

CLARiTY: A Vision Transformer for Multi-Label Classification and Weakly-Supervised Localization of Chest X-ray Pathologies

John M. Statheros , Hairong Wang , Richard Klein

School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa

Abstract

The interpretation of chest X-rays (CXRs) poses significant challenges, particularly in achieving accurate multi-label pathology classification and spatial localization. These tasks demand different levels of annotation granularity but are frequently constrained by the scarcity of region-level (dense) annotations. We introduce CLARiTY (Class Localizing and Attention Refining Image Transformer), a vision transformer-based model for joint multi-label classification and weakly-supervised localization of thoracic pathologies. CLARiTY employs multiple class-specific tokens to generate discriminative attention maps, and a SegmentCAM module for foreground segmentation and background suppression using explicit anatomical priors. Trained on image-level labels from the NIH ChestX-ray14 dataset, it leverages distillation from a ConvNeXtV2 teacher for efficiency.

Evaluated on the official NIH split, the CLARiTY-S-16-512 (a configuration of CLARiTY), achieves competitive classification performance across 14 pathologies, and state-of-the-art weakly-supervised localization performance on 8 pathologies, outperforming prior methods by 50.7%. In particular, pronounced gains occur for small pathologies like nodules and masses. The lower-resolution variant of CLARiTY, CLARiTY-S-16-224, offers high efficiency while decisively surpassing baselines, thereby having the potential for use in low-resource settings. An ablation study confirms contributions of SegmentCAM, DINO pretraining, orthogonal class token loss, and attention pooling. CLARiTY advances beyond CNN-ViT hybrids by harnessing ViT self-attention for global context and class-specific localization, refined through convolutional background suppression for precise, noise-reduced heatmaps.

Keywords: Chest X-ray, Weakly-supervised localization, Vision transformer, Multi-label classification, Anomaly detection

1. Introduction

X-ray imaging is a commonly used tool for disease diagnosis, enabling medical professionals to identify pathologies within the human body. Chest radiography, in particular, is the most frequently performed X-ray procedure worldwide. Chest X-rays (CXRs) are often employed to diagnose cardiopulmonary disorders, and their low radiation dosage makes them suitable as a screening or triage tool (UNSCEAR, 2010). CXR imaging has facilitated the diagnosis of numerous thoracic diseases, including COVID-19, pneumonia, tuberculosis, and various pneumoconioses (WHO, 2022; Metlay et al., 2019; WHO, 2016; ILO, 2022). However, diagnosis presents several challenges. As X-ray imaging projects 3D anatomy onto a 2D plane, pathologies may be obscured by superimposed organs and skeletal structures. Additionally, some pathologies manifest as small features or diffuse textures and patterns, leading to variability in diagnoses among radiologists (Calli et al., 2021).

The challenges of CXR interpretation, combined with the high volume of exams relative to the number of trained ra-

diologists, have strongly motivated the development of automated diagnostic tools. These include computer-aided detection (CADe) systems—algorithms that analyze medical images to assist clinicians in detecting and characterizing pathologies. In recent years, deep learning has emerged as the dominant and most effective approach for such automation (Hansun et al., 2023). Deep learning models can perform image-level tasks, such as the multi-label classification of diseases or abnormalities in CXRs, and pixel-level tasks, including segmenting or localizing anatomical and pathological regions using bounding boxes (Calli et al., 2021).

These models can achieve high performance but require abundant, high-quality labels for CXRs. Manual labeling by radiologists is expensive and time-intensive, particularly for dense labels like segmentation maps and bounding boxes used in pixel- or region-level tasks. To mitigate annotated data requirements and enhance utility, weakly-supervised methods are employed. These involve training on higher-level labels, such as anatomical region masks or image-level labels, to make predictions at lower levels, like pathology bounding box localization (Jin et al., 2023). A common and effective approach integrates CXR classification with weakly-supervised localization by adapting a classifier network to generate heatmap predictions, from which bounding boxes are derived around high-

*Corresponding author.

Email addresses: johnmstatheros@gmail.com (John M. Statheros , hairong.bau@wits.ac.za (Hairong Wang , richard.klein@wits.ac.za (Richard Klein

activation regions. This label-efficient strategy mitigates the need for ground-truth bounding boxes (Çalli et al., 2021).

Image-level labels in CXR datasets are often extracted automatically via natural language processing (NLP) of radiologists’ reports, reducing costs and enabling large-scale datasets. However, these extraction processes are imprecise and noisy, introducing errors and biases (Rafferty et al., 2025). Underrepresentation of demographic groups can lead to underdiagnosis in trained models, raising ethical concerns (Seyyed-Kalantari et al., 2021). Spurious visual elements have also been linked to generalization issues, manifesting as poor performance on out-of-distribution samples—a phenomenon known as shortcut learning (Ye et al., 2024). DeGrave et al. (2021) demonstrated shortcut signals in COVID-19 CXR diagnosis using explainable artificial intelligence (XAI) techniques. Heatmaps, a form of XAI, can validate model performance and detect shortcut learning. Some FDA-approved commercial CADe systems incorporate heatmaps to enhance explainability (Ait Nasser and Akhloufi, 2023).

Various techniques exist for generating localization heatmaps from CXR classification models, with class activation mapping (CAM) being one of the most prevalent (Feyisa et al., 2023). CAM methods fall into two main categories: gradient-free and gradient-based. Gradient-free CAM generates heatmaps through one or more forward passes without requiring gradient computations, often by aggregating activations from intermediate feature tensors (Zhou et al., 2016). These methods are typically architecture-specific and are commonly implemented in convolutional neural networks (CNNs) or hybrid CNN models. Gradient-based CAM, by contrast, calculates the gradient of a class prediction relative to an arbitrary intermediate feature tensor and uses it to compute a weighted sum across the tensor’s channels, allowing broader applicability to different architectures (Selvaraju et al., 2017).

Vision transformers (ViTs) inherently support heatmap generation via their self-attention mechanism. While approaches like attention rollout and gradient-based CAM have been adapted for ViTs in pathology localization, they frequently underperform. Despite strong image-level classification accuracy, the resulting heatmaps often lack precision, with activations scattered sparsely across the image rather than focused on pathological regions (Qiu et al., 2024). Pretraining strategies significantly influence these outcomes, as self-supervised methods have demonstrated superior heatmap quality for localization compared to supervised pretraining (Barekatain and Glocker, 2025). As a result, standalone ViT self-attention is seldom used for the localization of pathologies in CXRs. More common are hybrid CNN-ViT architectures that employ CAM, or ViTs functioning solely as feature extractors followed by a compact CNN classification head.

Hence, in this paper, we propose CLARiTy—a novel transformer-based model for multi-label classification and weakly-supervised localization of pathologies in chest X-rays. CLARiTy integrates multiple class-specific tokens within a vision transformer backbone. It also incorporates a specialized SegmentCAM module for foreground segmentation and background suppression, orthogonal regularization of class tokens,

and attention pooling. This design enables the model to achieve superior localization accuracy while maintaining competitive classification performance on the NIH ChestX-ray14 benchmark dataset. Our approach leverages anatomical priors derived from pre-existing segmentation models to constrain predictions to clinically relevant regions, reducing reliance on dense annotations. Through extensive experiments, including ablations and comparisons with state-of-the-art methods, we demonstrate CLARiTy’s effectiveness in producing precise, class-specific heatmaps and bounding boxes across pathologies of different sizes, with particularly robust performance for small lesions such as nodules and masses—even at lower image resolutions and under stricter intersection-over-union thresholds.

The key contributions of this study are summarized in the following list:

- A novel model architecture that employs multiple class tokens in a vision transformer to generate class-specific attention maps, enabling more discriminative feature extraction for multiple pathologies in CXRs.
- The introduction of the SegmentCAM module, which performs foreground segmentation and background activation suppression, enhancing localization precision without requiring pixel-level pathology labels.
- An orthogonal class token loss that promotes mutual orthogonality among class representations, complemented by attention pooling to allow embedding dimensions to attend to distinct pathological features, thereby improving both classification and localization performance.
- Empirical validation on the NIH ChestX-ray14 dataset, showing relative improvements in Macro IoU Accuracy of 50.7% over prior methods at stringent thresholds.

The rest of this paper is structured as follows: Sec. 2 reviews related work, Sec. 3 describes the methods and model architecture, Sec. 4 presents the experimental results, including ablations, and Secs. 5 and 6 provide the discussion and conclusion, respectively.

2. Related work

2.1. Deep learning for multi-label chest X-ray pathology classification

Prior to the dominance of deep learning, methods for medical imaging classification relied on hand-crafted features, such as texture descriptors derived from gray-level co-occurrence matrices and density measures (Petrosian et al., 1994), or filtering techniques like Difference of Gaussians (Giordano et al., 2007). These approaches were typically limited to binary classification tasks and struggled with the complexity of multiple pathologies in CXRs. The advent of large-scale public CXR datasets enabled the shift to deep learning for multi-label pathology detection. Key datasets include NIH ChestX-ray14 (Wang et al., 2017), comprising 112,120 images with 14 labels; CheXpert (Irvin et al., 2019), with 224,316 images and

uncertainty-aware labels; MIMIC-CXR (Johnson et al., 2019), featuring 377,110 images paired with radiology reports; PadChest (Bustos et al., 2020), with 160,868 annotated images; and VinDr-CXR (Nguyen et al., 2022), providing 18,000 images with expert-annotated bounding boxes.

Convolutional neural networks (CNNs) have been the predominant architecture for multi-label CXR classification (Çalli et al., 2021; Ait Nasser and Akhloufi, 2023). Wang et al. (2017) introduced baselines on the NIH dataset using ResNet-50, achieving a Macro AUC of 0.745 with the official train-test split. Models generating radiological reports from CXRs, such as TieNet (Wang et al., 2018), leverage text embeddings for feature transfer, enhancing multi-label performance. Subsequent improvements incorporated curriculum learning to model pathology severity (Tang et al., 2018), yielding a Macro AUC of 0.803, and pyramidal networks to retain long-distance spatial information (Xu and Duan, 2024; Alam et al., 2024). To mitigate background biases, some approaches segment and crop lung regions prior to classification (Liu et al., 2019a; Rahman et al., 2020; Sun et al., 2022), improving accuracy for region-specific pathologies. Location-aware supervision further refines attention to class-specific areas (Gündel et al., 2019; Agu et al., 2021; Hossain et al., 2024; Bassi et al., 2024), with ThoraX-PriorNet achieving a Macro AUC of 0.847 by integrating anatomical priors.

However, reported performance varies significantly depending on dataset partitioning strategies (i.e., train/validation/test split). Models trained on random splits often outperform those adhering to the official patient-wise NIH split. For instance, Gündel et al. (2019) reported a 3.4% absolute improvement in mean AUC under a random patient-wise split, while Thorax-Net gained 10.8% with an image-wise split (Wang et al., 2020). Such discrepancies underscore challenges in ensuring fair benchmarking and reproducibility. Moreover, many existing methods prioritize classification over localization, making them vulnerable to shortcut learning from spurious correlations. These limitations are compounded by heavy reliance on large scale labels, which are often derived from reports using NLP, that introduce additional layers of noise and bias (Rafferty et al., 2025).

2.2. Weakly-supervised localization in chest X-ray image analysis

Weakly-supervised localization is a technique designed to overcome the prohibitive cost and scarcity of dense annotations. It achieves this by deriving pixel- or region-level predictions, such as bounding boxes or heatmaps, using only image-level labels. Class activation mapping (CAM) is a cornerstone technique for this task. Gradient-based variants, particularly architecture-agnostic Grad-CAM, have been widely adopted as standard for localization and explainability (Selvaraju et al., 2017; Saporta et al., 2022). The approach was validated in many CXR studies: Irvin et al. (2019) applied Grad-CAM for localizing 14 pathologies in the CheXpert dataset, while others utilized Grad-CAM and Grad-CAM++ for multi-label localization, establishing them as essential baselines (Wang

et al., 2020; Viniavsky et al., 2020). Recent enhancements include isolating foreground pathology signals from background, where these methods typically integrate attention mechanisms on post-backbone features (Ouyang et al., 2021; Guan et al., 2021; Zhu et al., 2022; Wang et al., 2024). In particular, PCAN (Zhu et al., 2022) exemplifies this approach, reporting a Macro IoU Accuracy of 0.103 at T(IoU) = 0.5.

In contrast to gradient-based CAM, gradient-free variants, often architecture-specific to CNNs, aggregate intermediate activations for heatmaps (Zhou et al., 2016). Wang et al. (2017) adapted ResNet-50 with Log-Sum-Exp pooling for gradient-free CAM, generating bounding boxes from heatmaps. They achieved a benchmark Macro IoU Accuracy of 0.062 at T(IoU) = 0.5. Hybrid models like ResNet-DenseNet (Yao et al., 2018) and CNN-ViT (Li et al., 2022) extend this technique, with the latter achieving a state-of-the-art Macro IoU Accuracy of 0.243 at T(IoU) = 0.5. However, CNN-based CAM struggles with irregular or small pathologies; Grad-CAM heatmaps are often large and regular-shaped, leading to poor precision for multiple instances or complex lesions (Saporta et al., 2022). Small pathologies, such as nodules and masses, remain particularly challenging as low-resolution feature maps at the deeper network layers hinder precise localization (Sedai et al., 2018). Even when classification performance matches radiologists (Rajpurkar et al., 2018; Majkowska et al., 2020), localization accuracy often lags behind (Saporta et al., 2022).

