminator Training The discriminator is trained to:

- Classify real healthy images from Set A and Set B as real.
- Classify images generated from both Set A and Set B as fake.

The training uses a Wasserstein GAN (WGAN) loss with gradient penalty for improved training stability. This formulation addresses common GAN training issues such as mode collapse and oscillatory behavior.

$$\mathcal{L}_D^{\text{adv}} = \mathbb{E}_{x\in A}\left[D_{\text{real/fake}}(G(x))\right] - \mathbb{E}_{x\in B}\left[D_{\text{real/fake}}(x)\right] + \lambda_{\text{gp}}\,\mathbb{E}_{\hat{x}}\left[\left(\left\|\nabla_{\hat{x}} D_{\text{real/fake}}(\hat{x})\right\|_2 - 1\right)^2\right] \tag{2.8}$$

Generator Training The generator is trained using a composite loss function that integrates multiple objectives:

- **Adversarial Loss**: Encourages the generator to fool the discriminator by producing realistic healthy-like images.

$$\mathcal{L}_G^{\text{adv}} = \mathbb{E}_{x\in\{A,B\}}\left[D_{\text{real/fake}}(G(x))\right] \tag{2.9}$$

- **Identity Loss**: Ensures that when healthy images are passed through the generator, they are preserved, functioning effectively as an autoencoder.

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{x\in B}\left[\left\|G_{\text{int}}(x) - x\right\|_1\right] \tag{2.10}$$

- **Reconstruction Loss**: Promotes consistency between the original image and its output, particularly useful for retaining structure in healthy regions.

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x\in A, y\in A}\left[\left\|x - y\right\|_1\right] \tag{2.11}$$

- **Focus Loss**: Controls the size and sharpness of the spatial mask, ensuring it captures only relevant regions and converges toward binary values (0 or 1).

$$f = \lambda_{fs}\left(\frac{1}{n}\sum_{i=1}^{n} M_i\right)^2 + \lambda_{fz}\frac{1}{n}\sum_{i=1}^{n}\left|M_i - 0.5\right| \tag{2.12}$$

The final generator loss is a weighted sum of all the components, allowing precise tuning of the model's behavior based on the nature of the input data and desired sen-

sitivity to anomalies.

$$\mathcal{L}_D = \mathcal{L}_D^{\text{adv}} \tag{2.13}$$

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_f\mathcal{L}_f \tag{2.14}$$

Anomaly Detection Mechanism Once trained, the HealthyGAN model detects anomalies in a two-step process:

1. **Healthy Translation**: The input image (from either a healthy or diseased subject) is passed through the generator to produce its corresponding healthy-like version.
2. **Difference Map Computation**: The absolute voxel-wise difference between the input and the output is computed. This map highlights regions where the generator made significant changes—interpreted as potential abnormalities.

To quantify the presence of disease, an anomaly score is calculated by averaging the pixel values in the difference map. A higher score indicates a greater likelihood of pathology. This scalar score can be used for binary classification or ranked comparisons across patients.

Moreover, the difference map itself serves as a visual explanation, allowing clinicians to inspect which areas were flagged as anomalous. Although it does not provide precise segmentation, it offers valuable insight into localized structural deviations.

Datasets

HealthyGAN was trained and evaluated using the following three datasets:

X-ray14 Dataset

- **Modality**: X-ray (CXR), grayscale images.
- **Composition**:
  - *Known Healthy Set*: 10,000 healthyx-rays.
  - *Mixed Unannotated Set*: 5,000 healthyx-rays and 3,195 diseasedx-rays (with single disease labels, excluding multi-disease cases).
- **Task**: Detection of 14 thoracic diseases (e.g., pneumonia, pneumothorax, cardiomegaly).
- **Source**: Publicly available through the National Institutes of Health (NIH) **Wang2017**.

COVID-19 Dataset

- **Modality**: X-ray (CXR), grayscale images.
- **Composition**:
  - *Known Healthy Set*: Verified healthyx-rays (exact number not specified, likely from pre-COVID or non-COVID patients).
  - *Mixed Unannotated Set*: Includes both healthy and COVID-19-positivex-rays.
- **Task**: Detection of COVID-19-related abnormalities (e.g., ground-glass opacities, consolidation).
- **Source**: Likely derived from public repositories like the National COVID-19 Imaging Database (NCCID) or similar collections **Encord2023**.

IXI Dataset

- **Modality**: Brain Magnetic Resonance Imaging (MRI), typically T1-weighted or T2-weighted images.
- **Composition**:
  - *Known Healthy Set*: 424 healthy participants.
  - *Mixed Unannotated Set*: Includes both healthy and migraine-affected MRIs (exact number not specified).
- **Task**: Detection of migraine-related anomalies (subtle structural or textural changes).
- **Source**: Publicly available through Imperial College London **IXI**.

Key Contributions and Innovations

HealthyGAN's primary innovations can be summarized as follows:

- **Unsupervised Learning from Mixed Datasets**: It is one of the first GAN-based models designed to explicitly leverage unannotated, mixed datasets that resemble real-world clinical data pools.
- **One-Directional Translation**: Unlike cycle-consistent models, HealthyGAN eliminates the need for unrealistic backward mappings, simplifying the architecture and making it more practical.

- **Mask-Guided Reconstruction**: The introduction of attention masks allows for targeted anomaly correction and enhances the interpretability of model outputs.
- **Generalization Across Modalities**: HealthyGAN demonstrates consistent performance across both 2D x-rays and 3D brain MRI slices, showing versatility and adaptability.

Limitations and Future Directions

Despite its many strengths, HealthyGAN is not without limitations. The primary challenge lies in the imprecise localization of abnormalities. The difference maps, while informative, are coarse and do not provide pixel-level segmentation of disease regions. Moreover, the use of Fréchet Inception Distance (FID) for model selection, though effective in image quality assessment, may not align perfectly with anomaly detection performance.

2.4.6 Brainomaly[10]

Brainomaly is a cutting-edge, unsupervised method specifically developed to identify neurologic abnormalities in T1-weighted brain Magnetic Resonance Imaging (MRI) scans without requiring any manually annotated datasets. Introduced by Siddiquee et al. and presented at the 2024 Winter Conference on Applications of Computer Vision (WACV), Brainomaly addresses a critical gap in medical imaging: the need for reliable, annotation-free anomaly detection in clinical brain scans. By leveraging a novel Generative Adversarial Network (GAN)-based image-to-image translation framework, the method aims to detect a variety of neurologic conditions, including but not limited to Alzheimer's disease and different forms of headache disorders.

One of the key innovations of Brainomaly lies in its ability to operate on unannotated datasets that consist of both healthy and diseased MRI scans. Traditional anomaly detection models typically depend on clean, healthy-only datasets during training, which are difficult to obtain and do not always represent real-world conditions. Brainomaly, in contrast, is explicitly designed to work with mixed datasets—those containing both healthy and pathological scans without any explicit labeling—thereby making it highly applicable to real clinical environments where data annotation is costly, time-consuming, and often infeasible.

The model achieves superior performance through a combination of several innovative components: a novel additive map-based image translation approach, the inte-

gration of unannotated mixed datasets during training. These elements collectively enable Brainomaly to outperform existing unsupervised anomaly detection methods across multiple benchmark datasets.

Methodology

Brainomaly's operational pipeline is structured into three core stages: training, testing, and disease detection. The method is particularly optimized for use with registered T1-weighted brain MRIs, which are spatially aligned to a standardized anatomical template such as MNI152. This alignment facilitates precise voxel-level comparisons and allows the model to leverage spatial consistency across subjects.

