LANODET: SELF-SUPERVISED LUNG ANOMALY DETECTION WITH TRANSFORMER-BASED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting anomalies in chest X-rays is central to the timely diagnosis of lung disease, but labelled abnormal images are scarce and the abnormalities are often subtle. Reconstruction-based generative models trained only on normal images offer a label-free alternative in which an anomaly is revealed by the error between an image and its reconstruction. We identify three limitations of current diffusion-based detectors: convolutional U-Net backbones capture local texture but not the long-range structure of the thorax, whereas global self-attention costs quadratic time in the number of patches; the noise-prediction objective and stochastic sampler inject sampling variance into the very residual that serves as the anomaly score; and a fixed timestep code gives the network no control over how diffusion time is represented. We propose LAnoDet, in which each component answers one of the above limitations: (1) a time-embedded Transformer encoder–decoder whose attention is factorised into row and column multi-head attention (RC-MHA), reducing the cost of attention from $O(N^2 d)$ to $O(N^{1.5} d)$ while preserving structured spatial dependencies; (2) a deterministic reverse process that reconstructs the clean image at every timestep with a mean-squared-error objective instead of predicting and subtracting noise; and (3) a learnable timestep embedding added to the patch and positional embeddings. Trained only on normal chest X-rays from COVIDx CXR-4, LAnoDet reaches 96.2% accuracy, 97.3% sensitivity, 96.9% F1 score and 96.51% AUC on COVID-19 detection and transfers without retraining to a pneumonia dataset with 96.78% accuracy, 98.34% sensitivity, 97.71% F1 score and 97.16% AUC, matching or exceeding supervised classifiers reported on these datasets. A sixteen-variant ablation shows that row–column attention, skip connections and the learnable time embedding each contribute, the full model exceeding its global self-attention counterpart by 4.17 AUC points.

1 INTRODUCTION

Chest radiography is the first imaging examination for most thoracic disease because it is inexpensive and widely available, and lung diseases remain among the leading causes of death worldwide (GBD Chronic Respiratory Disease Collaborators, 2020). Deep networks trained with labels detect specific findings on chest X-rays at expert level (Rajpurkar et al., 2017; Wang et al., 2017; Irvin et al., 2019), but they require large annotated datasets, recognise only the findings they were trained on, and vary in performance across institutions (Zech et al., 2018). Anomaly detection takes the opposite route: a model is trained only on normal images, and any image that the model cannot explain is flagged (Ruff et al., 2021). The dominant family of such models is reconstruction-based. Auto-encoders (Baur et al., 2018; Chen & Konukoglu, 2018), variational auto-encoders (Zimmerer et al., 2018; 2019), generative adversarial networks (Schlegl et al., 2017; 2019), vector-quantised auto-encoders (Naval Marimont & Tarroni, 2021) and, most recently, denoising diffusion probabilistic models (Wyatt et al., 2022; Wolleb et al., 2022; Pinaya et al., 2022) learn to reproduce normal anatomy, and the pixel-wise difference between an image and its reconstruction is read as an anomaly map. Diffusion models are the strongest members of this family: their iterative denoising produces sharper reconstructions than auto-encoders and avoids the mode collapse and unstable training of adversarial models (Ho et al., 2020; Dhariwal & Nichol, 2021; Kazerouni et al., 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

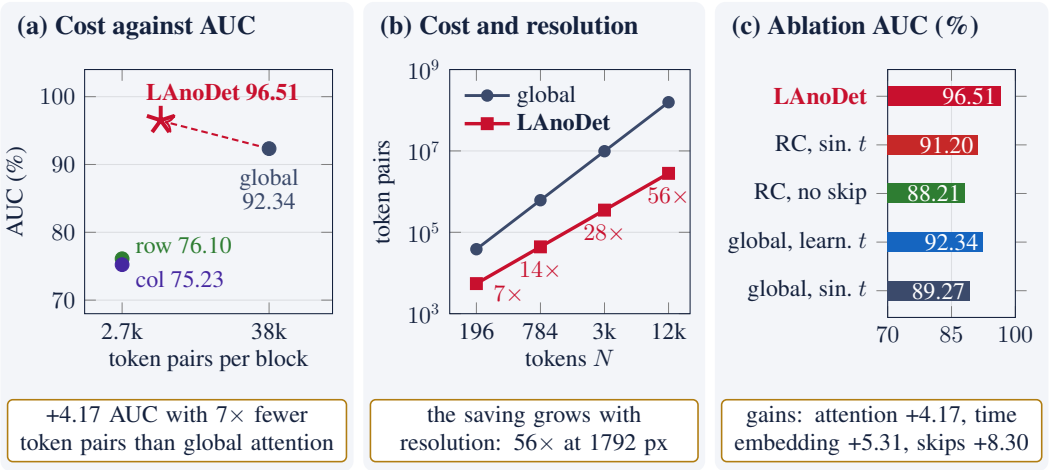

Figure 1: Motivational analysis: (a) attention cost against AUC for the four attention types; (b) attention cost with input resolution; (c) AUC of the ablation settings on COVIDx CXR-4 (RC: row–column attention without the other LAnoDet choices; sin./learn.: sinusoidal or learnable time embedding).

In our study, we find that the way current diffusion-based detectors are built limits how well the reconstruction residual can separate abnormal from normal chest X-rays. We identify three causes. ① *Backbone locality*: the convolutional U-Net (Ronneberger et al., 2015) used by almost all diffusion detectors aggregates information through a growing but still local receptive field, so it reproduces texture faithfully but represents the long-range structure of the thorax (the symmetry of the lung fields, the mediastinum and the diaphragm) only indirectly; the alternative, a Vision Transformer with global self-attention (Dosovitskiy et al., 2021), attends to all N patches at a cost of $O(N^2d)$ (Figure 1a,b) and spreads attention across the image rather than along the rows and columns where local anatomical structure lies. ② *Stochastic objective*: standard DDPMs are trained to predict the noise added to an image and reconstruct it by a stochastic sampler that injects fresh Gaussian noise at every step (Ho et al., 2020); when the reconstruction residual is the anomaly score, this sampling variance is added to the very quantity being measured. ③ *External timestep conditioning*: the timestep enters the network as a fixed sinusoidal code, so the network cannot adapt how diffusion time is represented to the reconstruction task, although the amount of noise to be removed differs by orders of magnitude between early and late steps.