Dataset splitting strategies further complicate fair comparisons, as some models are trained on annotated subsets to boost localization performance, potentially introducing information leakage even when explicit bounding boxes are not used (Li et al., 2018; Liu et al., 2019b; Zhao, 2021; Qi et al., 2022). Joint training of classification and localization under weak supervision (using only image-level labels, not bounding boxes)—combined with consistent dataset splits—is important because the tasks are positively correlated (Saporta et al., 2022). However, many such methods still produce noisy, diffuse heatmaps, limiting explainability and bias detection.

2.3. Vision transformer-based approaches for chest X-ray classification and localization

CNNs excel at local feature extraction but struggle to capture global context, and thus rely on deep network layers to expand receptive fields. Vision transformers (ViTs), on the other hand, are designed to model global dependencies directly (Dosovitskiy et al., 2021), though they typically require large amounts of data to outperform CNNs due to lack of built-in inductive biases. In CXR classification and localization tasks, ViTs are often combined with CNNs to exploit the complementary strengths of local and global context modeling, particularly when data are limited.

Hybrid CNN-ViT architectures vary. Series designs either place CNNs first for spatial feature encoding, followed by ViTs for global context modelling (Öztürk et al., 2025), or reverse the order (Fu et al., 2025). Parallel designs, in contrast, fuse outputs from parallel CNN and ViT branches, such as X-Pneumo, which concatenates DenseNet-121 and ViT representations (Pramanik et al., 2025). A notable example, RGT (Han

et al., 2023), employs dual ViT branches: one for classification and attention-based localization, and the other for extracting radiomics features from localized regions, achieving a Macro AUC of 0.839 across 8 pathologies.

Research that utilizes ViT self-attention for weakly-supervised localization and explainability is limited. RGT extracts attention maps for pathology regions (Han et al., 2023), while Wollek et al. (2023) found ViT attention rated higher than Grad-CAM by radiologists for pneumothorax classification explainability. However, most ViT-based methods rely on gradient-based CAM like Grad-CAM (Öztürk et al., 2025; Fu et al., 2025) or gradient-free variants with pooling (Gu et al., 2022). Sparse attention maps and high data requirements limit the direct utility of standalone ViTs, especially without anatomical constraints (Qiu et al., 2024).

2.4. Addressing biases, shortcut learning, and explainability in chest X-ray classification models

CXR datasets suffer from biases due to NLP-extracted labels and demographic underrepresentation, leading to underdiagnosis and ethical issues (Seyyed-Kalantari et al., 2021; Rafferty et al., 2025). Shortcut learning exacerbates this, with models exploiting spurious elements for poor generalization (DeGrave et al., 2021; Ye et al., 2024). XAI techniques like heatmaps detect shortcuts by validating focus on relevant regions.

Anatomical priors mitigate biases by constraining predictions, such as lung region cropping used in Liu et al. (2019a), or region-specific attention proposed in Hossain et al. (2024). Heatmaps enhance explainability but are underutilized in ViTs due to sparsity of attention (Barekatain and Glocker, 2025). These gaps highlight the need for the design of models like CLARiT, proposed in this study, which integrates class-specific tokens, background suppression, and priors for precise and interpretable localization outputs in CADe systems.

3. Methods

3.1. Model architecture

The architectural design of the proposed CLARiT model is shown in Fig. 1. The input chest X-ray image of shape $H_{in} \times W_{in}$ is partitioned into N^2 image patches, which are then projected into P patch tokens. A standard vision transformer often uses a single class token to encode the information of all predicted classes (Dosovitskiy et al., 2021). In CLARiT, however, C individual class tokens are used, where each class token (with an embedding size of D) encodes the information of a single class. Learned positional embeddings are added to the input class and patch tokens. A series of d transformer blocks then process all tokens, resulting in transformed class and patch tokens at the output. An attention pooling layer is applied to the class tokens to yield classification logits $\ell_C \in \mathbb{R}^C$. The output patch tokens are passed through the SegmentCAM module, where foreground classification logits $\ell_f \in \mathbb{R}^C$ and foreground segmentation map $S_f \in \{0, 1\}^{H \times W \times C}$ are produced. Finally, class-specific attention maps $A \in \{0, 1\}^{H \times W \times C}$ are produced by fusing the attention maps from the final p transformer blocks. During inference, the classification output is the average of ℓ_C and ℓ_f .

3.2. Weakly-supervised localization

This section describes our approach to weakly-supervised localization of pathologies in chest X-rays using a vision transformer with multiple class tokens. We first extract class-specific attention maps from the transformer’s self-attention matrix, averaged across the final model layers. These are then fused with foreground masks during inference to produce precise localization heatmaps, constrained to relevant anatomical regions via background suppression.

3.2.1. Class-specific attention maps

A vision transformer with multiple class tokens, where each class token corresponds to a single class, results in class-specific attention maps A . The class-to-patch sub-matrix in the transformer self-attention, as shown in Fig. 2, is extracted to obtain these attention maps. Using the class-to-patch attention, the class-specific attention maps are computed by averaging across the final p transformer layers:

$$A^{(i,j,c)} = \frac{1}{p} \sum_l^p A_l^{(i,j,c)}, \quad A_l \in (0, 1)^{H \times W \times C} \quad (1)$$

where $A^{(i,j,c)}$: class-specific attention map of class c at index (i, j) , $A_l^{(i,j,c)}$: class-specific attention map of class c at layer l and index (i, j) , l : layer index, i : image height index, j : image width index, and c : class index.

3.2.2. Heatmaps

During inference, localization heatmaps $\Xi \in (0, 1)^{H \times W \times C}$ are produced by fusing each class-specific foreground mask and attention map using element-wise multiplication:

$$\Xi^{(i,j,c)} = A^{(i,j,c)} S_f^{(i,j,c)} \quad (2)$$

where $\Xi^{(i,j,c)}$: localization heatmap of class c at index (i, j) , and $S_f^{(i,j,c)}$: foreground mask of class c at index (i, j) .

Each foreground mask ensures that the localization heatmap is constrained to class-specific foreground regions. Additionally, the background suppression methodology (Sec. 3.5.2) enforces a prior over potential locations of classes. The result is a highly precise localization of chest X-ray pathologies, which is seen in Fig. 3 with regards to the large mass in the upper-middle region of the left lung.

3.3. Class tokens

This section details the mechanisms involving class tokens in the CLARiT model, including classification losses, orthogonality regularization, and attention pooling. A further illustration of the proposed model is shown in Fig. 4. The class token logits ℓ_C are used in the classification loss \mathcal{L}_{CLS}^C . The SegmentCAM module produces a combined CAM loss \mathcal{L}_{CAM} . Finally, an orthogonal class token loss \mathcal{L}_{OCT} is applied to the class tokens produced from the final q transformer blocks.

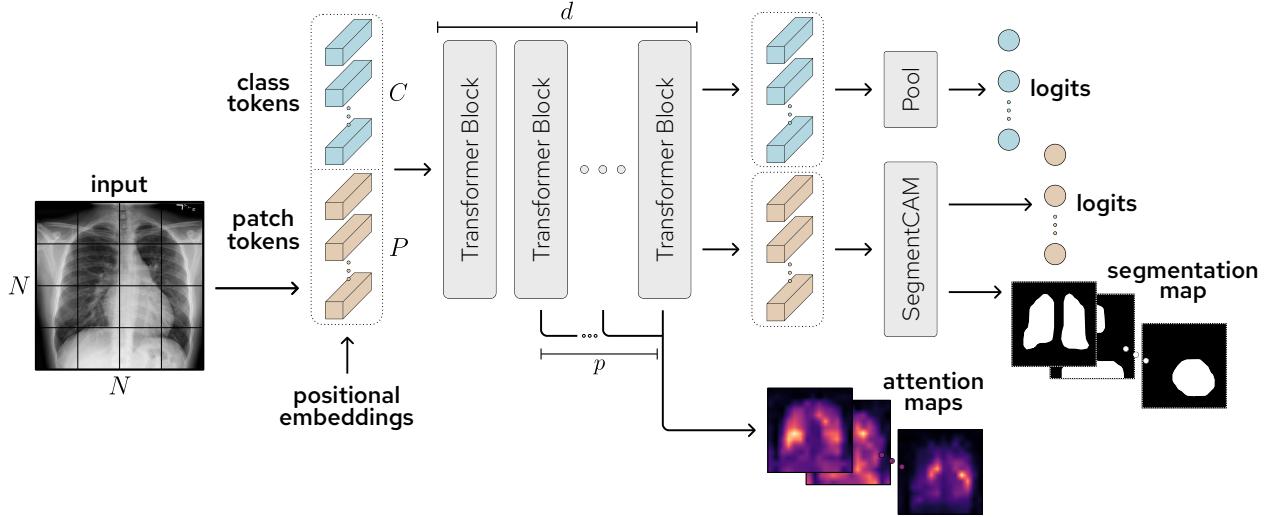


Figure 1: Illustration of the proposed CLARiT model. An input chest X-ray image is split into N^2 patches and embedded into P patch tokens, where they are concatenated with C class tokens. Learned positional embeddings are added to produce $C + P$ input tokens to the transformer. A series of d transformer blocks extract relevant information for classification and weakly-supervised localization. At the output, the C class tokens are passed through an attention pooling module to produce class token logits. The output P patch tokens are passed through the SegmentCAM module, where foreground logits and a segmentation map are produced. The self-attention maps from the final p transformer blocks are fused together to produce class-specific attention maps.

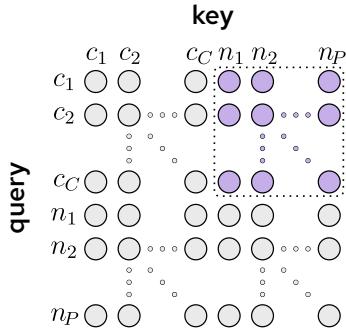


Figure 2: Transformer self-attention matrix with multiple class tokens. Each class token is denoted c_1, c_2, \dots, c_C , and each patch token is denoted n_1, n_2, \dots, n_P . The class-to-patch attention in the upper-right sub-matrix is extracted for the class-specific attention maps.

3.3.1. Classification

The classification loss is the weighted binary cross-entropy (WBCE):

$$\mathcal{L}_{\text{CLS}}(\ell, y) = - \sum_{c=1}^C \left[w_P y^{(c)} \ln \sigma(\ell^{(c)}) + w_N (1 - y^{(c)}) \ln (1 - \sigma(\ell^{(c)})) \right] \quad (3)$$

where \mathcal{L}_{CLS} : classification loss, $\ell \in \mathbb{R}^C$: predicted logits, $y \in \{0, 1\}^C$: ground-truth binary labels, w_P : positive-label weight, w_N : negative-label weight, $y^{(c)}$: label at index c , and $\ell^{(c)}$: predicted logit at index c .

Sigmoid activation is applied to logits for multi-label probabilities:

$$\sigma(x) = \frac{1}{1 + \exp(-x)}, \quad x \in \mathbb{R}. \quad (4)$$

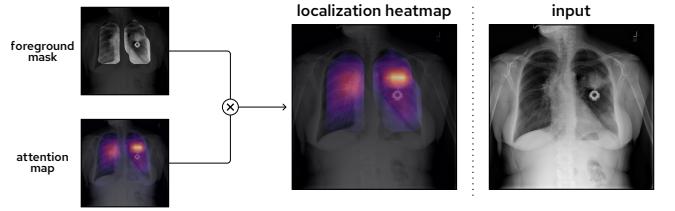


Figure 3: Weakly-supervised localization method of CLARiT. During inference, a foreground mask and attention map are produced for each class. Element-wise multiplication yields a class-specific localization heatmap that is both highly precise and confined to the class’ ground truth region. In this example chest X-ray, a mass (round opacity) is found in the upper-middle left lung of the patient. The attention map has high intensity directly over the mass, and the heatmap intensity is confined to the lung lobes.

Weights w_P and w_N are applied to positive and negative labels respectively, which dynamically handle class imbalances. Weights are calculated over all classes in a single batch:

$$w_P = \frac{|P| + |N|}{|P|}$$

$$w_N = \frac{|P| + |N|}{|N|}$$

where $|P|$: number of positive labels in the batch, and $|N|$: number of negative labels in the batch.

Using Eq. (3), the classification loss applied to the class token logits is

$$\mathcal{L}_{\text{CLS}}^C = \mathcal{L}_{\text{CLS}}(\ell_C, y) \quad (5)$$

where $\mathcal{L}_{\text{CLS}}^C$: class token classification loss.