Network Architecture

The Brainomaly framework is designed around a standard Generative Adversarial Network (GAN) architecture, comprising two primary components: a generator and a discriminator. Both networks operate on 2D slices extracted from volumetric T1-weighted brain MRIs. The final output for each subject is generated by stacking the processed 2D slices in the same order as they appear in the original 3D MRI volume.

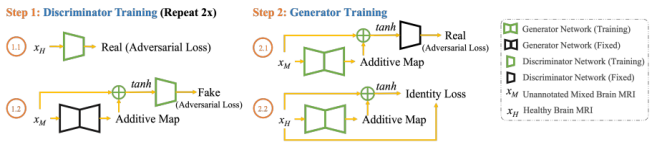


Figure 2.8: Pipeline of Brainomaly

Generator The generator follows an encoder-decoder architecture, similar in design to those used in prior generative anomaly detection models. The network accepts a single 2D T1-weighted brain MRI slice as input, agnostic to whether the subject is healthy or diseased. It generates an additive map, a 2D image where each voxel represents the estimated voxel-wise adjustment required to transform the input brain image into a healthy one.

The final reconstructed healthy MRI slice is obtained by performing a voxel-wise summation between the input MRI slice and the generated additive map, followed by a tanh activation function. This additive approach leverages the spatial alignment of registered MRIs, avoiding the need for strict cycle-consistency constraints commonly employed in other image-to-image translation models.

Discriminator The discriminator adopts the PatchGAN architecture, a model which classifies whether local patches within an image are real or generated, rather than assessing the image as a whole. This allows for more fine-grained discrimination and helps the generator produce realistic details at a local level.

The discriminator is trained to distinguish between real healthy brain MRIs and those synthesized by the generator. It takes as input either a true healthy MRI slice from the dataset or a generated MRI slice produced by the generator and returns a real/fake prediction for each patch in the input.

Training Procedure

The generator and discriminator are trained in an adversarial fashion, as is standard in GAN-based frameworks. The training alternates between optimizing the discriminator and the generator, where the discriminator is updated twice for every single update to the generator. This update schedule helps maintain the balance between the two networks and stabilizes the adversarial training process.

Discriminator Training During training, the discriminator receives real healthy brain MRI slices from a known healthy dataset (denoted as H) and generated slices created from an unannotated mixed dataset (M) using the generator. The discriminator is trained to maximize the distinction between these two types of inputs by minimizing the adversarial loss.

The original adversarial loss, which uses a binary cross-entropy formulation, is replaced by a Wasserstein GAN loss with gradient penalty, known for improved stability and convergence properties in GAN training. The loss function for the discriminator is:

$$\mathcal{L}_D^{\text{adv}} = \mathbb{E}_{x_M \sim M}\left[D_{\text{real/fake}}\left(\tanh(x_M + G(x_M))\right)\right] - \mathbb{E}_{x_H \sim H}\left[D_{\text{real/fake}}(x_H)\right] + \lambda_{\text{gp}}\, \mathbb{E}_{\hat{x}}\left[\left(\left\|\nabla_{\hat{x}} D_{\text{real/fake}}(\hat{x})\right\|_2 - 1\right)^2\right] \tag{2.15}$$

Here, $\hat{x}$ is a random linear interpolation between a real and a generated image, and λ_{gp} is the weight of the gradient penalty term.

Generator Training The generator's objective is twofold: to produce realistic healthy MRIs that can fool the discriminator and to preserve healthy MRIs when they are used as inputs. For this purpose, two distinct loss functions are employed:

Adversarial Loss: Encourages the generator to produce outputs that the discriminator cannot distinguish from real healthy MRIs. Applied to unannotated MRIs from

dataset M, this loss is defined as:

$$\mathcal{L}_G^{\text{adv}} = \mathbb{E}_{x_M \sim M}\left[D_{\text{real/fake}}\left(\tanh(x_M + G(x_M))\right)\right] \tag{2.16}$$

Identity Loss: Ensures that when the generator receives a truly healthy input (from dataset H), it behaves like an autoencoder and leaves the input unchanged. This regularization is essential for preserving anatomical fidelity in cases where no pathological transformation is needed. It is defined as:

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{x_H \sim H}\left[\left\|\tanh(x_H + G(x_H)) - x_H\right\|_1\right] \tag{2.17}$$

The final objective function for the generator combines the adversarial and identity losses, weighted by a hyperparameter λ_{id}, which controls the relative importance of maintaining identity reconstruction:

$$\mathcal{L}_D = \mathcal{L}_D^{\text{adv}} \tag{2.18}$$

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}} \tag{2.19}$$

This architecture and training strategy enable Brainomaly to effectively generate plausible healthy reconstructions from both normal and potentially abnormal MRIs. The difference between the original and generated images is later used to derive anomaly maps and scores for patient-level disease detection.

Disease Detection In the testing phase, Brainomaly receives an input MRI and uses the trained generator to produce a corresponding healthy version. The voxel-wise difference between the input image and the generated healthy image forms a "difference map," which visually and quantitatively highlights regions exhibiting structural anomalies. For healthy subjects, the difference map is expected to be minimal, whereas for diseased subjects, the map reveals distinct patterns of alteration corresponding to the disease pathology.

The model quantifies the anomaly using a disease detection score, calculated as the average activation in the difference map. Higher scores indicate a higher likelihood of neurologic abnormality. This approach facilitates not only slice-level but also subject-level disease detection, providing an interpretable and scalable solution for clinical use.

Evaluation on Datasets

Brainomaly was rigorously evaluated on two large-scale MRI datasets:

Alzheimer's Disease Dataset: Sourced from the ADNI initiative, this dataset includes 536 patients diagnosed with Alzheimer's disease and 1271 healthy controls. Brainomaly was tested using both clearly labeled healthy images and mixed subsets containing both healthy and diseased samples.

Headache Dataset: This dataset was curated from the Mayo Clinic and IXI repositories and includes several types of headache disorders—migraine, acute post-traumatic headache (APTH), and persistent post-traumatic headache (PPTH)—alongside a large number of healthy controls. Mixed sets were carefully constructed to test Brainomaly's ability to detect diverse and often subtle pathology.

In both settings, all images were spatially normalized to the MNI152 template, skull-stripped, and converted to 2D sagittal slices. Brainomaly demonstrated outstanding performance, achieving an average AUC of 0.6550 for Alzheimer's disease and 0.8960 for headache detection. In particular, it excelled in detecting sub-types such as migraine (precision 0.9375) and PPTH (precision 0.9600), though performance was somewhat reduced for APTH due to its heterogeneous nature.

Additional experiments in both transductive (test data seen during training) and inductive (completely unseen test data) settings showed that Brainomaly generalizes well, reinforcing its practical value in real-world clinical deployments.

Significance, Limitations, and Future Extensions

Brainomaly's primary contributions to the field of medical imaging include its capacity to function without annotated data, its unique use of additive maps for transformation, and the innovative AUCp-based model selection framework. These contributions make the method highly flexible, cost-effective, and suitable for a variety of clinical applications where labeled data are scarce.

Nonetheless, there are certain limitations. While the difference maps generated by Brainomaly provide valuable insights at the patient level, they do not precisely localize disease-specific abnormalities. The structural changes highlighted may not always correlate with the actual pathological regions. Additionally, Brainomaly's performance on structurally diverse conditions like APTH suggests a need for more robust handling of heterogeneous pathologies.