We propose LAnoDet, a self-supervised lung anomaly detector in which each component answers one of the above-discussed causes (Figure 2). (1) A *time-embedded RC-MHA Transformer* replaces the U-Net: the noisy image is split into patches, each patch is embedded together with a positional and a timestep embedding, and a Transformer encoder–decoder with skip connections reconstructs the image; in every block, global self-attention is replaced by *row–column multi-head attention* (RC-MHA), which attends along the rows and the columns of the patch grid and reduces the cost of attention from $O(N^2d)$ to $O(N^{1.5}d)$ while retaining structured spatial dependencies (cause ①). (2) A *deterministic reconstruction objective* trains the network to output the clean image from its noisy version at every timestep under a mean-squared-error loss, and inference applies the same network as a deterministic reconstruction chain, so the residual contains no sampling noise (cause ②). (3) A *learnable timestep embedding* is added to the patch and positional embeddings and is learned end to end with the reconstruction loss (cause ③). An anomaly score is the mean squared error between an image and its reconstruction, and an anomaly map is their absolute difference.

Our main contributions are summarised as follows:

- We propose LAnoDet, a self-supervised anomaly detector for chest X-rays that couples a deterministic, reconstruction-based reverse diffusion process with a time-embedded Transformer encoder–decoder, and we derive the $O(N^{1.5}d)$ cost of its row–column attention against the $O(N^2d)$ cost of global self-attention.
- We show with a sixteen-variant ablation that the three design decisions are complementary: replacing global self-attention by RC-MHA raises AUC from 92.34% to 96.51%, the learnable

timestep embedding raises it from 91.20% to 96.51%, and removing the skip connections lowers it to 88.21%.

- Trained only on 10,664 normal images of COVIDx CXR-4, LAnoDet reaches 96.51% AUC on COVID-19 detection and, without retraining, 97.16% AUC on a pneumonia dataset, matching or exceeding supervised classifiers reported on these datasets while using no labels.

2 Related Work

2.1 Reconstruction-based Anomaly Detection in Medical Imaging

Anomaly detection from normal data only is a long-standing problem (Ruff et al., 2021), and in medical imaging it is dominated by reconstruction: a generative model is fitted to normal images and the residual between an input and its reconstruction indicates pathology. Auto-encoders were the first such models (Baur et al., 2018), but with enough capacity they reproduce abnormal inputs as faithfully as normal ones, so constrained (Chen & Konukoglu, 2018) and context-encoding (Zimmerer et al., 2018) variants were introduced. Variational auto-encoders (Kingma & Welling, 2014; 2019) regularise the latent space and provide density-based scores (Zimmerer et al., 2019), at the price of blurred reconstructions that reduce localisation accuracy. Adversarial models learn the normal manifold with a discriminator, from AnoGAN (Schlegl et al., 2017) and its fast encoder-based successor f-AnoGAN (Schlegl et al., 2019) to one-class GAN classifiers for chest X-rays (Tang et al., 2019), but they are unstable to train and prone to mode collapse. Vector-quantised auto-encoders restore anomalous latents to the nearest codebook entries (Naval Marimont & Tarroni, 2021), and feature in-painting exploits the consistent anatomy of radiographs (Xiang et al., 2023). Unsupervised detection has been evaluated on chest radiographs (Nakao et al., 2021), and recent benchmarks show that the ranking of methods depends on the modality and the pathology (Bercea et al., 2023b; Cai et al., 2025). Diffusion models entered this field with AnoDDPM, which uses partial diffusion with simplex noise (Wyatt et al., 2022), with the classifier-guided weakly supervised detector of Wolleb et al. (2022), and with latent diffusion for brain imaging (Pinaya et al., 2022); masked (Iqbal et al., 2023), patched (Behrendt et al., 2023), automatically masked and re-sampled (Bercea et al., 2023a) and conditioned (Mousakhan et al., 2023) variants followed, and one-step norm-guided denoising has been proposed for industrial anomaly detection (Zhang et al., 2023). All of these detectors keep the convolutional U-Net backbone and the noise-prediction objective of the original DDPM; our work replaces both.

2.2 Diffusion Models and Transformer Backbones

Diffusion models were introduced as a thermodynamics-inspired generative process (Sohl-Dickstein et al., 2015) and made practical by DDPM (Ho et al., 2020), which trains a network to predict the added noise. Improved noise schedules and learned variances (Nichol & Dhariwal, 2021), deterministic implicit samplers (Song et al., 2021a), the score-based formulation (Song et al., 2021b), classifier guidance (Dhariwal & Nichol, 2021) and latent-space diffusion (Rombach et al., 2022) extended the framework, and Salimans & Ho (2022) discuss the alternative parameterisations in which the network predicts the clean image or a velocity rather than the noise. Transformer backbones for diffusion appeared with DiT (Peebles & Xie, 2023) and U-ViT (Bao et al., 2023), both with global self-attention over patch tokens and timestep conditioning through adaptive normalisation or an extra token. The quadratic cost of global attention (Vaswani et al., 2017; Dosovitskiy et al., 2021) has been addressed by windowed attention (Liu et al., 2021), by criss-cross attention (Huang et al., 2019) and by axial attention, which applies attention sequentially along the height and the width of the feature map (Ho et al., 2019; Wang et al., 2020a). LAnoDet belongs to this line: it factorises attention into a row branch and a column branch computed in parallel from the same normalised input and summed, it adds a learnable timestep embedding directly to the tokens, and it is, to our knowledge, the first diffusion detector for chest X-rays built on such a backbone.