3.3.2. Class token orthogonality

A regularization loss is applied to the class tokens, which drives them to mutual orthogonality. This affects the class-to-

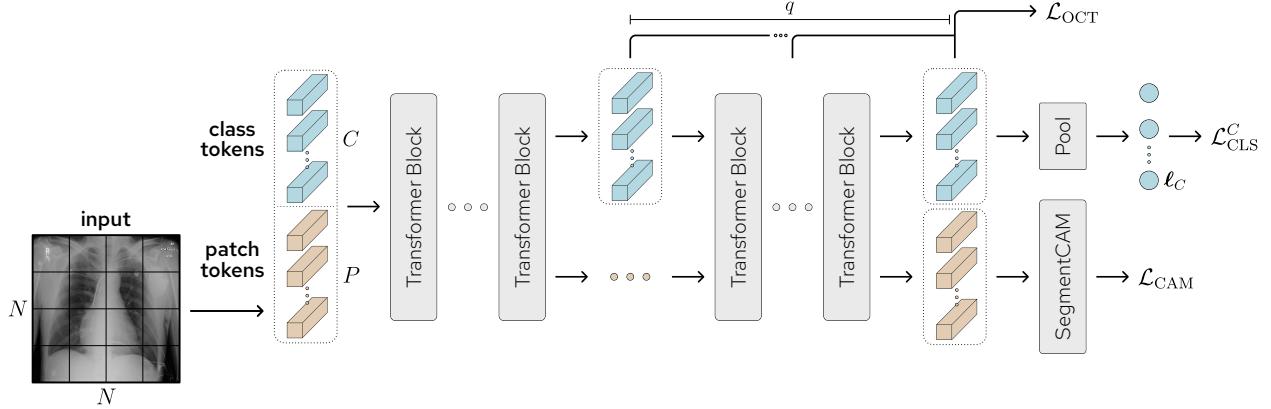


Figure 4: Illustration of the proposed CLARiT model. During training, the final q layers of class tokens are regularized using the orthogonal class token loss \mathcal{L}_{OCT} , which promotes orthogonality between class tokens. Following the attention pooling layer, the class token logits are used to calculate the class token classification loss $\mathcal{L}_{\text{CLS}}^C$. The outputs of the SegmentCAM module are used to calculate the combined CAM loss \mathcal{L}_{CAM} .

patch attention by causing each class token to attend to different image regions. To regularize for orthogonality, the class-to-class cosine similarity for each layer of class tokens is computed using

$$\boldsymbol{\Theta}_l = \frac{\mathbf{T}_l \mathbf{T}_l^\top}{\|\mathbf{T}_l\|_{2,\text{row}} \|\mathbf{T}_l\|_{2,\text{row}}^\top}, \quad \mathbf{T}_l \in \mathbb{R}^{C \times D}, \quad \boldsymbol{\Theta}_l \in [-1, 1]^{C \times C} \quad (6)$$

where $\boldsymbol{\Theta}_l$: cosine similarity matrix of layer l , \mathbf{T}_l : class token matrix of layer l , and $\|\cdot\|_{2,\text{row}}$: row-wise norm.

The diagonal (self-similarity) is masked with

$$\mathbf{M} = \mathbf{1} - \mathbf{I}_C, \quad \mathbf{M} \in \{0, 1\}^{C \times C}$$

where \mathbf{M} : mask matrix, $\mathbf{1}$: ones matrix, \mathbf{I}_C : identity matrix.

The orthogonal class token (OCT) loss is only applied to positive classes, and applied to all layers q :

$$\mathcal{L}_{\text{OCT}} = \frac{1}{(C-1) \left[q \sum_{c=1}^C \mathbf{y}^{(c)} + \varepsilon \right]} \sum_{l=1}^q \sum_{c=1}^C \sum_{k=1}^C \mathbf{M}^{(c,k)} (\boldsymbol{\Theta}^{(l,c,k)} - \theta)^2 \mathbf{y}^{(c)} \quad (7)$$

where \mathcal{L}_{OCT} : orthogonal class token loss, k : class index, $\mathbf{M}^{(c,k)}$: mask value at index (c, k) , $\boldsymbol{\Theta}^{(l,c,k)}$: cosine similarity of layer l at index (c, k) , θ : target cosine similarity ($\theta = 0$), and ε : small value to ensure numerical stability ($\varepsilon = 1 \times 10^{-8}$).

3.3.3. Attention pooling

Typically, global average pooling (GAP) is applied to class tokens to obtain classification logits (Xu et al., 2022). Attention pooling, however, complements the orthogonal class token loss by permitting each dimension of the embedding D to correspond to different classes. Fig. 5 shows the attention pooling module, which contains learned attention weights (invariant to the input image) applied to all class tokens. The softmax function is applied row-wise to linear weights:

$$\text{softmax}(\mathbf{X})^{(c,\varphi)} = \frac{\exp(\mathbf{X}^{(c,\varphi)})}{\sum_{\psi=1}^D \exp(\mathbf{X}^{(c,\psi)})}, \quad \mathbf{X} \in \mathbb{R}^{C \times D} \quad (8)$$

where $\text{softmax}(\mathbf{X})^{(c,\varphi)}$: attention weight at index (c, φ) , \mathbf{X} : learned linear weights, $\mathbf{X}^{(c,\varphi)}$ and $\mathbf{X}^{(c,\psi)}$: linear weight at index (c, φ) and (c, ψ) respectively, φ and ψ : embedding indices.

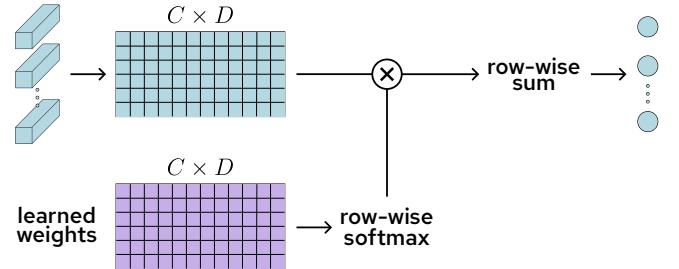


Figure 5: Illustration of the attention pooling module. Row-wise softmax activation is applied to a set of learned weights, which are invariant to the input image. This matrix is element-wise multiplied with the output class tokens after reshaping to a $C \times D$ matrix. Thereafter, a row-wise sum produces class-specific logits. Each dimension in D is able to attend to particular class features, which complements the orthogonal class token loss \mathcal{L}_{OCT} .

Now we can pool each class token into a single logit using:

$$\text{AttnPool}(\mathbf{T}, \mathbf{X})^{(c)} = \sum_{\varphi=1}^D \mathbf{T}^{(c,\varphi)} \text{softmax}(\mathbf{X})^{(c,\varphi)} = \mathbf{t}_C^{(c)} \quad (9)$$

where $\mathbf{T} \in \mathbb{R}^{C \times D}$: output class token matrix, $\mathbf{T}^{(c,\varphi)}$: class token matrix at index (c, φ) , $\text{AttnPool}(\mathbf{T}, \mathbf{X})^{(c)}$: attention pooling result of class token c , and $\mathbf{t}_C^{(c)}$: predicted logit of class token c .

3.4. SegmentCAM feature and heatmap extraction

This section outlines the first stage of our proposed SegmentCAM module (Fig. 6), inspired by Zhai et al. (2023), which involves feature and heatmap extraction from reshaped patch tokens, followed by the application of three specialized loss functions: mask proximity to penalize distant activations, mask confinement to restrict background intensity, and area constraint to limit foreground size. It also covers resampling of heatmaps and segmentation maps for computational efficiency.

In this first stage, the output patch tokens are reshaped to an $N \times N \times D$ tensor and passed through two convolutional

heads. The heads reduce the number of channels from D to the number of classes C . Each head is a fused inverse bottleneck with an expansion ratio of 2 (Fig. 8). The heatmap head uses sigmoid activation (Eq. (4)) to produce a heatmap tensor $\mathbf{H} \in (0, 1)^{N \times N \times C}$. A fixed threshold $t \in (0, 1)$ is applied to \mathbf{H} , which produces the foreground and background segmentation maps S_f and $S_b \in \{0, 1\}^{N \times N \times C}$, respectively. The feature head produces the feature tensor $\mathbf{F} \in \mathbb{R}^{N \times N \times C}$, which is used in the second stage. Finally, three loss functions are applied to \mathbf{H} : the mask proximity loss \mathcal{L}_{PRX} , the mask confinement loss \mathcal{L}_{CNF} , and the area constraint loss \mathcal{L}_{AC} .

3.4.1. Mask proximity

The mask proximity loss penalizes the model for producing heatmap intensity far from the foreground. Heatmap intensity within the foreground has a loss value of zero. The Euclidean distance transform is applied to the negation of the ground-truth segmentation map ($\neg S$), which calculates the distance from each background pixel to its nearest foreground pixel:

$$\mathbf{D} = \text{EDT}(\neg S), \quad S \in \{0, 1\}^{H \times W \times C}, \mathbf{D} \in \mathbb{R}_{\geq 0}^{H \times W \times C}$$

where \mathbf{D} : distance tensor, $\text{EDT}(\cdot)$: Euclidean distance transform function, and S : ground-truth segmentation map.

The distance tensor is then element-wise multiplied by the heatmap tensor and the similarity weight vector $\mathbf{w} \in \mathbb{R}_{\geq 1}^C$:

$$\mathcal{L}_{\text{PRX}} = \frac{8}{HWC(H+W)} \sum_{i=1}^H \sum_{j=1}^W \sum_{c=1}^C \mathbf{H}^{(i,j,c)} \mathbf{D}^{(i,j,c)} \mathbf{w}^{(c)} \quad (10)$$

where \mathcal{L}_{PRX} : mask proximity loss, $\mathbf{H}^{(i,j,c)}$: heatmap intensity at index (i, j, c) , $\mathbf{D}^{(i,j,c)}$: euclidean distance at index (i, j, c) , and $\mathbf{w}^{(c)}$: similarity weight of class c .

Since many chest X-ray pathologies exist in the same region (such as the lung lobes), \mathbf{w} is used to increase learning pressure on less-commonly masked regions. After vectorizing each mask in the ground-truth segmentation map, the cosine similarity is computed between each class' mask:

$$\begin{aligned} \mathbf{U} &= \text{vec}(S), \quad \mathbf{U} \in \{0, 1\}^{(HW) \times C} \\ \mathbf{V} &= \frac{\mathbf{U}^\top \mathbf{U}}{\|\mathbf{U}\|_{2,\text{row}}^\top \|\mathbf{U}\|_{2,\text{row}}}, \quad \mathbf{V} \in [-1, 1]^{C \times C} \end{aligned}$$

where \mathbf{U} : vectorized segmentation map, and \mathbf{V} : cosine similarity matrix.

The average inter-class similarity $\mu \in [-1, 1]^C$ is found by averaging across each row of \mathbf{V} (ignoring the diagonal). This represents the average similarity between each class c and all other classes k ($k \neq c$):

$$\mu^{(c)} = \frac{1}{C-1} \sum_{\substack{k=1 \\ k \neq c}}^C \mathbf{V}^{(c,k)}$$

where $\mu^{(c)}$: average inter-class similarity of class c , and $\mathbf{V}^{(c,k)}$: cosine similarity at index (c, k) .

To transform the vector μ from a similarity metric to weights, the range of values is inverted and the minimum value is set to 1:

$$\mathbf{w} = 2 - \mu - \min_c (\mu^{(c)}) \quad (11)$$

where \mathbf{w} : segmentation similarity weights.

When the mask of class c is very similar to all other masks of classes k , then its weight value will be near 1. For such a value, the segmentation loss for c is unaffected. If, however, the mask of c is highly dissimilar to all masks of k , then there is increased loss applied to c as the weight value is greater than 1.

3.4.2. Mask confinement

The mask confinement loss penalizes any heatmap intensity within the background. The model then learns to predict foreground that is confined within the ground-truth masked regions. It is defined as the ratio of heatmap intensity within the background to total heatmap intensity:

$$\mathcal{L}_{\text{CNF}} = \frac{1}{HWC} \sum_{i=1}^H \sum_{j=1}^W \sum_{c=1}^C \frac{\mathbf{H}^{(i,j,c)} \neg S^{(i,j,c)} \mathbf{w}^{(c)}}{\mathbf{H}^{(i,j,c)}} \quad (12)$$

where \mathcal{L}_{CNF} : mask confinement loss, and $S^{(i,j,c)}$: ground-truth segmentation map at index (i, j, c) .

3.4.3. Foreground area

The area constraint loss reduces the total heatmap intensity, which limits the foreground area:

$$\mathcal{L}_{\text{AC}} = \frac{1}{HWC} \sum_{i=1}^H \sum_{j=1}^W \sum_{c=1}^C \mathbf{H}^{(i,j,c)} \quad (13)$$

where \mathcal{L}_{AC} : area constraint loss.

3.4.4. Heatmap and segmentation map resampling

The heatmap tensor \mathbf{H} and ground-truth segmentation map S are both resampled to 64×64 when computing the losses \mathcal{L}_{PRX} , \mathcal{L}_{CNF} , and \mathcal{L}_{AC} . S is downsampled to reduce computational cost, while \mathbf{H} is upsampled to ensure it retains more spatial detail. Resampling is done via bilinear interpolation.

3.5. SegmentCAM background activation suppression

This section details the second stage of the SegmentCAM module (Fig. 7), where background activation suppression is achieved by creating tensors with selectively suppressed features, processing them through a convolutional classification head to derive specialized logits, and applying targeted classification and suppression losses to prioritize foreground-relevant information.