2.4.7 SAGAN[15]

SAGAN[15] is a novel framework designed to tackle semi-supervised anomaly detection in medical imaging, a field where identifying abnormalities (e.g., lung lesions, glaucoma) is critical but often hampered by limited labeled data. Unlike traditional methods, which either require extensive labeled datasets (fully supervised) or focus solely on modeling normal data (unsupervised), SAGAN[15] innovatively leverages both labeled and unlabeled data, including images with potential anomalies. By integrating spatial-aware encoding, attention mechanisms, and a generative adversarial network (GAN), SAGAN[15] achieves superior anomaly detection and localization, particularly in datasets like x-rays (VinDr-CXR, RSNA) and retinal scans (LAG).

This explanation dives into SAGAN[15]'s core innovations, training strategy, advantages, architecture, performance, limitations, and future directions, with an emphasis on clarity and depth.

Network Architecture

The Spatial-aware Attention Generative Adversarial Network (SAGAN**zhang2024Fspatialawareat**
is designed for semi-supervised anomaly detection in medical imaging, leveraging a sophisticated architecture that integrates spatial-aware position encoding and attention-based restoration. The network comprises two primary components: a generator and a discriminator, each tailored to address the challenges of handling mixed (normal and abnormal) unlabeled data while ensuring anatomical consistency and precise anomaly restoration.

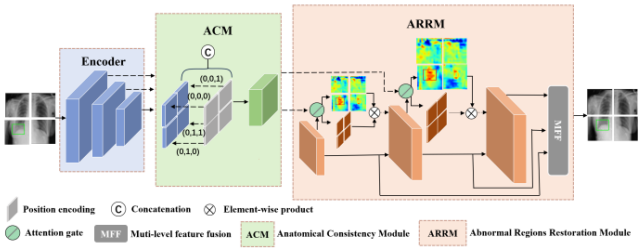


Figure 2.9: Pipeline of SAGAN[15]

Generator The generator is the core component responsible for transforming input medical images—whether normal, pseudo-anomalous, or unlabeled—into their healthy-like counterparts. It incorporates two innovative mechanisms: spatial-aware position encoding and attention-based restoration, which enable precise anomaly detection and localization.

Spatial-aware Position Encoding To exploit the consistent anatomical structures in medical images (e.g., the predictable location of lungs in x-rays), SAGAN[15] embeds spatial information into its feature representations. This ensures the model understands the spatial context of each region within the image.

- **Patch Division**: The input image is segmented into a grid of $N \times N$ non-overlapping patches (e.g., a 2×2 grid yielding four patches). Each patch represents a distinct region, such as the upper-left quadrant of a X-ray.
- **Positional Codes**: Each patch is assigned a binary positional code based on its grid location. For example, in a 2×2 grid, the top-left patch is encoded as $[0, 0]$, and the bottom-right patch as $[1, 1]$. These codes are compact and computationally efficient.
- **Feature Encoding**: The positional codes are integrated into the feature vectors processed by the network, resulting in a feature representation with the shape:

$$f_{pe} \in \mathbb{R}^{B \times \left(C + \lceil \log_2(N \times N) + 1 \rceil\right) \times H \times W} \tag{2.20}$$

  where:

  - B: Batch size (number of images processed simultaneously).
  - C: Number of input channels (e.g., 1 for grayscalex-rays, 3 for RGB retinal images).
  - $\lceil \log_2(N \times N) + 1 \rceil$: Number of bits required to encode patch positions, ensuring unique identification.
  - $H \times W$: Spatial dimensions of the feature map (height and width).

This encoding mechanism enforces anatomical consistency by ensuring that similar structures (e.g., the left lung) are processed similarly across different images, even in the presence of abnormalities. It is critical for distinguishing normal anatomical features from anomalies.

Attention-based Restoration To address the localized nature of anomalies (e.g., tumors or lesions) in medical images, SAGAN[15] employs attention mechanisms within the decoder to focus dynamically on abnormal regions while preserving normal structures.

- **Attention Gates**: At each decoder level, attention gates compute pixel-wise

attention coefficients (α) to identify regions requiring modification:

$$\alpha = S\left(C_3\left(R\left(C_1(f_l) + C_2(g_l)\right)\right)\right) \tag{2.21}$$

where:

- f_l: Features from the encoder, representing the input image.
- g_l: Features from skip connections, carrying information from earlier layers.
- C_1, C_2, C_3: Convolutional layers that process the features.
- R: ReLU activation function, introducing non-linearity.
- S: Sigmoid activation function, scaling α to $[0, 1]$.

- **Skip Connection Modulation**: The attention coefficients modulate the skip connections to prioritize abnormal regions:

$$g_l = \alpha \times g_l \tag{2.22}$$

This ensures that normal regions (where $\alpha \approx 0$) are preserved, while abnormal regions (where $\alpha \approx 1$) are targeted for restoration.

The attention-based restoration mechanism enables SAGAN[15] to selectively restore anomalies (e.g., converting a lung lesion to normal tissue) without altering unaffected areas, enhancing the precision of anomaly detection, especially for subtle abnormalities.

Discriminator The discriminator complements the generator by evaluating the realism of restored images. It adopts a PatchGAN architecture, which processes local image patches (e.g., 70×70 pixels) rather than the entire image. This focus on high-frequency details, such as textures and edges, ensures that restored images are realistic at a local level, which is critical for medical imaging applications where fine details (e.g., lesion boundaries) are significant.

Training Method

SAGAN[15]'s training method optimizes the generator and discriminator to generate realistic healthy images from a mix of normal images x_n, pseudo-anomalous images x_p, and unlabeled images x_u. The training leverages a three-pronged loss framework,

incorporating identity, reconstruction, and adversarial losses, with a gradient penalty to stabilize the discriminator.

Generator Training The generator is trained to produce healthy images x'_n, x'_p, and x'_u from normal, pseudo-anomalous, and unlabeled inputs, respectively, ensuring that x'_u and x'_p are close to realistic healthy images. The training objective is achieved through the following loss functions:

- **Identity Loss**: Ensures that normal images are preserved unchanged:

$$L_{id}^G = \mathbb{E}_{x_1 \in x_n}[\|x'_1 - x_1\|_1], \tag{2.23}$$

  where x_1 is a normal image and x'_1 is its generated output.

- **Reconstruction Loss**: Ensures that pseudo-anomalous images are restored to their corresponding normal state:

$$L_{rec}^G = \mathbb{E}_{x_2 \in x_p, x_1 \in x_n}[\|x'_2 - x_1\|_2], \tag{2.24}$$

  where x_2 is a pseudo-anomalous image generated from x_1, and x'_2 is its restored output.

- **Adversarial Loss**: Encourages the generator to produce realistic healthy images from unlabeled data:

$$L_{adv}^G = -\mathbb{E}_{x_3 \in x_u}[D(x'_3)], \tag{2.25}$$

  where D is the discriminator function and x'_3 is the generated image from an unlabeled input x_3.

The total generator loss combines these components with weighting factors:

$$L_G = L_{adv}^G + \lambda_{id} L_{id}^G + \lambda_{rec} L_{rec}^G, \tag{2.26}$$

where λ_{id} and λ_{rec} are the weights for the identity and reconstruction losses, respectively.

The generator's training is enhanced by the ACM and ARRM:

- The ACM embeds positional codes to maintain anatomical consistency, ensuring that features from corresponding regions (e.g., the left lung) are aligned across images.