2.3 Learning from Chest X-rays without Labels

Self-supervised representation learning provides features without annotation through pretext tasks such as rotation prediction (Gidaris et al., 2018), jigsaw puzzles (Noroozi & Favaro, 2016), context prediction (Doersch et al., 2015) and colourisation (Zhang et al., 2016), through contrastive and

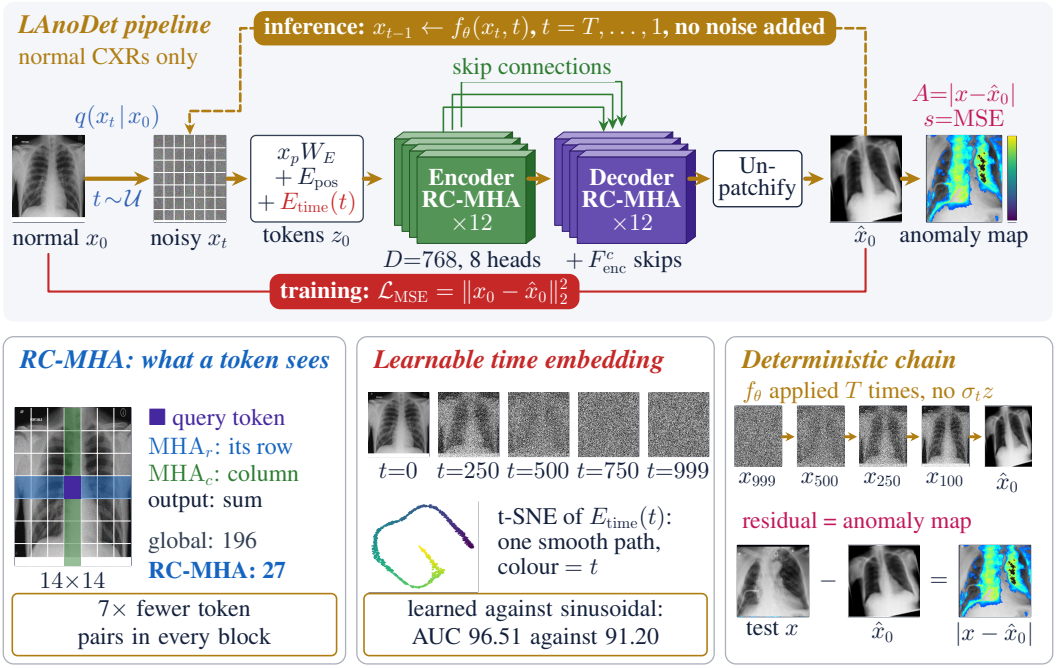

Figure 2: Overview of LAnoDet. Top: the reconstruction pipeline, its training loss and the deterministic inference loop. Bottom: the tokens a query attends to under RC-MHA, the learnable time embedding across the noise levels it encodes, and the deterministic reconstruction chain with the resulting residual.

bootstrapped objectives (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Caron et al., 2020; Khosla et al., 2020; Radford et al., 2021) and through masked image modelling (He et al., 2022). These methods produce an encoder that must still be fine-tuned with labels for diagnosis, whereas a reconstruction-based detector delivers a per-image score and a per-pixel map directly. The supervised classifiers against which we report our results, from transfer-learned CNNs (Sahinbas & Catak, 2021; Hashmi et al., 2020; Liang & Zheng, 2020) and ensembles (Deb & Jha, 2020; El Asnaoui, 2021; Bhatt & Shah, 2023) to COVID-Net (Wang et al., 2020b), residual networks (He et al., 2016; Saber et al., 2024), EfficientNetV2 (Ali et al., 2024) and Vision Transformers (Jiang & Chen, 2022), all use disease labels during training; LAnoDet uses none.

3 Methodology

Overview. As shown in Figure 2, LAnoDet has three components. The *time-embedded RC-MHA Transformer* (Section 3.2) is the denoising network: it embeds the patches of a noisy image together with positional and timestep embeddings and reconstructs the image with a Transformer encoder–decoder whose attention is factorised into row and column branches. The *deterministic reconstruction process* (Section 3.3) trains this network to output the clean image from its noisy version at any timestep and applies it at inference as a reconstruction chain. *Anomaly scoring* (Section 3.4) turns the residual between an image and its reconstruction into a per-image score and a per-pixel map. The model is trained only on normal chest X-rays.

3.1 Preliminaries

Problem setting. Let $\mathcal{X}_n$ be a set of normal chest X-rays $x_0 \in \mathbb{R}^{H \times W \times C}$. We learn a reconstruction function f_θ from $\mathcal{X}_n$ alone. At test time an image x of unknown status receives a score $s(x)$ that is large when x deviates from normal anatomy, and a map $A(x) \in \mathbb{R}^{H \times W}$ that indicates where.

Forward diffusion. Following DDPM (Ho et al., 2020), Gaussian noise is added to a clean image over T steps with a variance schedule $b_t \in (0, 1)$,

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-b_t}\, x_{t-1}, b_t \mathbf{I}\big), \qquad a_t = 1 - b_t, \qquad \bar{a}_t = \textstyle\prod_{s=1}^{t} a_s, \tag{1}$$

so that x_t can be sampled directly from x_0,

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{a}_t}\, x_0, (1-\bar{a}_t)\mathbf{I}\big), \qquad x_t = \sqrt{\bar{a}_t}\, x_0 + \sqrt{1-\bar{a}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

A standard DDPM trains a network $\epsilon_\theta(x_t, t)$ to predict ϵ and generates by the stochastic reverse chain $x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z$ with $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Because x_0 and ϵ are related through Equation (2), a network can equivalently be trained to predict x_0 (Salimans & Ho, 2022); LAnoDet adopts this parameterisation and removes the stochastic term from the reverse chain (Section 3.3).