In this second stage, foreground suppression is applied to the feature tensor \mathbf{F} by fusing it with the background segmentation map S_b . Both the suppressed and unsuppressed feature tensors are passed through the same convolutional classification head. This head is comprised of three fused inverse bottleneck blocks, each with an expansion ratio of 2 (Fig. 8). After the classification head, the unsuppressed feature tensor is fused

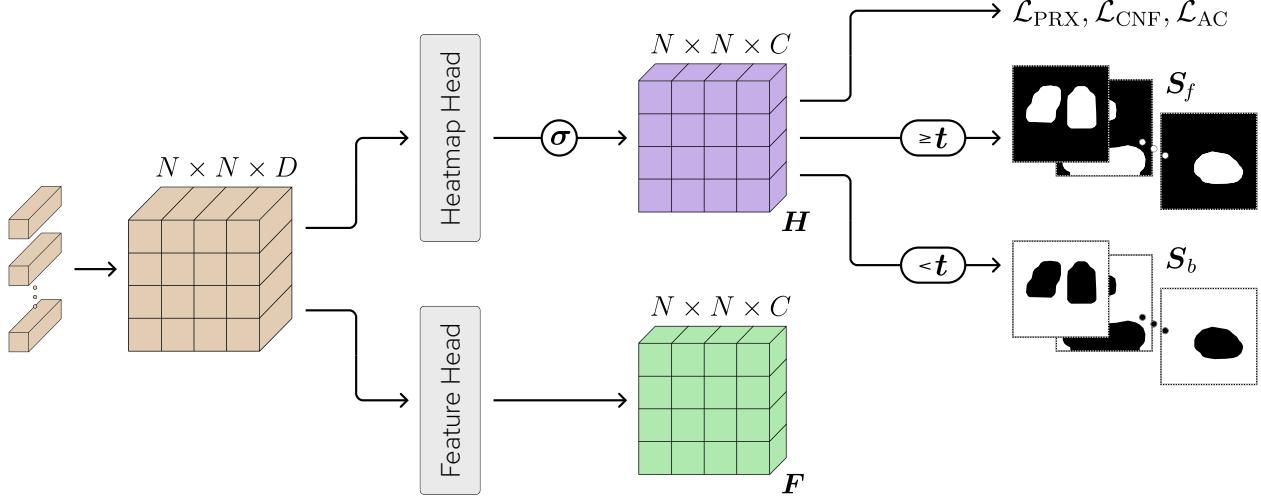


Figure 6: Illustration of the SegmentCAM module’s feature and heatmap extraction. The output patch tokens are first reshaped to an $N \times N \times D$ tensor. Then, in the heatmap branch, a convolutional head with sigmoid activation produces a heatmap tensor H of shape $N \times N \times C$. A threshold t is applied to H to produce foreground and background segmentation maps S_f and S_b respectively. Finally, H is used to calculate the mask proximity loss \mathcal{L}_{PRX} , the mask confinement loss \mathcal{L}_{CNF} , and the area constraint loss \mathcal{L}_{AC} . Then, in the feature branch, a convolutional head produces feature tensor F of shape $N \times N \times C$ which is used in the background activation suppression methodology.

with S_f in a separate branch. Global average pooling is then applied to all final feature tensors, producing classification logits. The foreground logits ℓ_f contain only foreground information, while the background logits ℓ_b contain only background information. The patch token logits ℓ_p contain information from the unmasked features. The classification head learns to only classify foreground-specific features using three loss functions. The foreground classification loss $\mathcal{L}_{\text{CLS}}^f$ is applied to ℓ_f , the patch token classification loss $\mathcal{L}_{\text{CLS}}^P$ is applied to ℓ_p , and the background activation suppression loss \mathcal{L}_{BAS} is applied to both ℓ_p and ℓ_b .

3.5.1. Classification

The foreground and patch token losses both use the weighted binary cross-entropy loss function (Eq. (3)). These losses guide the SegmentCAM module to only identify class-relevant information within foreground regions:

$$\mathcal{L}_{\text{CLS}}^f = \mathcal{L}_{\text{CLS}}(\ell_f, \mathbf{y}) \quad (14)$$

$$\mathcal{L}_{\text{CLS}}^P = \mathcal{L}_{\text{CLS}}(\ell_p, \mathbf{y}) \quad (15)$$

where $\mathcal{L}_{\text{CLS}}^f$: foreground classification loss, and $\mathcal{L}_{\text{CLS}}^P$: patch token classification loss.

3.5.2. Background activation suppression

The background activation suppression loss is designed to reduce the amount of information retained in the background logits relative to the patch token logits. Commonly, ReLU activation is applied to these logits to obtain non-zero activations (Zhai et al., 2023). To prevent numerical instability, softplus activation is used instead:

$$\text{softplus}(x) = \ln(1 + \exp(x)). \quad (16)$$

Applying only to positive classes, the background activation

suppression loss is

$$\mathcal{L}_{\text{BAS}} = \frac{1}{\max(\sum_{c=1}^C \mathbf{y}^{(c)}, 1)} \sum_{c=1}^C \mathbf{y}^{(c)} \frac{\text{softplus}(\ell_b^{(c)})}{\text{softplus}(\ell_p^{(c)}) + \varepsilon} \quad (17)$$

where \mathcal{L}_{BAS} : background activation suppression loss, $\ell_b^{(c)}$: background logit of class c , and $\ell_p^{(c)}$: patch token logit of class c .

3.6. Combined loss

All component loss functions are combined into a single loss using scalar weights α_C , α_P , α_f , β_1 , β_2 , γ_1 , γ_2 and δ :

$$\mathcal{L} = \alpha_C \mathcal{L}_{\text{CLS}}^C + \mathcal{L}_{\text{CAM}} + \delta \mathcal{L}_{\text{OCT}} \quad (18)$$

where \mathcal{L} : combined loss, and the CAM loss (\mathcal{L}_{CAM}) is defined as:

$$\mathcal{L}_{\text{CAM}} = \alpha_P \mathcal{L}_{\text{CLS}}^P + \alpha_f \mathcal{L}_{\text{CLS}}^f + \beta_1 \mathcal{L}_{\text{BAS}} + \beta_2 \mathcal{L}_{\text{AC}} + \gamma_1 \mathcal{L}_{\text{PRX}} + \gamma_2 \mathcal{L}_{\text{CNF}}. \quad (19)$$

3.7. Dataset

The CLARiT model was trained and evaluated on the official benchmark split of the NIH ChestX-ray14 dataset (Wang et al., 2017), which contains 112,120 frontal chest X-ray images of size 1024×1024 from 30,805 patients. There are 14 co-occurring pathology classes, and the official train-test split is 80%–20%. The official training set was further split such that 70% of all images in the dataset were used for training, and 10% used for validation. All splits were done patient-wise, so that all images from the same patient reside in the same dataset split.

A subset of the test set has been annotated with bounding boxes, which denote radiologist-identified pathology regions.

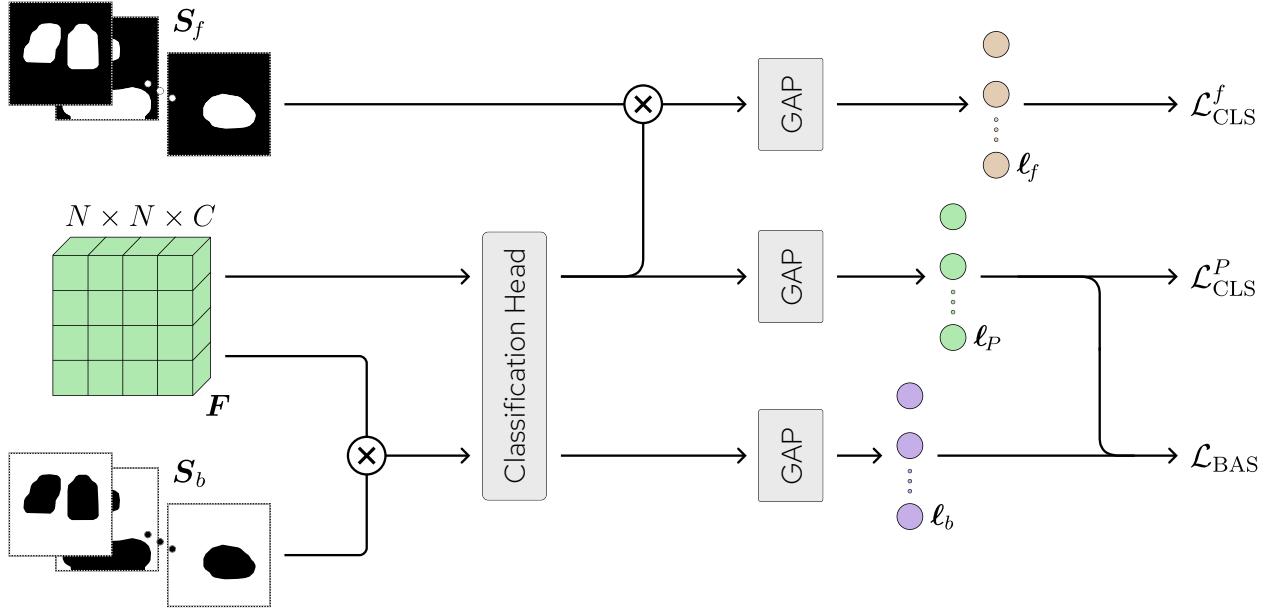


Figure 7: Illustration of the SegmentCAM module’s background activation suppression methodology. The feature tensor \mathbf{F} is passed through the convolutional classification head twice: with unmasked features, and with foreground features suppressed (after fusing with S_b). In the top branch, the output of the unmasked features has its background features suppressed with S_f , and global average pooling (GAP) is applied to produce foreground logits ℓ_f . In the middle branch, the output unmasked features are used to produce patch token logits ℓ_P using GAP. In the bottom branch, the output of the foreground-suppressed features produce background logits ℓ_b using GAP. Foreground and patch token losses $\mathcal{L}_{\text{CLS}}^f$ and $\mathcal{L}_{\text{CLS}}^P$ are calculated using ℓ_f and ℓ_P respectively. Finally, the background activation suppression loss \mathcal{L}_{BAS} is calculated using ℓ_P and ℓ_b . Using these losses, the classification head learns to classify \mathbf{F} while ignoring background features.

This subset includes 880 images and 984 bounding boxes. Bounding box annotations are only available for 8 of the 14 classes. These images were used to create a validation-test split of 50%–50% (patient-wise) for weakly-supervised localization.

3.8. Metrics

This section outlines the metrics used to evaluate model performance in classification and weakly-supervised localization tasks. These include the area under the receiver operating characteristic curve (AUC) for classification; Intersection over Union (IoU) Accuracy and the Multi-Scale Localization Index (MSLI) for localization; and the Unified Detection Proficiency (UDP) for joint assessment. Finally, comparisons with reproduced models highlight dataset split impacts.

3.8.1. Classification

The AUC is the primary metric for evaluating the classification performance for each class. It is macro-averaged across all classes to obtain the Macro AUC.

3.8.2. Weakly-supervised localization

The IoU Accuracy is the primary metric for evaluating weakly-supervised localization performance. It is calculated using the IoU between pairs of predicted and ground-truth bounding boxes. If $\text{IoU} > \text{T(IoU)}$ for a given bounding box pair, where $\text{T(IoU)} \in (0, 1)$ is a threshold, then this is considered a successful localization. The proportion of ground-truth bounding boxes which are successfully localized is named the IoU

Accuracy at the specified T(IoU). This metric can be macro-averaged across all classes to obtain the Macro IoU Accuracy at the specified T(IoU).

The Multi-Scale Localization Index (MSLI) is calculated by averaging the Macro IoU Accuracy over all T(IoU) in {0.1, 0.25, 0.5, 0.75, 0.9}. The MSLI evaluates the localization performance over multiple degrees of spatial precision, from low to high IoU tolerance.

The weakly-supervised localization metrics for each class are calculated only on images where that class is correctly predicted by the model (i.e., the model assigns a positive label for that class).

3.8.3. Unified performance

Both the classification and weakly-supervised localization performance must be maximized jointly. Consequently, the Unified Detection Proficiency (UDP) is defined as the average between the Macro AUC and the MSLI. The UDP is optimized in ablation experiments.

3.8.4. Performance comparisons

As mentioned earlier, model performance differs significantly with dataset split (Sec. 2). The highest reported Macro AUC for the NIH dataset in the literature is 0.853, achieved by the MLRFNet model (Li et al., 2023). We reproduced MLRFNet using the official NIH split, resulting in an absolute performance drop of 5.2% (Tab. 4). With our best efforts, while constrained by time, we also reproduced four other models—for which we could not verify that the published results used the official NIH split. We trained these using the official NIH

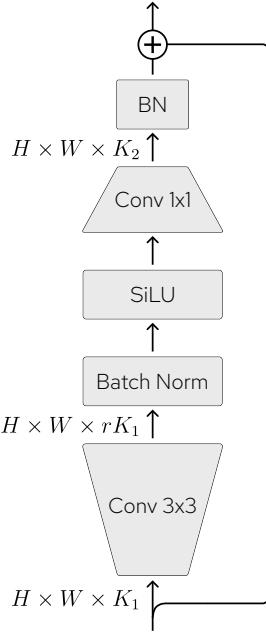


Figure 8: Fused inverse bottleneck block, as used in Xiong et al. (2021). An input tensor of shape $H \times W \times K_1$ is passed through a 3×3 convolution with an expansion ratio of $r \in \mathbb{R}_{>1}$. This increases the number of channels to rK_1 . Thereafter, batch normalization and SiLU activation is applied. Finally, a 1×1 convolution with batch normalization reduces the number of channels to K_2 . If $K_1 = K_2$, then a residual connection is added to the output.

split and tested only their classification performance. The absolute drop in Macro AUC for these models varied between 3.1%–6.3%. To enable comparison of model performance, all results are presented together irrespective of dataset split. However, when it is possible to verify that a model was (or was not) trained using the official NIH split, those models are grouped together.