- The ARRM uses attention gates to focus restoration efforts on abnormal regions, with modulated skip connections preserving normal structures.

Discriminator Training The discriminator is trained to classify normal images x_n as real and generated healthy images x'_u (from unlabeled images x_u) as fake, using an adversarial loss with a gradient penalty for stability:

$$L^D_{adv} = \mathbb{E}_{x_3 \in x_u}[D(x'_3)] - \mathbb{E}_{x_1 \in x_n}[D(x_1)] + \lambda_{gp}\mathbb{E}_{\hat{x}}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right], \tag{2.27}$$

where $\hat{x}$ is a random weighted average of normal and generated healthy images, and λ_{gp} is the gradient penalty weight.

The discriminator's PatchGAN architecture focuses on local patch realism, refining the generator's ability to produce high-fidelity restored images. The gradient penalty mitigates overfitting and stabilizes training, ensuring robust convergence.

Training Dynamics SAGAN[15] alternates between generator and discriminator updates, leveraging the ACM and ARRM to handle mixed datasets effectively. The generator's spatially-aware and attention-guided mechanisms enable precise anomaly restoration, while the discriminator ensures local realism. This training method allows SAGAN[15] to achieve superior anomaly detection and localization in medical imaging tasks.

Datasets

The Spatial-aware Attention Generative Adversarial Network (SAGAN[15]) was evaluated on three medical imaging datasets to assess its performance in semi-supervised anomaly detection: VinDr-CXR, RSNA, and LAG. These datasets, covering x-rays and retinal scans, were used to train and test SAGAN[15]'s ability to detect anomalies such as lung lesions, pneumonia, and glaucoma, leveraging labeled normal images, pseudo-anomalous images, and unlabeled mixed images. Below, we describe each dataset, including its modality, composition, task, performance metrics, and source, following the order presented in the performance evaluation **SAGAN\cite {zhang2024spatialawar**

VinDr-CXR Dataset. The VinDr-CXR dataset is a large-scale collection of x-rays focused on detecting thoracic abnormalities, specifically lung lesions in SAGAN[15]'s evaluation.

- *Modality*: X-ray (CXR), grayscale images.
- *Composition*:

  - Labeled Normal Set: Approximately 3,000–4,000 healthyx-rays, used to train the generator to preserve normal anatomy and the discriminator to recognize real healthy images.
  - Pseudo-anomalous Set: Synthetically generated anomalous images derived from normalx-rays, enabling restoration training.
  - Unlabeled Set: A mix of healthy and diseasedx-rays containing lung lesions, simulating real-world clinical data.
- *Task*: Detection of lung lesions (e.g., nodules, masses) via difference-based heatmaps.
- *Performance*: SAGAN[15] achieved an Average Precision (AP) of 89.30% and an Area Under the Curve (AUC) of 91.47%, outperforming Brainomaly (AP 84.37%, AUC 88.54%) **SAGAN\cite {zhang2024spatialawareattentiongenerativeadversarial}202**
- *Source*: Publicly available through the VinDr Lab platform, developed by the Vingroup Big Data Institute **VinDr2021**.

RSNA Dataset. The RSNA dataset, likely from the RSNA Pneumonia Detection Challenge, consists of x-rays targeting pneumonia-related abnormalities.

- *Modality*: X-ray (CXR), grayscale images.
- *Composition*:
  - Labeled Normal Set: Approximately 3,000–4,000 healthyx-rays, supporting generator and discriminator training.
  - Pseudo-anomalous Set: Synthetically created images with pneumonia-like anomalies, used for restoration training.
  - Unlabeled Set: A mix of healthy and pneumonia-affectedx-rays, reflecting clinical variability.
- *Task*: Detection of pneumonia (e.g., opacities, consolidation) through difference-based heatmaps.
- *Performance*: SAGAN[15] achieved an AP of 92.60% and an AUC of 92.45%, surpassing Brainomaly (AP 87.93%, AUC 91.01%). It maintained 92.6% AP with 0% anomalies in the unlabeled set, demonstrating robustness **SAGAN\cite {zhang2024spatialav**
- *Source*: Publicly available through the Radiological Society of North America (RSNA) Kaggle Challenge **RSNA2018**.

LAG Dataset. The LAG dataset comprises retinal scans for detecting glaucoma, a condition characterized by subtle structural changes in the optic nerve.

- *Modality*: Retinal scans (fundus photography or optical coherence tomography), typically RGB or grayscale images.
- *Composition*:
  - Labeled Normal Set: Approximately 3,000–4,000 healthy retinal scans, used for baseline training.
  - Pseudo-anomalous Set: Synthetically generated images with glaucoma-like features, supporting restoration training.
  - Unlabeled Set: A mix of healthy and glaucoma-affected retinal scans, simulating clinical archives.
- *Task*: Detection of glaucoma-related anomalies (e.g., optic disc cupping, nerve fiber layer defects) via difference-based heatmaps.
- *Performance*: SAGAN[15] achieved an AP of 96.96% and an AUC of 96.98%, outperforming Brainomaly (AP 94.71%, AUC 95.92%).
- *Source*: Likely derived from the Large-scale Attention-based Glaucoma (LAG) dataset or similar retinal imaging datasets **LAG2020**.

These datasets were critical for training SAGAN[15]'s generator and discriminator, with labeled normal sets establishing baselines, pseudo-anomalous sets enabling restoration, and unlabeled sets testing real-world applicability. Their diverse modalities and clinical relevance underscore SAGAN[15]'s versatility in anomaly detection

Key Advantages Over Prior Work

SAGAN[15] offers several improvements over existing methods like CycleGAN, Brainomaly, DDAD, and AMAE, making it particularly suited for medical anomaly detection.

- **Efficient Use of Unlabeled Data**:
  - Unlike CycleGAN, which requires cyclic consistency (forcing images to be reversible), SAGAN[15] relaxes this constraint, allowing flexible handling of unpaired normal and abnormal images.
  - Achieves 89.3% Average Precision (AP) on VinDr-CXR, compared to 84.37% for Brainomaly, demonstrating superior performance with mixed data.

- **Anomaly Localization**:
  - SAGAN[15] generates heatmaps by computing the difference between the original and restored images, visually highlighting anomalies (e.g., lesions in lungx-rays).
  - Outperforms DDAD and AMAE by 4–5% in AUC across datasets, thanks to its attention-based restoration.
- **Robustness to Anomaly Ratios**:
  - Performs well even when unlabeled data contains 0% anomalies (simulated), achieving 92.6% AP on the RSNA dataset. This robustness is critical for real-world scenarios where anomaly prevalence varies.

Limitations and Future Directions

While SAGAN[15] is a significant advancement, it has some limitations that pave the way for future improvements.

Limitations

- **Pixel-level Localization**:
  - SAGAN[15] struggles to delineate precise boundaries for small or subtle anomalies (e.g., tiny lesions), as its attention mechanism operates at a coarser level.
- **Computational Cost**:
  - The dual-path generator (handling both normal and anomalous regions) increases training time by 25% compared to Brainomaly, posing challenges for large-scale deployment.

Future Directions

- **Transformer Integration**:
  - Incorporate transformer-based architectures to model long-range dependencies in images, potentially improving localization of subtle anomalies.
- **Efficiency Improvements**:
  - Optimize the generator to reduce computational overhead, enabling faster training and deployment on resource-limited devices.