Global self-attention. A Vision Transformer (Dosovitskiy et al., 2021) splits $x \in \mathbb{R}^{H \times W \times C}$ into $N = HW/P^2$ patches of size $P \times P$, maps each to a D-dimensional token, adds a positional embedding E_{pos} and a class token, and applies multi-head self-attention (MHSA) in every block,

$$\mathrm{MHSA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O, \qquad \mathrm{head}_i = \mathrm{Softmax}\Big(\tfrac{Q_i K_i^\top}{\sqrt{d_k}}\Big) V_i, \tag{3}$$

where $Q, K, V \in \mathbb{R}^{N \times D}$ are linear projections of the tokens. Since every token attends to every other token, the cost of one block is $O(N^2 d)$ in the token count.

3.2 Time-embedded RC-MHA Transformer

Observation 1. *Global attention is quadratic and diffuse; anatomical structure is row- and column-aligned.* The cost in Equation (3) grows quadratically with N, which becomes the bottleneck for the high-resolution images used in radiology, and the attention it learns is spread over the whole image. The structures that a chest X-ray detector must reproduce, the lung fields, the ribs, the mediastinum and the diaphragm, are organised along the horizontal and vertical axes of the image, and an abnormality is a local departure from this organisation. Attention that is restricted to the row and the column of each token therefore covers the relevant context at a fraction of the cost, and the ablation in Section 4.3 confirms that it also detects anomalies better than global attention (96.51% against 92.34% AUC, Figure 1a).

Token embedding with time. The noisy image x_t is split into N patches $x_p \in \mathbb{R}^{N \times (P^2 C)}$ and embedded as

$$z_0 = x_p W_E + E_{\text{pos}} + E_{\text{time}}(t), \tag{4}$$

where W_E is the patch projection, E_{pos} the positional embedding and $E_{\text{time}}(t)$ a learnable embedding of the diffusion timestep produced by an embedding module of width 512 followed by an MLP. No class token is used, because the output of the network is an image rather than a label. In contrast to diffusion models that inject a fixed sinusoidal code of t, the timestep representation is learned end to end with the reconstruction loss; Section 4.4 shows that the learned embeddings form a smooth trajectory over t, and Section 4.3 shows that they raise AUC from 91.20% to 96.51% relative to the sinusoidal code.

Row–column multi-head attention. Let the tokens of a block be arranged on their $h \times w$ grid, $F \in \mathbb{R}^{h \times w \times D}$ with $N = hw$. RC-MHA applies multi-head attention within each row and within each column of the grid and sums the two results,

$$\text{RC-MHA}(F) = \mathrm{MHA}_r(F) + \mathrm{MHA}_c(F), \tag{5}$$

where MHA_r attends among the w tokens of each of the h rows and MHA_c among the h tokens of each of the w columns, both computed from the same layer-normalised input (Ba et al., 2016). Each block is $F' = F + \text{RC-MHA}(\mathrm{LN}(F))$, $F'' = F' + \mathrm{MLP}(\mathrm{LN}(F'))$. The two branches are computed in parallel and summed rather than composed sequentially as in axial attention (Ho et al., 2019; Wang et al., 2020a), so a token receives its row context and its column context on equal footing in a single step.

Complexity. Row attention costs $O(h \cdot w^2 \cdot d)$ over the h rows of length w, and column attention $O(w \cdot h^2 \cdot d)$ over the w columns of height h, so a block costs $O\big((hw^2 + wh^2)\, d\big)$. For an approximately square grid, $h \approx w \approx \sqrt{N}$, this becomes

$$O\big(\sqrt{N} \cdot N \cdot d + \sqrt{N} \cdot N \cdot d\big) = O(N^{1.5}\, d), \tag{6}$$

against $O(N^2 d)$ for global MHSA in Equation (3). For the 14×14 grid of a 224×224 input with $P = 16$, a global block computes $196^2 = 38{,}416$ token pairs per head and an RC-MHA block $2 \times 14 \times 14^2 = 5{,}488$, a sevenfold saving that reaches 14 at 448×448 input and 56 at 1792×1792 (Figure 1b); the ablation in Section 4.3 shows that this saving comes together with, not instead of, an accuracy gain.

Encoder, decoder and skip connections. The encoder consists of 12 RC-MHA blocks and the decoder mirrors it with 12. To retain the spatial detail the encoder captures, the output of encoder block c is added to the input of the corresponding decoder block,

$$F^c_{\text{dec}} = \text{Decoder}_c\big(F^{c-1}_{\text{dec}}\big) + F^c_{\text{enc}}, \qquad \hat{x}_{t-1} = \text{Unpatchify}(z_L), \tag{7}$$

where z_L is the output of the last decoder block and Unpatchify reassembles the patch tokens into an image. Removing these connections lowers AUC from 96.51% to 88.21% (Section 4.3).

3.3 Deterministic Reconstruction Process

Motivation. When the reconstruction residual is the anomaly score, everything that separates the reconstruction from the input counts as evidence of abnormality. In a standard DDPM two sources of discrepancy have nothing to do with pathology: the network is trained to predict noise rather than anatomy, so its errors are expressed in noise space and reach the image only through the reverse update, and the sampler adds a fresh Gaussian term $\sigma_t z$ at every step, so two runs on the same image give different reconstructions. LAnoDet removes both sources by training the network to output the clean image directly and by reconstructing deterministically.