3.9. Bounding boxes

To obtain bounding boxes from the localization heatmaps, thresholding is used. A class-specific threshold $\xi^{(c)} \in (0, 1)$ is applied to a class’ localization heatmap $\Xi^{(c)}$ which produces connected regions. A single bounding box is created for each connected region. The optimal threshold $\xi^{(c)}$ is found by maximizing the MSLI on the localization-validation set.

3.10. Image resolution and augmentation

Models were trained at an image resolution of 224×224 or 512×512 by downsampling the original 1024×1024 images via the Lanczos method. Augmentations were applied to images during training, namely: random rotation, random cropping, random Poisson noise, and random brightness. All augmentations were generated using a pseudo-random generator, where the seed is unique for each combination of image and epoch number.

3.11. Ground-truth segmentation maps

The segmentation model provided in the TorchXrayVision library was used to produce the ground-truth segmentation

maps (Cohen et al., 2022, 2024). This model produces anatomical segmentations for chest X-ray images. A large language model (Grok) was prompted to list the possible anatomical regions (given by the segmentation model) wherein each NIH pathology could be located (Tab. 1). The anatomical masks for each pathology are combined into a single region-of-interest mask. These region-of-interest masks are then combined into a segmentation map for each image. The anatomical regions predicted by Grok were validated by a diagnostic radiologist. The prompt given to Grok is shown in Appendix A.

Table 1: Anatomical regions of pathologies in the NIH dataset, as given by Grok and validated by a diagnostic radiologist.

Pathology	Anatomical Regions
Atelectasis	Left Lung, Right Lung
Cardiomegaly	Heart
Consolidation	Left Lung, Right Lung
Edema	Left Lung, Right Lung
Effusion	Left Lung, Right Lung
Emphysema	Left Lung, Right Lung
Fibrosis	Left Lung, Right Lung
Hernia	Facies Diaphragmatica, Mediastinum
Infiltration	Left Lung, Right Lung
Mass	Left Lung, Right Lung, Mediastinum
Nodule	Left Lung, Right Lung
Pleural Thickening	Left Lung, Right Lung
Pneumonia	Left Lung, Right Lung
Pneumothorax	Left Lung, Right Lung

3.12. Distillation

The ConvNeXtV2-B is used as the teacher model for distillation. Using ImageNet pretrained weights, it was trained on the training set, and the best performing model was selected during training via the Macro AUC, by evaluating on the validation set after each epoch. The teacher was trained with image augmentations, using the weighted binary cross-entropy loss (Eq. (3)), for 100 epochs, with a batch size of 1024, learning rate of 2.5×10^{-5} , using the AdamW optimizer, weight decay of 0.05, classifier head dropout rate of 0.1 and stochastic depth rate of 0.1. Finally, a cosine learning schedule was used with a warmup of 3 epochs. Two versions were trained at different resolutions: the ConvNeXtV2-B-224 at 224×224 and the ConvNeXtV2-B-512 at 512×512 . Macro AUCs of 0.833 and 0.850 on the validation set were achieved by the ConvNeXtV2-B-224 and ConvNeXtV2-B-512, respectively.

To reduce the training time of the distillation, the TinyViT method was used (Wu et al., 2022). This is an efficient method of distilling a teacher model into a vision transformer. The teacher probabilities are the learning target for classification, which are pre-computed and stored as logits ahead of distillation. Dense logits are stored, as opposed to the sparse logits used in the original TinyViT method. Teacher logits are pre-computed for each combination of epoch and augmented image.

3.13. Model specification

CLARiTy is built on the ViT-S-16 backbone, which has 12 layers, 6 self-attention heads per layer, embedding size of 384

and a patch size of 16. After an ablation study (Sec. 4.5), the final 8 self-attention layers were used to create the class-specific attention maps, and the orthogonal class token loss was applied to the final 8 layers of class tokens.

The CLARiT-S-16-224 configuration takes 224×224 images as input, resulting in a spatial resolution of 14×14 for the output localization heatmaps. The heatmaps are upsampled to 224×224 at inference using bilinear interpolation.

The CLARiT-S-16-512 configuration takes 512×512 images as input, resulting in a spatial resolution of 32×32 for the output localization heatmaps. The heatmaps are likewise upsampled to 512×512 at inference using bilinear interpolation.

3.14. Training

Using DINO pretrained weights, the CLARiT-S-16-224 was trained for 200 epochs, with a batch size of 1024, learning rate of 1×10^{-4} using the AdamW optimizer, weight decay of 0.05, attention pooling dropout rate of 0.1, stochastic depth rate of 0.1, MLP dropout rate of 0.1, and self-attention dropout rate of 0.1. Finally, a cosine learning schedule was used with a warmup of 5 epochs. The best performing model was selected during training via the UDP, by evaluating on the validation set after each epoch.

The CLARiT-S-16-512 was trained using the same methodology. However, instead of using DINO pretrained weights, it used the trained weights of the CLARiT-S-16-224 and interpolated the positional embeddings for 512×512 images.

3.15. Hyperparameter tuning

Hyperparameter tuning used the CLARiT-S-16-224 model to optimize loss weights α_C , α_P , α_f , β_1 , β_2 , γ_1 , γ_2 and δ . We employed a sampler with Optuna’s Tree-structured Parzen Estimator (TPE) algorithm (Akiba et al., 2019, 2025), with 20 trials, 10 startup trials, and 100 candidate samples per trial. Each scalar weight was searched over $[0, 5]$. The optimization objective maximized classification and weakly-supervised localization performance on the validation set using the UDP. The optimal loss weights are given in Tab. 2.

Table 2: Final optimal scalar loss weights selected to maximize validation classification and weakly-supervised localization. These values are used for all experiments.

Loss Weight	Value
α_C	2.2674
α_P	2.4678
α_f	3.9670
β_1	0.4377
β_2	1.4000
γ_1	0.0898
γ_2	2.1624
δ	2.6375

4. Results

This section presents the experimental results of the proposed CLARiT model, including quantitative comparisons on

classification and weakly-supervised localization tasks, qualitative visualizations, resource requirements, and ablation studies. Models performing both classification and weakly-supervised localization are compared in Tab. 3. Detailed comparisons of classification and localization performance are provided in Tabs. 4 and 5, respectively. Ablation results are given in Tab. 6.

4.1. CLARiT-S-16-224

The CLARiT-S-16-224 achieves superior weakly-supervised localization performance when compared against the current state of the art, at a resolution of 224×224 . When comparing Macro IoU Accuracy at a T(IoU) in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, the relative improvement over ThoraX-PriorNet-224 is 16.1%, 20.6%, 26.1%, 33.1%, and 50.7%, respectively. CLARiT-S-16-224 is able to retain high localization accuracy even at higher T(IoU) thresholds, such as 0.5, showing its ability to accurately localize pathologies of various clinical presentations. The primary reason for the improved performance is the use of multiple class tokens, which let the transformer allocate separate token representations per class. This produces highly precise and class-specific activations in the transformer attention maps.

Compared with models operating at a higher resolution of 512×512 , CLARiT-S-16-224 remains competitive and even surpasses them in localization accuracy at higher T(IoU) thresholds. Its relative performance improvement against ThoraX-PriorNet-512 is -3.74%, -1.92%, 2.87%, 18.0%, and 39.6% as T(IoU) increases from 0.1 to 0.5. These results show that CLARiT-S-16-224 is able to accurately localize pathologies even at a low resolution.

The CLARiT-S-16-224 achieves typical classification results when compared to prior work that uses the official NIH split, as the Macro AUC of 0.799 falls within the 0.745–0.830 range. It achieves similar classification performance to the ConvNeXtV2-B-224 teacher model, with an absolute drop in Macro AUC of 0.8%.

4.2. CLARiT-S-16-512

The CLARiT-S-16-512 achieves superior weakly-supervised localization performance when compared against the current state of the art, at a resolution of 512×512 . The relative improvement in Macro IoU Accuracy over ThoraX-PriorNet-512 is 1.4%, 9.9%, 20.5%, 40.7%, and 46.5% as T(IoU) increases from 0.1 to 0.5. CLARiT-S-16-512 is able to achieve improved localization performance at higher resolutions, showing that the weakly-supervised localization methodology in CLARiT is adaptable for different precision requirements, and scales with input resolution.

The CLARiT-S-16-512 achieves competitive classification results when compared to prior work that uses the official NIH split, and the Macro AUC of 0.818 falls within the upper end of the 0.745–0.830 range. It achieves similar classification performance to the ConvNeXtV2-B-512 teacher model, with an absolute drop in Macro AUC of 0.7%. The CLARiT-S-16-512 has an absolute improvement in Macro AUC of 1.9% over

Table 3: Classification and weakly-supervised localization performance of different models evaluated on the NIH test set. Classification results are macro averaged across all 14 pathologies, and the localization results are macro averaged across the 8 pathologies with bounding box labels. The highest metric values are shown in bold. The results of cited models are taken from their publications.

Model	Image Size	NIH Training Set Size	Macro AUC	Macro IoU Accuracy at T(IoU)				
				0.1	0.2	0.3	0.4	0.5
RGT ^{1‡}	224	70%	0.839 [†]	0.591	0.424	0.281	0.173	0.090
ThoraX-PriorNet-224 ^{2‡}	224	70%	0.844	0.666	0.509	0.398	0.296	0.201
ThoraX-PriorNet-512 ^{2‡}	512	70%	0.847	0.803	0.626	0.488	0.334	0.217
Wang et al. (2017)	1024	70%	0.745	0.568	0.373	0.221	0.116	0.062
Li et al. (2022)	512	80%	0.829	0.639	0.527	0.397	0.314	0.243
PCAN ³	512	80%	0.830	0.778	0.574	0.364	0.207	0.103
CLARiT-S-16-224	224	70%	0.799	0.773	0.614	0.502	0.394	0.303
CLARiT-S-16-512	512	70%	0.818	0.814	0.688	0.588	0.470	0.318

[†] Classification performance evaluated on 8 pathologies: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax.

[‡] We could not verify that published model results used the official NIH dataset split.

¹ Han et al. (2023) ² Hossain et al. (2024) ³ Zhu et al. (2022)

Table 4: Classification performance comparison across different models evaluated on the NIH test set. The highest values are shown in bold for models that classify all 14 pathologies. The results of cited models are taken from their publications. Shaded results signify models reproduced using the official NIH dataset split.

Model	Pathology AUC														
	Atel.	Card.	Cons.	Edem.	Effu.	Emph.	Fibr.	Hern.	Infi.	Mass	Nodu.	Pleu.	Pn. 1	Pn. 2	Mean
Yao et al. (2017) [‡]	0.772	0.904	0.788	0.882	0.859	0.829	0.767	0.914	0.695	0.792	0.717	0.765	0.713	0.841	0.798
Li et al. (2018) [‡]	0.800	0.870	0.800	0.880	0.870	0.910	0.780	0.770	0.700	0.830	0.750	0.790	0.670	0.870	0.806
Ho and Gwak (2019) [*]	0.795	0.887	0.786	0.892	0.875	0.875	0.756	0.836	0.703	0.835	0.716	0.774	0.742	0.863	0.810
Ho and Gwak (2019)	0.726	0.865	0.717	0.812	0.810	0.740	0.748	0.803	0.682	0.773	0.686	0.730	0.660	0.796	0.753
RGT ^{1‡}	0.800 [†]	0.920 [†]	—	—	0.780 [†]	—	—	—	0.860 [†]	0.880 [†]	0.880 [†]	—	0.790 [†]	0.810 [†]	0.839 [†]
MLRFNet ^{2‡}	0.833	0.915	0.826	0.905	0.884	0.941	0.821	0.963	0.717	0.858	0.799	0.810	0.760	0.900	0.853
MLRFNet	0.756	0.888	0.742	0.840	0.821	0.894	0.816	0.910	0.684	0.808	0.735	0.760	0.711	0.850	0.801
Kufel et al. (2023) [‡]	0.817	0.911	0.815	0.908	0.879	0.935	0.824	0.890	0.716	0.853	0.771	0.812	0.769	0.898	0.843
Kufel et al. (2023)	0.766	0.863	0.736	0.846	0.830	0.926	0.818	0.846	0.699	0.794	0.740	0.776	0.725	0.873	0.803
ThoraX-PriorNet-224 ^{3‡}	0.826	0.906	0.819	0.910	0.884	0.924	0.818	0.919	0.723	0.864	0.780	0.800	0.770	0.880	0.844
ThoraX-PriorNet-512 ^{3‡}	0.827	0.902	0.812	0.908	0.884	0.927	0.826	0.905	0.723	0.867	0.807	0.813	0.764	0.890	0.847
Hung-Nguyen (2024) [*]	0.803	0.906	0.776	0.874	0.851	0.940	0.842	0.915	0.735	0.858	0.790	0.789	0.746	0.887	0.837
Hung-Nguyen (2024)	0.768	0.895	0.744	0.839	0.827	0.902	0.814	0.898	0.700	0.816	0.753	0.768	0.709	0.855	0.806
HydraViT ^{4‡}	0.810	0.904	0.822	0.882	0.878	0.908	0.845	0.908	0.755	0.840	0.800	0.830	0.758	0.876	0.841
HydraViT	0.732	0.876	0.720	0.822	0.804	0.889	0.795	0.844	0.695	0.776	0.690	0.736	0.679	0.837	0.778
Wang et al. (2017)	0.700	0.810	0.703	0.805	0.759	0.833	0.786	0.872	0.661	0.693	0.669	0.684	0.658	0.799	0.745
Yao et al. (2018)	0.733	0.856	0.711	0.806	0.806	0.842	0.743	0.775	0.673	0.777	0.718	0.724	0.684	0.805	0.761
Tang et al. (2018)	0.756	0.887	0.728	0.848	0.819	0.908	0.818	0.875	0.689	0.814	0.755	0.765	0.729	0.850	0.803
Gündel et al. (2019)	0.767	0.883	0.745	0.835	0.828	0.895	0.818	0.896	0.709	0.821	0.758	0.761	0.731	0.846	0.807
SDFN ⁵	0.781	0.885	0.743	0.842	0.832	0.921	0.835	0.911	0.700	0.815	0.765	0.791	0.719	0.866	0.815
Ma et al. (2019)	0.777	0.894	0.750	0.846	0.829	0.908	0.827	0.934	0.696	0.838	0.771	0.779	0.722	0.862	0.817
Thorax-Net ⁶	0.751	0.871	0.742	0.835	0.818	0.843	0.804	0.902	0.682	0.799	0.715	0.746	0.694	0.825	0.788
CRAL ⁷	0.781	0.880	0.754	0.850	0.829	0.908	0.830	0.917	0.702	0.834	0.773	0.778	0.729	0.857	0.816
Chen et al. (2020)	0.786	0.893	0.751	0.850	0.832	0.944	0.834	0.929	0.699	0.840	0.800	0.795	0.739	0.876	0.826
Guan et al. (2021)	0.785	0.899	0.763	0.850	0.835	0.924	0.831	0.922	0.699	0.838	0.775	0.776	0.738	0.871	0.822
Ouyang et al. (2021)	0.770	0.870	0.740	0.840	0.830	0.940	0.830	0.910	0.710	0.830	0.790	0.790	0.720	0.880	0.819
SwinCheX ⁸	0.781	0.875	0.748	0.848	0.824	0.914	0.826	0.855	0.701	0.822	0.780	0.778	0.713	0.871	0.810
Li et al. (2022)	0.797	0.872	0.779	0.858	0.852	0.935	0.825	0.907	0.711	0.843	0.795	0.800	0.742	0.894	0.829
PCAN ⁹	0.791	0.887	0.759	0.854	0.841	0.944	0.819	0.928	0.711	0.839	0.809	0.806	0.746	0.881	0.830
ConvNeXtV2-B-224	0.763	0.894	0.750	0.853	0.827	0.908	0.805	0.901	0.689	0.814	0.751	0.773	0.718	0.857	0.807
ConvNeXtV2-B-512	0.785	0.900	0.755	0.851	0.833	0.940	0.842	0.919	0.701	0.839	0.797	0.780	0.737	0.877	0.825
CLARiT-S-16-224	0.760	0.895	0.755	0.844	0.826	0.887	0.817	0.839	0.692	0.808	0.736	0.765	0.707	0.854	0.799
CLARiT-S-16-512	0.784	0.899	0.759	0.855	0.837	0.939	0.831	0.854	0.700	0.827	0.783	0.779	0.735	0.874	0.818