Chapter 3

Proposed Methodology

Medical anomaly detection from x-rays is essential for identifying abnormalities associated with diseases such as pneumonia, tuberculosis, and lung cancer. Existing unsupervised approaches, particularly the Spatial-aware Attention GAN (*SAGAN* [15]), have demonstrated promising results by incorporating spatial awareness to preserve anatomical consistency during reconstruction. SAGAN introduces two key modules: the Anatomical Consistency Module (ACM), which enforces consistency of normal anatomical regions, and the Abnormal Region Restoration Module (ARRM), which reconstructs anomalous regions by leveraging learned normal patterns. This architecture allows SAGAN to focus on clinically relevant areas while mitigating the effect of noise or irrelevant features.

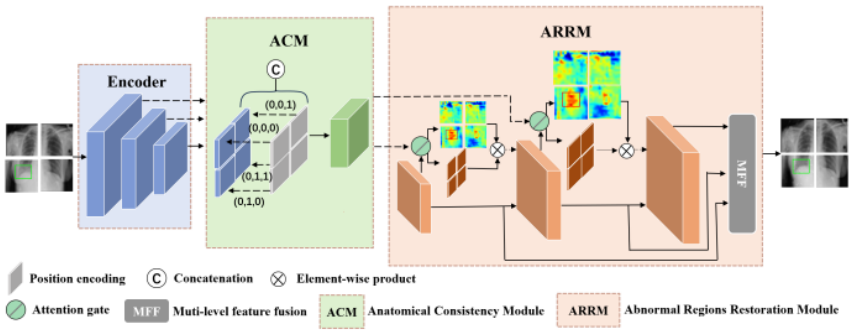
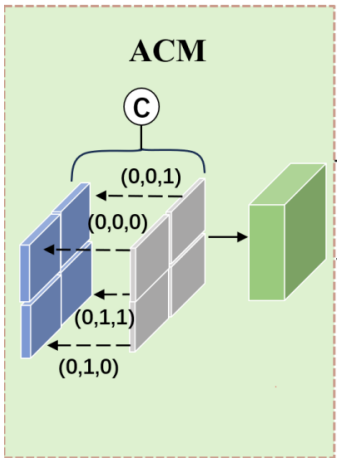


Figure 3.1: Overview of the SAGAN architecture showing the generator, discriminator, Anatomical Consistency Module (ACM), and Abnormal Region Restoration Module (ARRM).

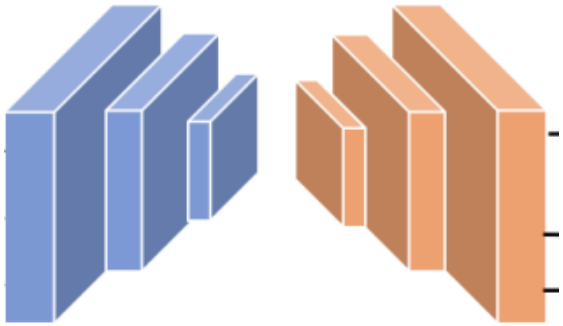
Despite these advancements, SAGAN has several limitations: its binary positional encoding in ACM may not fully capture fine-grained spatial relationships, and the

reliance on convolutional networks constrains the modeling of long-range dependencies. These issues reduce sensitivity to subtle or distributed anomalies, which are critical in early-stage disease detection.

To overcome these limitations, this research proposes a spatially aware anomaly detection framework with two primary enhancements: (1) improved spatial representations that capture both positional and structural relationships more effectively, and (2) a Swin Transformer-based encoder–decoder that enables multi-scale feature extraction and global context modeling. By integrating these modifications, the proposed framework aims to enhance anomaly detection accuracy while maintaining interpretability, and the subsequent sections provide a detailed, step-by-step explanation of each component and its rationale.



(a) Where enhanced positional encoding will be incorporated.



(b) Where Swin Transformer-based encoder–decoder will be incorporated.

Figure 3.2: Proposed enhancements to the SAGAN architecture.

3.1 Enhanced Positional Encodings

The SAGAN[15] model employs binary positional encoding, dividing the input image into $N \times N$ non-overlapping patches, each assigned a binary code based on its position (e.g., for $N = 2$, the first patch is encoded as $[0, 0, 0]$). While this approach provides basic spatial awareness, it may oversimplify the intricate anatomical structures inx-rays, potentially limiting the model's ability to preserve spatial relationships critical for accurate anomaly detection. To address this, we propose advanced positional encodings, categorized into absolute and relative encodings, to enhance SAGAN[15]'s performance.

1. Sinusoidal Positional Encoding[13]

Description: Sinusoidal positional encoding, introduced in the Transformer model [13], represents spatial positions using sine and cosine functions at multiple frequencies. For a 2D image, each pixel at position (x, y) is mapped to a vector of sine and cosine values, with the frequency varying across embedding dimensions:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right), \tag{3.1}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{3.2}$$

where pos is the position, i is the dimension index, and d is the embedding dimension. For images, separate encodings are computed along the x and y axes.

Incorporation: In this work, sinusoidal encodings were incorporated in place of the original binary positional encodings. They were added to input embeddings and concatenated with intermediate feature maps in both the generator and discriminator, with the expectation that they would provide more informative spatial cues for X-ray anomaly detection.

Predicted Impact: By replacing the original binary positional encoding with sinusoidal encodings, we expected to provide the model with a continuous and smooth representation of spatial positions, offering finer-grained awareness across anatomical variations and image resolutions. This encoding was predicted to capture both local anatomical details and broader thoracic structures, potentially improving the detection of subtle or distributed anomalies. Additionally, it was anticipated to help the model maintain spatial consistency, making deviations such as lesions or fractures easier to identify.

2. Learned Positional Embeddings

Description: Learned positional embeddings treat positional information as trainable parameters. Instead of using fixed functions (like sinusoids), the model learns a unique embedding vector for each position (x, y) in the image during training. These embeddings are typically initialized randomly and optimized via backpropagation to capture task-specific spatial relationships.

Incorporation: We replaced the binary positional encoding with a grid of learnable embeddings. Each patch or window in the input image is assigned a trainable vector stored in an embedding layer. These vectors are broadcasted across the spatial dimensions and concatenated with the feature maps of both the generator and discrim-

inator. This approach allows the network to learn spatial representations optimized for anomaly detection and supports multi-scale feature integration in a transformer-based or attention-enhanced architecture.

Predicted Impact: By replacing the static binary positional encoding in SAGAN with learnable embeddings, we predicted that the model would gain a more flexible, data-driven representation of spatial information tailored to x-rays. We expected these embeddings to adapt to dataset-specific spatial patterns, capturing key anatomical structures and regions prone to anomalies, such as the lungs and heart, more effectively than fixed binary codes. Unlike static encodings, learnable embeddings can model non-uniform spatial relationships, accommodating variability in patient anatomy and X-ray acquisition. Additionally, by emphasizing regions likely to contain abnormalities, the network is predicted to better localize and reconstruct anomalous areas. The approach also supports scalability, allowing embeddings to be applied to high-resolution images without relying on predefined patterns, enabling detailed analysis across varying image sizes. Overall, we hypothesized that these embeddings would enhance the generator's ability to reconstruct anatomically consistent images with improved fidelity and enable the discriminator to better distinguish real from reconstructed images based on meaningful spatial patterns, addressing the limitations of the original binary encoding.