Training. For a normal image x_0 and a timestep t sampled uniformly from $\{1, \ldots, T\}$, the noisy image x_t is formed by Equation (2) and the network is trained to reconstruct x_0,

$$\hat{x}_0 = f_\theta(x_t, t), \qquad \mathcal{L}_{\text{MSE}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \big\| x_0^{(i)} - \hat{x}_0^{(i)} \big\|_2^2, \tag{8}$$

where N_b is the batch size. Because the loss is measured in image space at every timestep, the network is optimised for exactly the quantity that is later used as the anomaly score, and the gradient does not pass through a stochastic sampling step.

Inference. At test time the reverse process is a deterministic chain of reconstructions,

$$\hat{x}_{t-1} = f_\theta(x_t, t), \quad t = T, \ldots, 1, \qquad \hat{x}_0 = f_\theta\big(f_\theta(\ldots f_\theta(x_T, T) \ldots, 2), 1\big), \tag{9}$$

in which each output is fed back as the input of the next step and no noise is added. The iterative refinement of diffusion is retained, but the trajectory is a function of the input alone. Algorithm 1 in Section A summarises training and inference.

3.4 Anomaly Scoring

A model trained only on normal images reconstructs normal anatomy with a small error and abnormal regions with a large one. The per-image score is the mean squared error between input and reconstruction, and the anomaly map is the absolute pixel-wise difference,

$$s(x) = \frac{1}{HW} \big\| x - \hat{x}_0 \big\|_2^2, \qquad A(x) = \big| x - \hat{x}_0 \big|. \tag{10}$$

The score is a continuous measure of deviation from normal chest X-ray patterns and is used directly for AUC; for binary decisions, a threshold τ is selected on the validation set as the cutoff that maximises the F1 score, and an image is classified as abnormal when $s(x) > \tau$. The anomaly map is rendered as a colour overlay on the input.

Table 1: Comparison with supervised classifiers on COVIDx CXR-4 (COVID-19 against normal) and on the pneumonia dataset (pneumonia against normal). LAnoDet is trained once on normal COVIDx CXR-4 images and evaluated on both datasets without retraining. Mean ± standard deviation, in %. Without any abnormal label, LAnoDet matches or exceeds every supervised classifier on both datasets.

Dataset	Method	Acc ↑	Sen ↑	F1 ↑	AUC ↑
COVIDx CXR-4	VGG16 (Sahinbas & Catak, 2021)	80.21 ± 2.3	80.32 ± 2.1	80.00 ± 2.4	80.00 ± 2.5
	Ensemble (Deb & Jha, 2020)	90.64 ± 1.5	82.00 ± 2.0	64.81 ± 2.7	90.64 ± 1.7
	CNN (Gazda et al., 2021)	84.93 ± 2.3	77.43 ± 2.1	90.67 ± 2.4	83.44 ± 2.5
	COVID-Net (Wang et al., 2020b)	93.32 ± 1.0	91.00 ± 1.2	91.00 ± 1.1	92.12 ± 1.3
	COVIDX-Net (Hemdan et al., 2020)	90.00 ± 1.8	91.12 ± 1.5	91.00 ± 1.3	91.32 ± 1.4
	ResNet-101 (Saber et al., 2024)	95.19 ± 0.6	95.59 ± 0.5	94.19 ± 0.6	94.86 ± 0.6
	ResNet-50 (Saber et al., 2024)	96.04 ± 0.5	96.70 ± 0.5	94.90 ± 0.6	95.77 ± 0.5
	Channel-Attention (Zhao et al., 2023)	90.00 ± 1.8	91.12 ± 1.5	91.00 ± 1.3	91.34 ± 1.4
	LAnoDet (ours)	96.20 ± 1.6	97.30 ± 0.6	96.90 ± 0.8	96.51 ± 0.9
Pneumonia	CNN (Bhatt & Shah, 2023)	85.58 ± 2.1	83.33 ± 2.0	96.15 ± 2.2	89.29 ± 2.2
	EfficientNetV2L (Ali et al., 2024)	94.02 ± 1.5	94.40 ± 1.0	97.29 ± 1.7	95.50 ± 1.7
	ViT (Jiang & Chen, 2022)	91.19 ± 1.3	91.82 ± 1.1	89.36 ± 2.1	90.34 ± 1.5
	ResNet-50 (Saber et al., 2024)	91.67 ± 0.6	92.04 ± 0.5	94.87 ± 0.5	93.43 ± 0.6
	Transfer learning (Hashmi et al., 2020)	93.00 ± 1.5	89.20 ± 1.8	86.00 ± 2.3	81.40 ± 2.2
	GAN (Joby et al., 2021)	80.20 ± 2.1	91.40 ± 1.5	89.50 ± 1.6	85.00 ± 2.2
	Transfer learning (Liang & Zheng, 2020)	90.50 ± 1.5	96.70 ± 1.5	94.90 ± 1.6	95.30 ± 1.5
	Ensemble learning (El Asnaoui, 2021)	95.00 ± 1.8	94.40 ± 1.5	94.80 ± 1.3	96.30 ± 1.4
	LAnoDet (ours)	96.78 ± 0.7	98.34 ± 0.8	97.71 ± 0.6	97.16 ± 0.9

4 Experiments

4.1 Settings

Datasets. We train on COVIDx CXR-4 (Wu et al., 2023), a multi-institutional benchmark of 83,173 chest X-rays with 67,863 training, 8,473 validation and 8,482 test images labelled normal or COVID-19. Because LAnoDet is trained only on normal anatomy, we use the 10,664 normal training images for training and 4,232 normal validation images for validation, and we test on both normal and COVID-19 images. To assess generalisation, the same trained model is evaluated without retraining on the Kaggle pneumonia dataset of Kermany et al. (2018). All images are resized to 224×224, augmented with random horizontal flips (probability 0.5) and random rotations within $15°$, and normalised to mean 0.5 and standard deviation 0.5.