* Pathologies are: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, and Pneumothorax.

[†] Classification performance evaluated on 8 pathologies: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax.

[‡] We could not verify that published model results used the official NIH dataset split.

¹ Han et al. (2023) ² Li et al. (2023) ³ Hossain et al. (2024) ⁴ Öztürk et al. (2025) ⁵ Liu et al. (2019a) ⁶ Wang et al. (2020) ⁷ Guan and Huang (2020) ⁸ Taslimi et al. (2022) ⁹ Zhu et al. (2022)

Table 5: Weakly-supervised localization performance comparison across different models. Models are evaluated on the localization portion of the NIH dataset. The highest values for each T(IoU) are shown in bold. The results of cited models are taken from their publications.

T(IoU)	Model	Pathology IoU Accuracy								
		Atel.	Card.	Effu.	Infi.	Mass	Nodu.	Pn. 1	Pn. 2	Mean
0.1	RGT ^{1‡}	0.610	0.950	0.650	0.820	0.500	0.130	0.790	0.280	0.591
	ThoraX-PriorNet-224 ^{2‡}	0.628	1.000	0.791	0.862	0.518	0.127	0.825	0.577	0.666
	ThoraX-PriorNet-512 ^{2‡}	0.761	1.000	0.837	0.870	0.741	0.582	0.867	0.763	0.803
	Wang et al. (2017)	0.689	0.938	0.660	0.707	0.400	0.139	0.633	0.378	0.568
	Li et al. (2022)	0.635	1.000	0.748	0.788	0.694	0.070	0.786	0.394	0.639
	PCAN ³	0.839	1.000	0.856	0.935	0.824	0.194	0.900	0.375	0.778
0.2	CLARIty-S-16-224	0.759	1.000	0.653	0.887	0.906	0.483	1.000	0.500	0.774
	CLARIty-S-16-512	0.671	1.000	0.789	0.806	0.871	0.786	1.000	0.590	0.814
	RGT ^{1‡}	0.410	0.910	0.410	0.590	0.260	0.050	0.570	0.190	0.424
	ThoraX-PriorNet-224 ^{2‡}	0.461	1.000	0.621	0.626	0.318	0.013	0.692	0.340	0.509
	ThoraX-PriorNet-512 ^{2‡}	0.567	0.897	0.693	0.724	0.577	0.253	0.692	0.608	0.626
	Wang et al. (2017)	0.472	0.685	0.451	0.478	0.259	0.051	0.350	0.235	0.373
0.3	Li et al. (2022)	0.404	1.000	0.664	0.737	0.429	0.014	0.691	0.277	0.527
	PCAN ³	0.650	0.760	0.575	0.789	0.612	0.190	0.767	0.250	0.574
	CLARIty-S-16-224	0.414	1.000	0.387	0.710	0.719	0.345	1.000	0.342	0.614
	CLARIty-S-16-512	0.541	1.000	0.553	0.597	0.742	0.750	0.938	0.385	0.688
	RGT ^{1‡}	0.280	0.790	0.220	0.380	0.120	0.010	0.410	0.050	0.283
	ThoraX-PriorNet-224 ^{2‡}	0.300	0.986	0.458	0.439	0.212	0.000	0.542	0.247	0.398
0.4	ThoraX-PriorNet-512 ^{2‡}	0.467	0.795	0.490	0.463	0.506	0.165	0.558	0.464	0.488
	Wang et al. (2017)	0.244	0.459	0.301	0.276	0.153	0.038	0.167	0.133	0.221
	Li et al. (2022)	0.205	1.000	0.441	0.525	0.265	0.000	0.548	0.192	0.397
	PCAN ³	0.428	0.336	0.333	0.569	0.482	0.038	0.600	0.125	0.364
	CLARIty-S-16-224	0.264	0.982	0.133	0.516	0.562	0.345	1.000	0.211	0.502
	CLARIty-S-16-512	0.341	1.000	0.263	0.532	0.677	0.679	0.875	0.333	0.588
0.5	RGT ^{1‡}	0.170	0.540	0.130	0.180	0.070	0.010	0.260	0.020	0.173
	ThoraX-PriorNet-224 ^{2‡}	0.172	0.945	0.261	0.309	0.165	0.000	0.392	0.124	0.296
	ThoraX-PriorNet-512 ^{2‡}	0.322	0.616	0.294	0.268	0.377	0.051	0.400	0.340	0.334
	Wang et al. (2017)	0.094	0.281	0.203	0.122	0.071	0.013	0.075	0.071	0.116
	Li et al. (2022)	0.103	0.979	0.273	0.465	0.184	0.000	0.381	0.128	0.314
	PCAN ³	0.228	0.096	0.163	0.366	0.365	0.013	0.325	0.104	0.207
	CLARIty-S-16-224	0.161	0.946	0.040	0.403	0.531	0.276	0.667	0.132	0.394
	CLARIty-S-16-512	0.212	0.952	0.118	0.387	0.581	0.643	0.688	0.179	0.470
	RGT ^{1‡}	0.080	0.320	0.050	0.090	0.050	0.000	0.120	0.010	0.090
	ThoraX-PriorNet-224 ^{2‡}	0.106	0.726	0.124	0.252	0.118	0.000	0.200	0.083	0.201
	ThoraX-PriorNet-512 ^{2‡}	0.172	0.390	0.144	0.179	0.294	0.013	0.317	0.227	0.217
	Wang et al. (2017)	0.050	0.178	0.111	0.065	0.012	0.013	0.033	0.031	0.062
	Li et al. (2022)	0.045	0.873	0.133	0.343	0.123	0.000	0.333	0.096	0.243
	PCAN ³	0.094	0.007	0.105	0.187	0.211	0.000	0.167	0.052	0.103
	CLARIty-S-16-224	0.069	0.839	0.000	0.258	0.406	0.103	0.667	0.079	0.303
	CLARIty-S-16-512	0.106	0.871	0.013	0.242	0.419	0.393	0.375	0.128	0.318

* Pathologies are: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax.

‡ We could not verify that published model results used the official NIH dataset split.

¹ Han et al. (2023) ² Hossain et al. (2024) ³ Zhu et al. (2022)

Table 6: Ablations on the CLARiT-S-16-224. Classification and weakly-supervised localization performance was evaluated on the validation sets. The highest values are shown in bold.

SegmentCAM	$\mathcal{L}_{\text{PROX}}, \mathcal{L}_{\text{CONF}}$	Pretrained Weights		Classification Loss		Class Token Pooling		Class Token Loss		Validation Metric		
		ImageNet	DINO	MLSM	WBCE	GAP	Attention	CCT	OCT	Macro AUC	MSLI	UDP
✓	—	—	✓	—	✓	—	✓	—	✓	0.822	0.282	0.552
—	—	—	✓	—	✓	—	✓	—	✓	0.835	0.350	0.592
✓	✓	✓	—	—	✓	—	✓	—	✓	0.815	0.307	0.561
✓	✓	—	✓	✓	—	—	✓	—	✓	0.825	0.342	0.584
✓	✓	—	✓	—	✓	✓	—	—	✓	0.829	0.294	0.562
✓	✓	—	✓	—	✓	—	✓	✓	—	0.833	0.340	0.586
✓	✓	—	✓	—	✓	—	✓	—	✓	0.830	0.358	0.594

the CLARiT-S-16-224, showing that the model’s classification performance scales with input resolution.

4.3. Qualitative results

Some localization outputs of the CLARiT-S-16-224 model are provided in Fig. 9. The model is able to produce highly precise heatmaps and bounding boxes for pathologies of all sizes. The foreground masks ensure heatmaps and bounding boxes are constrained to clinically relevant regions, improving accuracy.

4.4. Computing resources

The computing resources required for the CLARiT-S-16-224 and CLARiT-S-16-512 models and their components are given in Tab. 7. For the CLARiT-S-16-224 model, the combination of the SegmentCAM and attention pooling modules increases the number of parameters and FLOPS over the vision transformer backbone by 24.7% and 21.3%, respectively. For the CLARiT-S-16-512 model, the increase in parameters and FLOPS is 24.4% and 17.0%, respectively.

Table 7: Computing resources of the CLARiT-S-16-224, CLARiT-S-16-512, and their components.

Image Size	Component	Parameters	FLOPS
224	Vision Transformer Backbone	21.68 M	4.94 G
	Attention Pooling Module	5.38 K	5.38 K
	SegmentCAM Module	5.35 M	1.05 G
	CLARiT-S-16-224	27.03 M	5.99 G
512	Vision Transformer Backbone	21.99 M	32.35 G
	Attention Pooling Module	5.38 K	5.38 K
	SegmentCAM Module	5.35 M	5.49 G
	CLARiT-S-16-512	27.35 M	37.84 G

4.5. Ablation study

This section investigates the contributions of key components in the CLARiT model through ablation studies on the SegmentCAM module, pretrained weights, classification loss, class token pooling, class token regularization, and the number of transformer layers for attention fusion and orthogonality enforcement. All ablations were performed on CLARiT-S-16-224. For each variant, we selected the best configuration according to the validation UDP. The chosen model

is the variant that achieved the highest UDP, and reported test results correspond to this best validation model.

4.5.1. SegmentCAM

Two ablations were performed for the SegmentCAM module. Turning off the module completely (using only transformer self-attention for localization) resulted in a relative drop in UDP of 0.4%, a relative drop in MSLI of 2.2% and a relative increase in Macro AUC of 0.6% (Tab. 6). The SegmentCAM module is important for accurate localization, but this comes at a small cost of classification performance. Since the foreground segmentation map is generated from the low-resolution (14×14) heatmap tensor of the CLARiT-S-16-224, there is imprecise coverage of each foreground mask over pathology-specific regions (Figs. 3 and 9). The small unmasked pathology regions (which would have been masked with a higher resolution heatmap tensor) become suppressed with the background activation suppression methodology, thus resulting in the drop in Macro AUC.

Retaining the module but turning off the mask proximity and mask confinement losses caused significant relative drops in the Macro AUC, MSLI and UDP of 1.0%, 21.2% and 7.1%, respectively. The segmentation map supervision is necessary for the model to learn the correct foreground locations of each pathology.

4.5.2. Pretrained weights

An ablation was performed to investigate the impact of pretrained weights used for the CLARiT-S-16-224. The ImageNet pretrained weights led to relative reductions in Macro AUC (1.8%), MSLI (14.2%) and UDP (5.6%), as shown in Tab. 6. The DINO weights, however, resulted in more discriminative patch embeddings, yielding more precise attention maps and higher classification accuracy.