3.1.1 Relative Positional Encodings

1. Rotary Positional Encoding (RoPE)[11]

Description: Rotary Positional Encoding (RoPE), proposed by Su et al. (2021), introduces positional information by applying rotation matrices to the query and key vectors within the attention mechanism. This technique enables the model to represent relative positional relationships in a continuous and geometrically meaningful manner. In this design, both horizontal and vertical spatial coordinates are encoded, allowing the attention mechanism to exploit structured spatial context.

Incorporation: In the updated generator architecture, RoPE is integrated into a newly introduced `MultiHeadSelfAttention` (MHSA) module placed in the bottleneck stage. The query and key tensors are reshaped and passed through the `RotaryPositionalEncoding` class, which applies position-dependent rotations derived from each spatial location's (x, y) coordinates. These rotated tensors are then used in the scaled dot-product attention calculation, embedding spatial relationships directly in the attention computation. This modification replaces the previous binary memory-conditioning mech-

anism, simplifying the forward path while enhancing the model's ability to capture global context.

Predicted Impact: The integration of RoPE was expected to enhance the model's spatial reasoning capabilities by encoding relative positional information in a continuous and geometrically meaningful way. This improvement would allow the network to better capture spatial relationships between anatomical structures, such as the relative positioning of the lungs and ribs, while also strengthening its ability to attend to clinically relevant regions, including subtle anomalous areas. Furthermore, the continuous nature of RoPE would make the model more robust to variations in image resolution and patient anatomy, potentially improving generalization across diverse X-ray datasets. By embedding positional information directly into the attention mechanism, RoPE would help overcome the limitations of binary encodings, which could not represent relational positions.

2. ALiBi (Attention with Linear Biases)[5]

Description: ALiBi (Attention with Linear Biases) is a relative positional encoding method that applies a non-learned linear bias to attention scores based on the relative spatial distance between positions in a 2D image. Instead of using embeddings, it penalizes distant positions with a negative bias, encouraging the attention mechanism to prioritize nearby pixels or patches. The bias is typically proportional to the Manhattan distance, making it simple and computationally efficient.

Incorporation: In the updated generator, ALiBi is integrated via a Multi-Head Self-Attention (MHSA) layer in the bottleneck of the network. Specifically, the encoder output is fed directly into the MHSA layer without reshaping or appending binary positional encodings. ALiBi computes head-specific linear slopes and applies Manhattan-distance-based biases directly to the attention score matrix before the softmax, modifying the attention mechanism as:

$$\text{Attention Score}_{i,j} = \frac{q_i \cdot k_j}{\sqrt{d_k}} - m_h \cdot d_{i,j} \tag{3.3}$$

where $d_{i,j} = |x_j - x_i| + |y_j - y_i|$ is the Manhattan distance between spatial positions, and m_h is the slope specific to each attention head. This integration preserves all existing skip connections and attention gates in the decoder while removing the previous binary memory-conditioning, simplifying the forward path and enhancing positional awareness in the attention mechanism.

Predicted Impact: We predicted that integrating Alibi into the model would potentially improve its handling of spatial information in x-rays. By capturing relative spatial relationships, the model would be able to maintain better anatomical consistency, while its invariance to absolute position shifts might enhance robustness to variations in patient positioning or X-ray alignment. Alibi's emphasis on local regions was expected to help attention mechanisms, such as those in SAGAN or Swin Transformer-based generators, focus on areas critical for anomaly detection, potentially improving precision in identifying lesions or other deviations from typical anatomical layouts. Furthermore, since Alibi introduces no additional learnable parameters or lookup tables, it was anticipated to offer these benefits with minimal computational overhead, making it a lightweight and scalable approach for high-resolution image generation and anomaly detection.

3. Attention Bias (Relative Position Bias)[3]

Description: Attention Bias Encoding, also known as Relative Position Bias, adds a learnable scalar bias to the attention scores based on the relative distance between positions in a 2D image. Used in models like Transformer-XL, it adjusts the attention mechanism to prioritize certain spatial relationships, capturing relative positional dependencies without modifying token embeddings.

Incorporation: In the generator, attention bias is integrated into a multi-head self-attention layer placed approximately at the midpoint of the bottleneck residual blocks. For positions (x_i, y_i) and (x_j, y_j), compute the relative distance $(\Delta x, \Delta y) = (x_j - x_i, y_j - y_i)$ and retrieve a learnable scalar bias $b_{\Delta x, \Delta y}$ from a precomputed lookup table. This bias is then added directly to the attention scores:

$$\text{Attention Score}_{i,j} = \frac{q_i \cdot k_j}{\sqrt{d_k}} + b_{\Delta x, \Delta y} \tag{3.4}$$

The output of this attention layer is followed by normalization and optional dropout for stability. This approach replaces the previous memory-conditioned positional encoding and allows the network to capture long-range spatial dependencies more effectively without reshaping or concatenating binary positional codes.

Predicted Impact: Integrating attention bias into the generator was expected to improve the model's ability to capture contextual spatial relationships, such as the relative distances between anatomical structures like ribs and spine, thereby potentially enhancing anatomical consistency in generated images. By focusing on relative positions rather than absolute coordinates, the model would likely become more robust

to shifts in X-ray alignment or variations in patient positioning. Additionally, attention bias could refine the focus of attention mechanisms in models like SAGAN[15] or Swin Transformer, helping the network emphasize relevant spatial regions and detect anomalies that disrupt expected anatomical patterns, such as fractures in unusual locations. Compared to previous memory-conditioned or binary positional encodings, this approach might have offered a more memory-efficient way to encode spatial dependencies, providing moderate computational overhead while allowing the model to reason about relative spatial layouts more effectively and potentially improve anomaly detection performance.

These advanced encodings are expected to improve the model's ability to maintain anatomical consistency and accurately restore abnormal regions, addressing the limitations of binary encoding.

3.2 Swin Transformer Architecture

Description:The Swin Transformer is a hierarchical vision transformer that computes self-attention within local windows while allowing cross-window connections through shifted windows. It produces multi-scale feature maps, making it suitable for dense prediction tasks. Swin blocks include window-based multi-head self-attention and feed-forward layers, and the architecture efficiently models spatial hierarchies without relying on fixed convolutional kernels or explicit positional embeddings.

Incorporation: The generator design was restructured from its earlier CNN-based, memory-conditioned form to a transformer-driven approach. Instead of convolutional encoders and a memory-augmented bottleneck, the new version employs a hierarchical transformer backbone for feature extraction and a redesigned decoder that emphasizes transformer-style attention during upsampling. Skip connections between encoder and decoder were preserved, but the explicit positional encodings and memory-conditioning mechanisms were removed. The result is a more streamlined encoder–decoder pipeline that relies on transformer attention rather than handcrafted conditioning for representation learning.

Predicted Impact: These modifications were predicted to enhance the model's ability to capture hierarchical spatial relationships and maintain anatomical consistency in generated images. The Swin-based architecture was predicted to enable more flexible and data-driven spatial modeling and improve robustness to variations in input alignment. Without explicit memory-conditioning, some control over localized feature preservation was predicted to be reduced, so the improvements were mainly ex-

pected to rely on the Swin encoder's inherent hierarchical representation.

Chapter 4

Results and Discussion

The results and discussion focus on evaluating the proposed spatially aware unsupervised anomaly detection model on X-ray images. The datasets used in this study are briefly revisited to provide context for interpreting the findings, highlighting their composition, preprocessing procedures, and relevance to the anomaly detection task.