Implementation. The encoder uses a patch size of 16×16, 12 Transformer blocks with eight attention heads and an embedding dimension of 768; the decoder uses 12 RC-MHA blocks with skip connections from the encoder; the learnable time-embedding module has width 512. Diffusion uses $T = 1{,}000$ steps with a linear noise schedule from 1×10^{-4} to 2×10^{-2}. The model is trained for 1,000 epochs in PyTorch (Paszke et al., 2019) on an NVIDIA GeForce RTX 4090 GPU with 64 GB of system RAM, with a batch size of 64 and the Adam optimiser (Kingma & Ba, 2015) at a learning rate of 1×10^{-4}, minimising Equation (8).

Metrics and compared methods. We report accuracy, sensitivity, F1 score and AUC, as mean ± standard deviation, on the binary task of separating abnormal (COVID-19 or pneumonia) from normal images, and compare with the eight supervised classifiers per dataset listed in Table 1. All compared models are trained with disease labels; LAnoDet is trained on normal images only.

4.2 Main Results

COVIDx CXR-4. Table 1 (top) compares LAnoDet with supervised classifiers on COVID-19 detection. VGG16 (Sahinbas & Catak, 2021) and the CNN of Gazda et al. (2021) reach moderate performance, with 80.21% accuracy for VGG16 and a high F1 score of 90.67% but reduced sensitivity for the CNN. The ensemble (Deb & Jha, 2020), COVIDX-Net (Hemdan et al., 2020) and the channel-attention network (Zhao et al., 2023) improve AUC and balance sensitivity against F1. ResNet-101 and ResNet-50 (Saber et al., 2024) are the strongest supervised baselines, ResNet-50 reaching 96.04% accuracy and 95.77% AUC. LAnoDet, trained without any COVID-19 image, attains the highest

Table 2: Ablation of the attention mechanism, the skip connections and the time embedding on COVIDx CXR-4 (AUC / F1, in %). All variants use 12-layer encoder and decoder; the proposed configuration is highlighted.

	With skip connections		Without skip connections	
Attention	Learnable time	Sinusoidal time	Learnable time	Sinusoidal time
RC-MHA (LAnoDet)	96.51 / 96.90	91.20 / 90.36	88.21 / 89.52	85.45 / 84.69
Global self-attention	92.34 / 91.08	89.27 / 90.52	83.56 / 84.97	82.50 / 84.35
Row only	76.10 / 75.32	75.21 / 75.32	71.25 / 73.10	70.51 / 71.32
Column only	75.23 / 75.10	72.52 / 73.20	70.25 / 71.56	69.52 / 70.00

value on every metric, with 96.2% accuracy, 97.3% sensitivity, 96.9% F1 score and 96.51% AUC, exceeding ResNet-50 by 0.16 accuracy, 0.60 sensitivity, 2.00 F1 and 0.74 AUC points. The margin is small, but it is obtained under different supervision: ResNet-50 is trained on labelled COVID-19 images, whereas LAnoDet uses normal images only and also produces a pixel-level anomaly map. For chest X-ray screening, where abnormal images are scarce, costly to annotate and unavailable for new diseases, reaching supervised accuracy without abnormal labels is the relevant result.

Pneumonia dataset. Table 1 (bottom) evaluates generalisation to an unseen dataset and an unseen pathology. The CNN of Bhatt & Shah (2023) obtains a high F1 score (96.15%) but low accuracy (85.58%), sensitivity and AUC. EfficientNetV2L (Ali et al., 2024) and the Vision Transformer (Jiang & Chen, 2022) are more balanced; ResNet-50 (Saber et al., 2024) and transfer learning (Liang & Zheng, 2020) are strong, the latter with 96.70% sensitivity and 94.90% F1, and ensemble learning (El Asnaoui, 2021) reaches 95.00% accuracy and 96.30% AUC. All of these models were trained with pneumonia labels. LAnoDet, which never saw a pneumonia image or a label, attains 96.78% accuracy, 98.34% sensitivity, 97.71% F1 score and 97.16% AUC, the highest on all four metrics. Ensemble learning and EfficientNetV2L come within 0.86 and 1.66 AUC points, but both are trained on labelled images of this dataset, whereas LAnoDet is applied without retraining to a pathology it has never seen; this transfer is the setting in which a label-free detector is most useful, and it supports the generalisability of the learned notion of normal anatomy.

4.3 Ablation Study

Table 2 and Figure 1c evaluate the three design decisions on COVIDx CXR-4 with AUC and F1 score. Every attention type (RC, global self-attention, row only, column only) is combined with and without skip connections and with a learnable or a sinusoidal time embedding, sixteen variants in all, under the same training settings.

Attention mechanism. With skip connections and the learnable time embedding, RC-MHA reaches 96.51% AUC and 96.90% F1, against 92.34% and 91.08% for global self-attention, the best of its variants. Row-only and column-only attention fall to 76.10% and 75.23% AUC at best: each branch alone sees one axis of the patch grid, and it is their sum in Equation (5) that recovers, and then exceeds, global attention. The gain of RC-MHA over self-attention holds in every setting of the other two factors.

Time embedding. Replacing the learnable time embedding by the fixed sinusoidal code lowers RC-MHA from 96.51% to 91.20% AUC and from 96.90% to 90.36% F1, and lowers self-attention from 92.34% to 89.27% AUC. The learned embedding is therefore the second largest single factor, which agrees with the smooth temporal structure it acquires (Section 4.4); row-only and column-only attention gain less from it, since their bottleneck is the attention pattern.

Skip connections. Removing the encoder–decoder skip connections lowers RC-MHA from 96.51% to 88.21% AUC with the learnable embedding and from 91.20% to 85.45% with the sinusoidal one, and the same drop of 4 to 9 points appears for every other attention type. Skip connections pass the high-resolution spatial features of the encoder to the decoder and support gradient flow through the 24 blocks; both are needed for a reconstruction accurate enough for the residual to isolate the abnormality.