4.5.3. Classification loss

An ablation experiment tested the commonly-used Multi-Label Soft Margin (MLSM) loss, and it resulted in relative drops in Macro AUC (0.6%), MSLI (4.5%), and UDP (1.7%), compared to the WBCE loss (Tab. 6). The MLSM has the same definition as the WBCE in Eq. (3), except that $w_P = w_N = 1$.

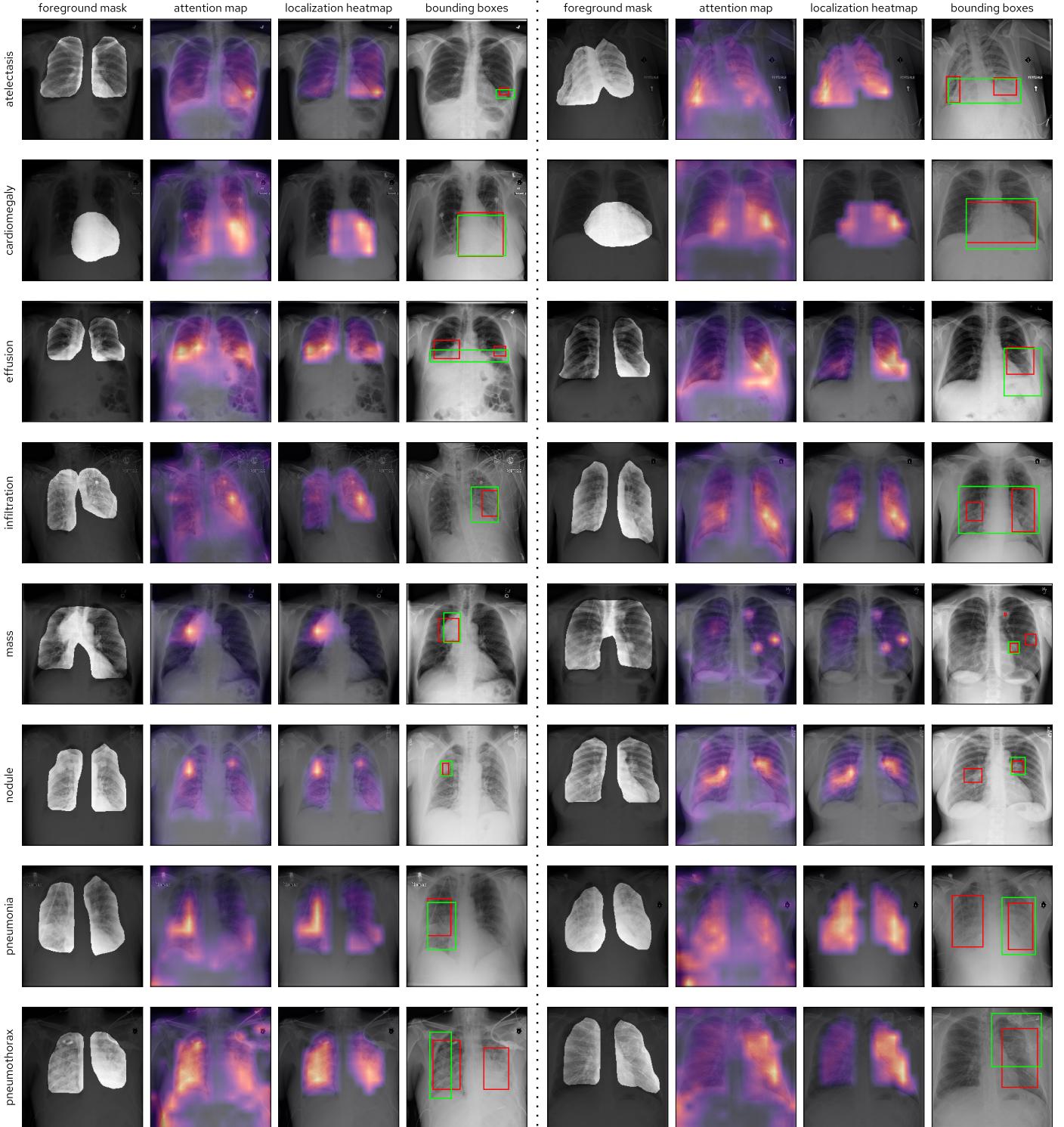


Figure 9: Selected localization outputs from the CLARiT-S-16-224. For each pathology, two chest X-rays at 224×224 are shown. All chest X-rays belong to the localization-test set. The model correctly predicted the presence of the pathology in each image. Class-specific outputs: foreground mask, attention map, localization heatmap, and bounding boxes. Each output is overlaid on the input chest X-ray. Ground-truth bounding boxes: green. Predicted bounding boxes: red.

The WBCE loss is able to better handle rare classes in multi-label chest X-ray classification using its dynamic weighting scheme.

4.5.4. Class token pooling

An ablation was performed for the class token pooling method. The commonly-used GAP pooling yielded very similar Macro AUC compared to attention pooling (0.829 versus 0.830), however the MSLI and UDP had considerable relative drops of 17.9% and 5.4%, respectively (Tab. 6). Attention pooling is superior for weakly-supervised localization because each dimension of a class token can correspond to different pathological features. With GAP pooling, all dimensions must have uniformly high activation values to positively classify a pathology (Wang et al., 2017). For classification, the different kinds of pooling have a small impact.

4.5.5. Class token regularization

An ablation was performed on the class token regularization loss. The contrastive class token (CCT) loss from Xu et al. (2024) resulted in a relative decrease in MSLI (5.0%) and UDP (1.3%), but a relative increase in Macro AUC (0.4%), compared to the OCT loss (Tab. 6). CCT is worse for localization as the cosine similarity between class tokens can be pushed to being negative via cross-entropy, which is less ideal than zero similarity. CCT can promote opposition between the attention maps of different class tokens, while OCT promotes dissimilarity.

4.5.6. Class token regularization

An ablation was performed on the class token regularization loss. The contrastive class token (CCT) loss from Xu et al. (2024) resulted in a relative decrease in MSLI (5.0%) and UDP (1.3%), but a relative increase in Macro AUC (0.4%), compared to the OCT loss (Tab. 6).

To compute the CCT loss, the cross-entropy loss is applied between the class token similarity at each layer (Eq. (6)), and the identity matrix, where softmax is applied row-wise:

$$\mathcal{L}_{\text{CCT}} = \frac{1}{q} \sum_{l=1}^q \text{CrossEntropy}(\boldsymbol{\Theta}_l, \mathbf{I}_C) \quad (20)$$

where CCT is applied to the final $q = 12$ layers of class tokens, and negative classes are masked (Xu et al., 2024).

Although CCT improves classification performance marginally, it degrades localization. This can be attributed to the nature of the contrastive objective: the cross-entropy formulation encourages class token embeddings to have cosine similarity significantly lower than zero, potentially pushing embeddings into opposing hemispheres of the hyperspherical space. In contrast, the OCT loss explicitly targets zero cosine similarity, enforcing orthogonality without inducing active repulsion. Consequently, CCT may over-separate class-specific attention patterns, leading to antagonistic rather than merely dissimilar attention maps, whereas OCT promotes strict dissimilarity while preserving angular equidistance in the embedding space.

4.5.7. Number of transformer layers p and q

An ablation was performed on the number of layers p to fuse for the attention maps, and the number of layers q to apply the orthogonal class token loss. A sweep over all tuples (p, q) where $p \in \{1, 2, \dots, 12\}$ and $q \in \{0, 1, \dots, 12\}$, and evaluating the validation UDP, proved that the best configuration is (8, 8).

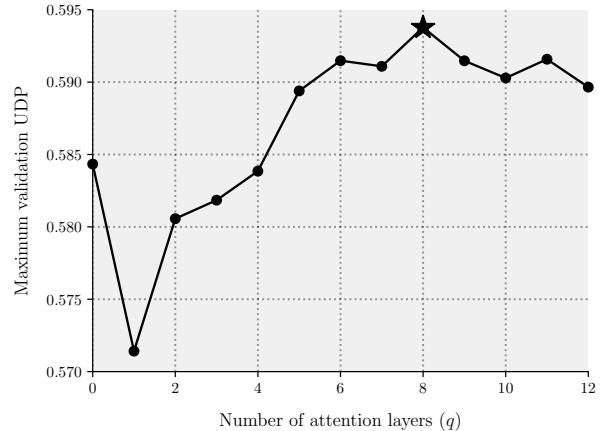


Figure 10: Plot of the number of layers q to apply the orthogonal class token loss, and the impact on the validation UDP (maximized over the number of attention layers p that are fused). The maximum UDP of 0.594 is reached when $q = 8$, and the minimum UDP of 0.571 is reached when $q = 1$.

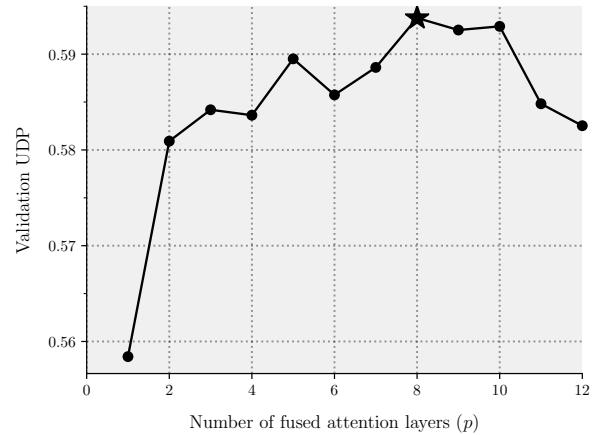


Figure 11: Plot of the number of layers p to fuse together when $q = 8$, and the impact on the validation UDP. The maximum UDP of 0.594 is reached when $p = 8$, and the minimum UDP of 0.558 is reached when $p = 1$.

Fig. 10 shows how the validation UDP (maximized over all p) varies with q . Applying the OCT loss to $q \leq 4$ decreases model performance compared to not applying the loss to any layers. The UDP peaks at $q = 8$ and there is a downward trend for $q \geq 9$. This shows that the early transformer layers tend to extract features common across pathologies, while the later layers tend to extract pathology-specific features.

Fig. 11 shows how the validation UDP varies with p when q is fixed at 8. Generally, fusing together more attention maps leads to better performance, until $p = 8$. When $p > 10$, performance degrades significantly, which shows that the early trans-

former layers have high attention in regions non-specific to individual pathologies.

5. Discussion

CLARiTY addresses key limitations in CXR pathology analysis by integrating multi-label classification with weakly-supervised localization in a unified vision transformer framework. Our results demonstrate superior localization performance, with CLARiTY-S-16-512 achieving a Macro IoU Accuracy of 0.318 at $T(\text{IoU}) = 0.5$, surpassing ThoraX-PriorNet-512 (0.217) and PCAN (0.103) by 46.5% and 209%, respectively, on the NIH test set. This improvement stems from class-specific attention maps fused across transformer layers, which capture pathology-specific features more discriminatively than CNN-based CAM methods (Wang et al., 2017; Li et al., 2022). Notably, gains are pronounced for small pathologies like nodules and masses, where low-resolution heatmaps in prior ViT hybrids (Han et al., 2023) falter. Classification remains competitive, with a Macro AUC of 0.818, close to state-of-the-art (0.847 for ThoraX-PriorNet-512, 0.830 for PCAN), highlighting the model’s balanced proficiency as quantified by UDP (0.594 peak in ablations).

The CLARiTY-S-16-224 variant demonstrates high efficiency at lower input resolutions, achieving a Macro AUC of 0.799 and Macro IoU Accuracy of 0.303 at $T(\text{IoU}) = 0.5$. While scaling to 512 resolution yields modest gains (+2.4% AUC, +5.0% IoU), the 224 model decisively outperforms baselines like ThoraX-PriorNet-224 (0.201 IoU, +50.7% gain) and RGT (0.090 IoU, +236% gain), even for small pathologies such as nodules and masses. This efficiency is potentially advantageous for low-resource settings, enabling deployment on devices with limited computational power without substantial performance trade-offs.

However, fair comparisons are complicated by inconsistencies in dataset splits. Models such as RGT (Han et al., 2023), MLRFNet (Li et al., 2023), ThoraX-PriorNet (Hossain et al., 2024), and HydraViT (Öztürk et al., 2025) could not be verified as using the official NIH patient-wise split, which can yield inflated performance gains—up to 10.8% absolute Macro AUC as shown in related works (Gündel et al., 2019; Wang et al., 2020). This underscores the need for standardized evaluation protocols to ensure reproducibility and generalizability.

The SegmentCAM module in CLARiTY is pivotal, enforcing anatomical priors to suppress background activations and refine foreground masks, yielding a 2.2% MSLI boost in ablations. This could mitigate shortcut learning, a prevalent issue in NLP-labeled datasets (Rafferty et al., 2025), by constraining predictions to clinically plausible regions—e.g., lungs for atelectasis. The orthogonal class token loss further promotes token dissimilarity, improving MSLI by 5.0% over contrastive alternatives, while attention pooling allows dimension-specific feature encoding, outperforming GAP by 17.9% in MSLI. DINO pretraining enhances discriminative embeddings, contributing 14.2% MSLI gain versus ImageNet, aligning with findings on self-supervised benefits for medical tasks (Barekatian and Glocker, 2025).