Experimental outcomes are reported, emphasizing the effects of the implemented spatial learning enhancements, such as positional encoding and the integration of Swin Transformer architectures in place of conventional convolutional components within the GAN framework.

Comparative analyses with baseline models demonstrate the effectiveness of the proposed improvements. Both quantitative metrics and qualitative assessments are examined to provide a comprehensive understanding of the model's performance, while identifying strengths, limitations, and potential directions for future enhancement.

4.1 Dataset

RSNA Pneumonia Detection Challenge Dataset

The RSNA dataset is used for pneumonia detection from x-rays. It consists of the following sets:

- **Normal Training Set:** 3,851 images.
- **Unlabeled Training Set:** 9,012 images, including 4,000 normal and 5,012 abnormal images.
- **Testing Set:** 2,000 images, with 1,000 normal and 1,000 abnormal images.

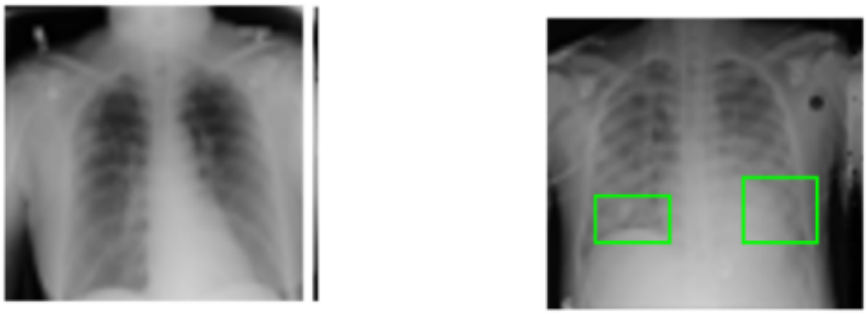


Figure 4.1: Example X-ray from the RSNA Pneumonia Detection Challenge dataset.

VinBigData X-ray Abnormalities Detection Challenge Dataset (VinDr-CXR)

The VinDr-CXR dataset is used for detecting a wide range of abnormalities from X-ray images. The dataset contains:

- **Normal Training Set:** 4,000 images.
- **Unlabeled Training Set:** 9,000 images, including 5,606 normal and 3,394 abnormal images.
- **Testing Set:** 2,000 images, with 1,000 normal and 1,000 abnormal images.

Figure 4.2: Example X-ray from the VinDr-CXR dataset.

Large-scale Attention-based Glaucoma (LAG) Dataset

The LAG dataset focuses on glaucoma detection using retinal fundus images. It provides the following sets:

- **Normal Training Set:** 1,500 images.

- **Unlabeled Training Set:** 1,732 images, including 832 normal and 900 abnormal samples.
- **Testing Set:** 1,622 images, with 811 normal and 811 abnormal images.

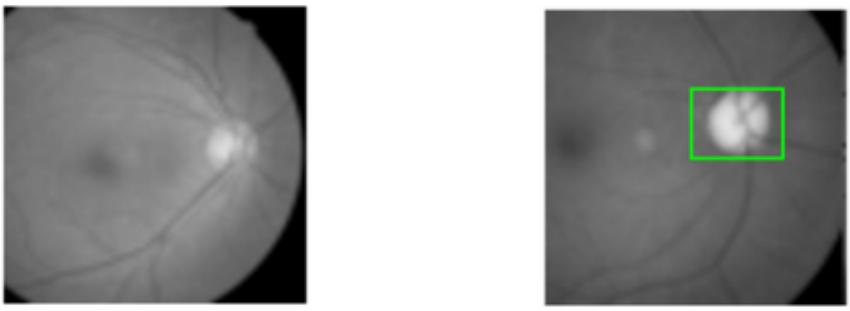


Figure 4.3: Example fundus image from the LAG dataset.

4.2 Evaluation Strategy

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a widely used metric for evaluating the performance of binary classification models. It measures the area under the ROC curve, which plots the True Positive Rate (TPR, or sensitivity) against the False Positive Rate (FPR, or 1-specificity) across different decision thresholds. Mathematically, AUC represents the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance.

Interpretation of AUC

- **AUC = 1:** Perfect discrimination — the model correctly distinguishes all positive and negative cases.
- **AUC = 0.5:** No discriminative power — equivalent to random guessing.
- **AUC < 0.5:** Worse than random guessing — predictions are inversely correlated with the true labels.

AUC in Medical Anomaly Detection

In medical anomaly detection, AUC-ROC is particularly suitable because:

- It provides a **threshold-independent** assessment of model performance, which is critical when the optimal operating threshold has not yet been clinically defined.

- It handles **class imbalance** effectively, as it evaluates ranking quality rather than absolute classification accuracy.
- It summarizes the trade-off between **sensitivity (recall)** and **specificity**, both of which are essential in clinical decision-making.

Limitations and Complementary Metrics

While AUC is a robust summary metric, it does not convey information about probability calibration or performance at specific clinical thresholds. Two models with similar AUC values may behave differently at clinically relevant decision points. For highly imbalanced datasets, the Precision–Recall AUC may provide additional insight, and reporting sensitivity, specificity, or F1-score at a chosen threshold can complement the AUC-based evaluation.

In this work, AUC-ROC is used as the primary evaluation metric to compare models on their ability to distinguish normal from abnormal X-ray images, with higher AUC values indicating superior discriminative performance.

4.3 Quantitative Analysis

The results of incorporating positional encodings into various anomaly detection models are summarized in Table 4.1. Classical GAN-based models (**AnoGAN**, **ALAD**, **GANomaly**) show moderate AUCs, while **Brainomaly** and **SAGAN** achieve higher scores through enhanced feature extraction and attention mechanisms.

The **Custom** models with positional encodings (Sinusoidal, Learned, RoPE, ALiBi, Attention Bias) show only marginal improvements in mean AUC values over SAGAN, which are effectively negated by the variability within the reported ranges, suggesting limited to no benefit from these encodings. TThe **SWIN**-based model, despite its advanced multi-scale attention architecture, performs slightly worse than SAGAN, suggesting that its hierarchical global view does not contribute significant additional value beyond the spatial modeling already captured by SAGAN.

Table 4.1: Comparison of AUC (%) with variability ranges for different models across three datasets.

Model	VinDr-CXR	RSNA	LAG
ANOGAN	79.13 ± 1.4	80.11 ± 1.9	82.37 ± 1.3
ALAD	82.12 ± 1.8	82.89 ± 1.5	87.31 ± 1.7
GANOMALY	81.92 ± 1.2	83.22 ± 1.6	86.77 ± 1.3
BRAINOMALY	88.50 ± 1.5	90.03 ± 1.9	95.88 ± 1.4
SAGAN	91.32 ± 1.7	92.21 ± 1.3	96.93 ± 1.6
CUSTOM (Sinusoidal)	**91.47** ±1.9	**92.75** ±1.4	97.10 ±1.7
CUSTOM (Learned)	91.39 ± 1.3	92.13 ± 1.5	97.02 ± 1.8
CUSTOM (RoPE)	91.09 ± 1.6	92.69 ± 1.4	96.95 ± 1.5
CUSTOM (Alibi)	91.43 ± 1.4	92.65 ± 1.9	97.08 ± 1.2
CUSTOM (Attention Bias)	91.45 ± 1.5	92.22 ± 1.8	**97.13** ±1.4
CUSTOM (SWIN)	88.45 ± 1.8	90.13 ± 1.4	94.55 ± 1.7

4.4 Qualitative Analysis

To further assess the impact of positional encodings and architectural changes, we qualitatively compared anomaly localization outputs across **SAGAN** and the **Custom** models. Representative examples are shown in Figure 4.4. Across these samples, the highlighted anomaly regions are visually very similar, and differences between models are often subtle or indistinguishable.