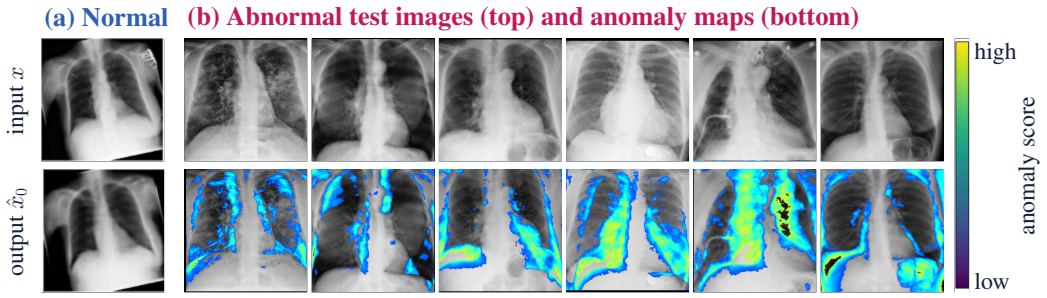

Figure 3: Qualitative results: (a) a normal chest X-ray and its reconstruction; (b) abnormal chest X-rays from the test set (top) and their anomaly maps overlaid on the image (bottom).

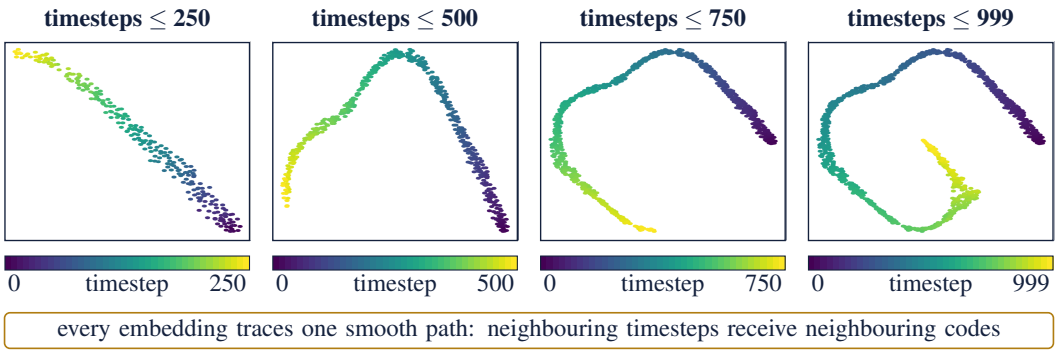

Figure 4: t-SNE of the learned time embeddings for timesteps up to 250, 500, 750 and 999; colour encodes the timestep.

4.4 Analysis

Reconstruction and anomaly maps. Figure 3a shows a normal test image and its reconstruction: the model, trained only on normal chest X-rays, reproduces the lung fields, the cardiac silhouette and the diaphragm with a small residual, so it has learned the underlying normal patterns and removes the noise iteratively without distorting anatomy. Figure 3b shows six abnormal test images with their anomaly maps, obtained by applying a colour map to the pixel-wise absolute difference between input and reconstruction and overlaying it on the input; blue marks low and green to yellow increasing anomaly intensity. The reconstruction departs from the input where the model encounters appearance it has not learned, so the residual is large in the regions it cannot explain, which is how the model both detects and localises the anomaly.

Learned time embedding. We extracted the output of the learnable time-embedding module for each of the 1,000 diffusion timesteps and projected it with t-SNE (van der Maaten & Hinton, 2008) at perplexity 30 (Figure 4). The embeddings form one continuous, smooth trajectory along which the timestep increases monotonically, whether the first 250, 500, 750 or all 999 steps are projected. A smooth trajectory means that neighbouring timesteps receive neighbouring conditioning vectors, so the amount of noise the network expects to remove changes gradually across the reverse chain, the property required for stable step-by-step reconstruction, which a fixed sinusoidal code cannot adapt to the task.

5 Conclusion

We presented LAnoDet, a self-supervised anomaly detector for chest X-rays that combines a deterministic reconstruction-based reverse diffusion process with a time-embedded Transformer encoder–decoder built on row–column multi-head attention. Trained only on normal images, it reaches 96.51% AUC on COVIDx CXR-4 and, without retraining, 97.16% AUC on a pneumonia dataset, matching or exceeding supervised classifiers without any abnormal label, and the ablation shows that row–column attention, the learnable time embedding and the skip connections each contribute.

Ethics statement

This work uses the publicly released COVIDx CXR-4 (Wu et al., 2023) and Kaggle pneumonia (Kermany et al., 2018) chest X-ray datasets, which contain de-identified images; no new data were collected and no patient can be identified from the images used. The detector is intended for research use and not for clinical decision-making.

Reproducibility statement

Section 4.1 specifies the datasets, the training and validation splits, the preprocessing, the architecture, the diffusion schedule and the optimisation settings, and Algorithm 1 gives the complete training and inference procedure. The pneumonia results use the model trained on COVIDx CXR-4 without retraining.

AI use statement

In this work, we used generative AI tools to interpret results, by drafting and revising the text that reports them. We have not used generative AI tools to design the methodology or experiments or to implement methods, and generating synthetic data, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, translation, dataset cleaning and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit parts of the paper, to suggest its structure and to format references. We have reviewed all AI-assisted work: every number in the text, tables and figures was checked against the experiment outputs, and every figure against the method. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Mudasir Ali, Mobeen Shahroz, Urooj Akram, Muhammad Faheem Mushtaq, Stefanía Carvajal Altamiranda, Silvia Aparicio Obregon, Isabel De La Torre Díez, and Imran Ashraf. Pneumonia detection using chest radiographs with novel EfficientNetV2L model. *IEEE Access*, 12:34691–34707, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In *International MICCAI Brainlesion Workshop*, pp. 161–169. Springer, 2018.

Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain MRI. In *Medical Imaging with Deep Learning*, 2023.

Cosmin I Bercea, Michael Neumayr, Daniel Rueckert, and Julia A Schnabel. Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In *ICML Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023a.

Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. In *Medical Imaging with Deep Learning*, 2023b.

Harsh Bhatt and Manan Shah. A convolutional neural network ensemble model for pneumonia detection using chest X-ray images. *Healthcare Analytics*, 3:100176, 2023.

Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MedIAnomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, 102:103500, 2025.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.

Sagar Deep Deb and Rajib Kumar Jha. COVID-19 detection from chest X-ray images using ensemble of CNN models. In *International Conference on Power, Instrumentation, Control and Computing (PICC)*, pp. 1–5. IEEE, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Khalid El Asnaoui. Design ensemble deep learning model for pneumonia disease classification. *International Journal of Multimedia Information Retrieval*, 10(1):55–68, 2021.

Matej Gazda, Ján Plavka, Jakub Gazda, and Peter Drotar. Self-supervised deep convolutional neural network for chest X-ray classification. *IEEE Access*, 9:151972–151982, 2021.

GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Respiratory Medicine*, 8(6):585–596, 2020.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Mohammad Farukh Hashmi, Satyarth Katiyar, Avinash G Keskar, Neeraj Dhanraj Bokde, and Zong Woo Geem. Efficient pneumonia detection in chest X-ray images using deep transfer learning. *Diagnostics*, 10(6):417, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Ezz El-Din Hemdan, Marwa A Shouman, and Mohamed Esmail Karar. COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images. *arXiv preprint arXiv:2003.11055*, 2020.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 603–612, 2019.

Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. Unsupervised anomaly detection in medical images using masked diffusion model. In *International Workshop on Machine Learning in Medical Imaging*, pp. 372–381. Springer, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Zheng Jiang and Liang Chen. Multisemantic level patch merger vision transformer for diagnosis of pneumonia. *Computational and Mathematical Methods in Medicine*, 2022(1):7852958, 2022.

Joshma Joby, Robin Jose Raju, Roshan Job, Sandra Merin Thomas, and Arun KS. PneumoGAN: A GAN based model for pneumonia detection. In *International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, pp. 102–106. IEEE, 2021.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

Gaobo Liang and Lixin Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*, 187:104964, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023.

Takahiro Nakao, Shouhei Hanaoka, Yukihiro Nomura, Masaki Murata, Tomomi Takenaga, Soichiro Miki, Takeyuki Watadani, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. Unsupervised deep anomaly detection in chest radiographs. *Journal of Digital Imaging*, 34(2):418–427, 2021.

Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1764–1767. IEEE, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pp. 117–126. Springer, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Alireza Saber, Pouria Parhami, Alimohammad Siahkarzadeh, Mansoor Fateh, and Amirreza Fateh. Efficient and accurate pneumonia detection using a novel multi-scale transformer approach. *arXiv preprint arXiv:2408.04290*, 2024.

Kevser Sahinbas and Ferhat Ozgur Catak. Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images. In *Data Science for COVID-19*, pp. 451–466. Elsevier, 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Yu-Xing Tang, You-Bao Tang, Mei Han, Jing Xiao, and Ronald M Summers. Abnormal chest X-ray identification with generative adversarial one-class classifier. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pp. 1358–1361. IEEE, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pp. 108–126. Springer, 2020a.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10(1):19549, 2020b.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45. Springer, 2022.

Yifan Wu, Hayden Gunraj, Chi-en Amy Tai, and Alexander Wong. COVIDx CXR-4: An expanded multi-institutional open-source benchmark dataset for chest X-ray image-based computer-aided COVID-19 diagnostics. *arXiv preprint arXiv:2311.17677*, 2023.

Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 650–656, 2022.

Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. SQUID: Deep feature in-painting for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23890–23901, 2023.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018.

Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection. *arXiv preprint arXiv:2303.08730*, 2023.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pp. 649–666. Springer, 2016.

Aite Zhao, Huimin Wu, Ming Chen, and Nana Wang. DCACorrCapsNet: A deep channel-attention correlative capsule network for COVID-19 detection based on multi-source medical images. *IET Image Processing*, 17(4):988–1000, 2023.

David Zimmerer, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 289–297. Springer, 2019.

A Training and Inference Procedure

Algorithm 1 summarises the training and inference procedure of LAnoDet.

Algorithm 1 LAnoDet: training and inference

Require: normal images $\mathcal{X}_n$, network f_θ, schedule $\{b_t\}_{t=1}^{T}$, threshold τ

Training

for each mini-batch $\{x_0\} \subset \mathcal{X}_n$ **do**
  sample $t \sim \mathcal{U}\{1,\dots,T\}$, $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$; $x_t \leftarrow \sqrt{\bar{a}_t}\, x_0 + \sqrt{1-\bar{a}_t}\, \epsilon$ ▷ Equation (2)
  $z_0 \leftarrow x_p W_E + E_{\text{pos}} + E_{\text{time}}(t)$; $\hat{x}_0 \leftarrow f_\theta(x_t, t)$ ▷ Equations (4) and (7)
  update θ with $\nabla_\theta \mathcal{L}_{\text{MSE}}(\hat{x}_0, x_0)$ ▷ Equation (8)
end for

Inference for a test image x

$x_T \leftarrow$ forward diffusion of x to step T
for $t = T,\dots,1$ **do**
  $x_{t-1} \leftarrow f_\theta(x_t, t)$ ▷ Equation (9)
end for
$s(x) \leftarrow \text{MSE}(x, \hat{x}_0)$; $A(x) \leftarrow |x - \hat{x}_0|$; abnormal if $s(x) > \tau$ ▷ Equation (10)