The **Custom** models, which incorporated various positional encodings—Sinusoidal, Learned, RoPE, ALiBi, and Attention Bias—produced results closely resembling SAGAN. While these encodings were designed to inject additional spatial context, the similarity of outputs suggests that SAGAN's attention mechanisms and skip connections already capture the key spatial relationships in x-rays.

Positional encodings, in theory, should enhance reasoning about the relative positions of anatomical structures. However, the nearly identical heatmaps and bounding regions across models (Fig. 4.4) indicate that this additional information was largely redundant. This is consistent with the quantitative results, where only marginal gains were observed, and variability within AUC ranges weakened the case for consistent improvement.

Interestingly, the **SWIN**-based Custom model also produced outputs comparable to SAGAN but, in some cases, less precise. Although the hierarchical transformer design

provides strong global context, it appeared to diffuse attention across broader structural patterns, occasionally at the expense of fine-grained anomaly localization. For x-rays, where subtle local deviations often signal anomalies, strong localized attention and effective skip connections remain more critical than multi-scale global modeling.

In summary, the qualitative analysis reinforces the quantitative findings: across different design choices, outputs are visually very similar and difficult to differentiate. The strongest performance drivers remain local attention mechanisms and skip connections, whereas positional encodings or hierarchical context alone do not substantially improve anomaly detection in x-rays.

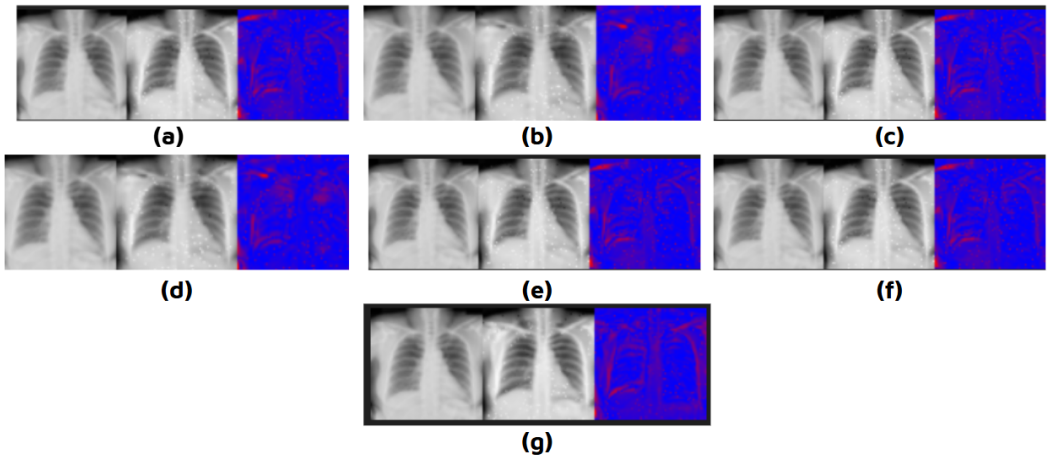


Figure 4.4: Qualitative comparison of anomaly localization across **SAGAN** and **Custom** models. Starting from (a) to (g), the outputs correspond to the same input image but across different models: (a) SAGAN, (b) Custom (Sinusoidal), (c) Custom (Learned), (d) Custom (RoPE), (e) Custom (ALiBi), (f) Custom (Attention Bias), and (g) Custom (SWIN). Despite the architectural differences, the highlighted anomaly regions are visually very similar, with only marginal differences that are hard to differentiate. In this example, all models produce a false result, illustrating that positional encodings and hierarchical global context provide little additional benefit beyond strong local attention and skip connections.

Chapter 5

Conclusion

This study evaluates enhancements to the Spatial-aware Attention Generative Adversarial Network (SAGAN[15]) through the integration of positional encodings (Sinusoidal, Learned, RoPE, ALiBi, Attention Bias) and Swin Transformer backbones. Across the VinDr-CXR, RSNA, and LAG datasets, classical GAN-based methods such as AnoGAN, ALAD, and GANomaly achieve moderate AUCs, while Brainomaly and SAGAN provide stronger baselines through improved feature extraction and attention mechanisms.

The Custom models with positional encodings show only marginal AUC improvements over SAGAN, with variability ranges negating consistent gains. Qualitative comparisons further reveal that outputs from Custom models remain visually similar to SAGAN, making differences difficult to distinguish. This indicates that SAGAN's native attention mechanisms and skip connections already capture the spatial dependencies necessary for anomaly detection in x-rays. The Swin-based Custom model performs slightly worse, suggesting that hierarchical global context does not contribute significant additional value for detecting localized anomalies in this domain.

Overall, the findings indicate that the key drivers of performance in unsupervised X-ray anomaly detection are strong localized attention and skip connections. Additional spatial encodings or hierarchical modeling, while theoretically promising, do not yield substantial improvements in practice. These insights highlight the importance of balancing architectural complexity with practical performance gains when advancing computer-aided diagnosis systems.

References

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," *arXiv preprint arXiv:1805.06725*, 2018. [Online]. Available: https://arxiv.org/abs/1805.06725

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *arXiv preprint arXiv:1701.07875*, 2017. [Online]. Available: https://arxiv.org/abs/1701.07875

[3] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, *Transformer-xl: Attentive language models beyond a fixed-length context*, 2019. arXiv: 1901.02860 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1901.02860

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2017. [Online]. Available: https://arxiv.org/abs/1611.07004

[5] O. Press, N. A. Smith, and M. Lewis, *Train short, test long: Attention with linear biases enables input length extrapolation*, 2022. arXiv: 2108.12409 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2108.12409

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. [Online]. Available: https://arxiv.org/abs/1511.06434

[7] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," *CoRR*, vol. abs/1802.09088, 2018. [Online]. Available: http://arxiv.org/abs/1802.09088

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *arXiv preprint arXiv:1606.03498*, 2016. [Online]. Available: https://arxiv.org/abs/1606.03498

[9] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," *CoRR*, vol. abs/1703.05921, 2017. [Online]. Available: http://arxiv.org/abs/1703.05921

[10] M. M. R. Siddiquee, J. Cho, A. K. Mondal, and T. F. Syeda-Mahmood, "Brainomaly: Unsupervised neurologic disease detection utilizing unannotated t1-weighted brain mr images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 8527–8536. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11078334/

[11] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, *Roformer: Enhanced transformer with rotary position embedding*, 2023. arXiv: 2104.09864 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2104.09864

[12] Y. Sun, N. Zhao, and E. Adeli, "Healthygan: Learning from unannotated medical images to detect anomalies associated with human disease," *Medical Imaging with Deep Learning*, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11062325/

[13] A. Vaswani et al., *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762

[14] H. Zenati, M. Romain, C. S. Foo, B. Lecouat, and V. R. Chandrasekhar, "Adversarially learned anomaly detection," *CoRR*, vol. abs/1812.02288, 2018. [Online]. Available: http://arxiv.org/abs/1812.02288

[15] Z. Zhang et al., *Spatial-aware attention generative adversarial network for semi-supervised anomaly detection in medical image*, 2024. arXiv: 2405.12872 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2405.12872