Compared to related works, CLARiTY advances beyond CNN-ViT hybrids (Li et al., 2022; Öztürk et al., 2025) by leveraging pure ViT self-attention for localization without gradient-based CAM, while reducing noise in heatmaps (Qiu et al., 2024). It also outperforms location-aware and multi-resolution pyramidal methods (Yao et al., 2018; Gündel et al., 2019) in handling complex pathologies via multi-token design. Anatomical priors echo ThoraX-PriorNet (Hossain et al., 2024) but are integrated more seamlessly via SegmentCAM, avoiding separate segmentation networks.

Limitations include reliance on the NIH dataset, which may embed biases from NLP labels and demographic skews (Seyyed-Kalantari et al., 2021). While patient-wise splits mitigate leakage, external validation on datasets like CheXpert or MIMIC-CXR could confirm generalizability. Heatmaps enhance explainability, but clinical validation by radiologists is needed to assess utility in bias detection.

Future work could extend CLARiTY to 3D CT scans or multimodal inputs (such as radiological reports via TieNet Wang et al., 2018), incorporate uncertainty estimation for rare pathologies, or fine-tune on underrepresented demographics to address ethical concerns. Overall, CLARiTY’s label-efficiency and precision position it as a robust tool for automated CXR screening, with potential uses in resource-limited settings.

6. Conclusion

In conclusion, CLARiTY represents a significant advancement in automated CXR analysis, enabling joint multi-label pathology classification and weakly-supervised localization using only image-level labels and anatomical priors. By leveraging multiple class tokens, SegmentCAM for prior-guided background suppression, orthogonal regularization, and attention pooling in a vision transformer, CLARiTY achieves state-of-the-art localization (50.7% relative Macro IoU Accuracy gain over priors) while maintaining strong classification on the NIH benchmark. An ablation study confirms each component’s contributions to improved classification and localization performance. These advances yield interpretable heatmaps that could facilitate the detection of biases and shortcut learning, thereby supporting more reliable computer-aided detection systems in clinical practice. Future extensions to diverse datasets and modalities will further broaden the impact of CLARiTY.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

John M. Statheros: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. **Hairong Bau:** Conceptualization, Methodology, Writing – Review & Editing,

Supervision. **Richard Klein:** Conceptualization, Methodology, Writing – Review & Editing, Supervision.

Data availability

The NIH ChestX-ray14 dataset is available at <https://nihcc.app.box.com/v/ChestXray-NIHCC>.

Computer code

Code for models and scripts will be available upon publication.

Acknowledgments

Funding for infrastructure related to this work was provided by the Canadian Institutes of Health Research (CIHR) under project grant titled “Using locally developed computer-assisted detection to promote social justice for a population with a high burden of lung disease: A participatory equity-sensitive approach” (RN520904 – 506651).

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT and Grok in order to improve clarity and readability of prose. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Prompt given to Grok

I am training a segmentation-aware classification model based on the NIH ChestX-ray14 dataset. The segmentation model produces maps for the following regions: ['Left Clavicle', 'Right Clavicle', 'Left Scapula', 'Right Scapula', 'Left Lung', 'Right Lung', 'Left Hilus Pulmonis', 'Right Hilus Pulmonis', 'Heart', 'Aorta', 'Facies Diaphragmatica', 'Mediastinum', 'Weasand', 'Spine']. The labels for the ChestX-ray14 dataset are: ['Atelectasis', 'Cardiomegaly', 'Consolidation', 'Edema', 'Effusion', 'Emphysema', 'Fibrosis', 'Hernia', 'Infiltration', 'Mass', 'Nodule', 'Pleural Thickening', 'Pneumonia', 'Pneumothorax']. Tell me, which segmentation regions do I need for each pathology? In other words, where would I find these pathologies in a chest X-ray from the regions given to you?

References

Agu, N.N., Wu, J.T., Chao, H., Lourentzou, I., Sharma, A., Moradi, M., Yan, P., Hendler, J., 2021. AnaXNet: Anatomy aware multi-label finding classification in chest X-ray, in: de Bruijne, M., Cattin, P.C., Cotin, S., Paday, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer

International Publishing, Cham. pp. 804–813. URL: http://dx.doi.org/10.1007/978-3-030-87240-3_77, doi:10.1007/978-3-030-87240-3_77.

Ait Nasser, A., Akhloufi, M.A., 2023. A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics* 13, 159. URL: <http://dx.doi.org/10.3390/diagnostics13010159>, doi:10.3390/diagnostics13010159.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 2623—2631. URL: <https://dl.acm.org/doi/10.1145/3292500.3330701>, doi:10.1145/3292500.3330701.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2025. Optuna (Version 4.4.0) [software]. GitHub. URL: <https://github.com/optuna/optuna>.

Alam, M.S., Wang, D., Sowmya, A., 2024. AMFP-net: Adaptive multi-scale feature pyramid network for diagnosis of pneumoconiosis from chest X-ray images. *Artificial Intelligence in Medicine* 154, 102917. URL: <https://www.sciencedirect.com/science/article/pii/S0933365724001593>, doi:<https://doi.org/10.1016/j.artmed.2024.102917>.

Barekatain, L., Glocker, B., 2025. Evaluating the explainability of vision transformers in medical imaging. arXiv preprint. URL: <https://arxiv.org/abs/2510.12021>, doi:10.48550/ARXIV.2510.12021.

Bassi, P.R.A.S., Dertkil, S.S.J., Cavalli, A., 2024. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications* 15. URL: <http://dx.doi.org/10.1038/s41467-023-44371-z>, doi:10.1038/s41467-023-44371-z.

Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M., 2020. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 66, 101797. URL: <http://dx.doi.org/10.1016/j.media.2020.101797>, doi:10.1016/j.media.2020.101797.

Chen, B., Li, J., Lu, G., Yu, H., Zhang, D., 2020. Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification. *IEEE Journal of Biomedical and Health Informatics* 24, 2292–2302. URL: <http://dx.doi.org/10.1109/JBHI.2020.2967084>, doi:10.1109/JBHI.2020.2967084.

Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarnera, M., Lungren, M.P., Chaudhari, A., Brooks, R.,

- Hashir, M., Bertrand, H., 2022. TorchXRayVision: A library of chest X-ray datasets and models. *Medical Imaging with Deep Learning*. URL: <https://arxiv.org/abs/2111.00595>, doi:10.48550/arXiv.2111.00595.
- Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guerrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H., 2024. TorchXRayVision (Version 1.2.4) [software]. GitHub. URL: <https://github.com/mlmed/torchxrayvision>.
- DeGrave, A.J., Janizek, J.D., Lee, S.I., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* 3, 610—619. URL: <http://dx.doi.org/10.1038/s42256-021-00338-7>, doi:10.1038/s42256-021-00338-7.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- Feyisa, D.W., Ayano, Y.M., Debelee, T.G., Schwenker, F., 2023. Weak localization of radiographic manifestations in pulmonary tuberculosis from chest X-ray: A systematic review. *Sensors* 23. URL: <https://www.mdpi.com/1424-8220/23/15/6781>, doi:10.3390/s23156781.
- Fu, M., Tantithamthavorn, C., Le, T., 2025. DAViT: A domain-adapted vision transformer for automated pneumonia detection and explanation using chest X-ray images. *IEEE Access* 13, 103033–103044. URL: <http://dx.doi.org/10.1109/ACCESS.2025.3579314>, doi:10.1109/ACCESS.2025.3579314.
- Giordano, D., Leonardi, R., Maiorana, F., Scarciofalo, G., Spampinato, C., 2007. Epiphysis and metaphysis extraction and classification by adaptive thresholding and DoG filtering for automated skeletal bone age analysis, in: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. p. 6551–6556. URL: <http://dx.doi.org/10.1109/IEMBS.2007.4353861>, doi:10.1109/iembs.2007.4353861.
- Gu, H., Wang, H., Qin, P., Wang, J., 2022. Chest L-Transformer: Local features with position attention for weakly supervised chest radiograph segmentation and classification. *Frontiers in Medicine* 9. URL: <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.923456>, doi:10.3389/fmed.2022.923456.
- Guan, Q., Huang, Y., 2020. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters* 130, 259–266. URL: <https://www.sciencedirect.com/science/article/pii/S0167865518308559>, doi:10.1016/j.patrec.2018.10.027.
- Guan, Q., Huang, Y., Luo, Y., Liu, P., Xu, M., Yang, Y., 2021. Discriminative feature learning for thorax disease classification in chest X-ray images. *IEEE Transactions on Image Processing* 30, 2476–2487. URL: <http://dx.doi.org/10.1109/TIP.2021.3052711>, doi:10.1109/TIP.2021.3052711.
- Gündel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D., 2019. Learning to recognize abnormalities in chest X-rays with location-aware dense networks, in: Vera-Rodriguez, R., Fierrez, J., Morales, A. (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Springer International Publishing, Cham. pp. 757—765. URL: http://dx.doi.org/10.1007/978-3-030-13469-3_88, doi:10.1007/978-3-030-13469-3_88.
- Han, Y., Holste, G., Ding, Y., Tewfik, A., Peng, Y., Wang, Z., 2023. Radiomics-guided global-local transformer for weakly supervised pathology localization in chest X-rays. *IEEE Transactions on Medical Imaging* 42, 750–761. URL: <http://dx.doi.org/10.1109/TMI.2022.3217218>, doi:10.1109/tmi.2022.3217218.
- Hansun, S., Argha, A., Liaw, S.T., Celler, B.G., Marks, G.B., 2023. Machine and deep learning for tuberculosis detection on chest X-rays: Systematic literature review. *Journal of Medical Internet Research* 25. URL: <https://www.sciencedirect.com/science/article/pii/S1438887123005010>, doi:<https://doi.org/10.2196/43154>.
- Ho, T.K.K., Gwak, J., 2019. Multiple feature integration for classification of thoracic disease in chest radiography. *Applied Sciences* 9. URL: <https://www.mdpi.com/2076-3417/9/19/4130>, doi:10.3390/app9194130.
- Hossain, M.I., Zunaed, M., Ahmed, M.K., Hossain, S.M.J., Hasan, A., Hasan, T., 2024. ThoraX-PriorNet: A novel attention-based architecture using anatomical prior probability maps for thoracic disease classification. *IEEE Access* 12, 3256–3273. URL: <http://dx.doi.org/10.1109/ACCESS.2023.3346315>, doi:10.1109/ACCESS.2023.3346315.
- Hung-Nguyen, M., 2024. Patch-level feature selection for thoracic disease classification by chest X-ray images using information bottleneck. *Bioengineering* 11, 316. URL: <http://dx.doi.org/10.3390/bioengineering11040316>, doi:10.3390/bioengineering11040316.
- ILO, 2022. Guidelines for the use of the International Labour Organization (ILO) International Classification of Radiographs of Pneumoconioses: Revised edition 2022. ILO, Geneva. URL: <https://www.ilo.org/media/365991/download>.

- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-IIcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpan-skaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y., 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint. URL: <https://arxiv.org/abs/1901.07031>, doi:10.48550/ARXIV.1901.07031.
- Jin, C., Guo, Z., Lin, Y., Luo, L., Chen, H., 2023. Label-efficient deep learning in medical image analysis: Challenges and future directions. arXiv preprint. URL: <https://arxiv.org/abs/2303.12484>, doi:10.48550/ARXIV.2303.12484.
- Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.D., Lungren, M.P., Stumpe, M.C., Deng, Y., Peng, Y., Lu, Z., Mark, R.G., Snyder, M., Ilse, M., Chute, C.G., Moore, W.J., Rajpurkar, P., 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6, 317. URL: <https://www.nature.com/articles/s41597-019-0322-0>, doi:10.1038/s41597-019-0322-0.
- Kufel, J., Bielówka, M., Rojek, M., Mitręga, A., Lewandowski, P., Cebula, M., Krawczyk, D., Bielówka, M., Kondol, D., Bargieł-Łączek, K., Paszkiewicz, I., Czogalik, Ł., Kaczyńska, D., Woław, A., Gruszczyńska, K., Nawrat, Z., 2023. Multi-label classification of chest X-ray abnormalities using transfer learning techniques. *Journal of Personalized Medicine* 13. URL: <https://www.mdpi.com/2075-4426/13/10/1426>, doi:10.3390/jpm13101426.
- Li, F., Zhou, L., Wang, Y., Chen, C., Yang, S., Shan, F., Liu, L., 2022. Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray. *Quantitative Imaging in Medicine and Surgery* 12. URL: <https://qims.amegroups.org/article/view/92467>, doi:10.21037/qims-21-1117.
- Li, Q., Lai, Y., Adamu, M.J., Qu, L., Nie, J., Nie, W., 2023. Multi-level residual feature fusion network for thoracic disease classification in chest X-ray images. *IEEE Access* 11, 40988–41002. URL: <http://dx.doi.org/10.1109/ACCESS.2023.3269068>, doi:10.1109/ACCESS.2023.3269068.
- Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L., 2018. Thoracic disease identification and localization with limited supervision, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 8290–8299. URL: [http://dx.doi.org/10.1109/cvpr.2018.00865](http://dx.doi.org/10.1109/CVPR.2018.00865), doi:10.1109/cvpr.2018.00865.
- Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., Pu, J., 2019a. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Computerized Medical Imaging and Graphics* 75, 66–73. URL: <https://www.sciencedirect.com/science/article/pii/S0895611118306177>, doi:10.1016/j.compmedimag.2019.05.005.
- Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., Yu, Y., 2019b. Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10631–10640. URL: <http://dx.doi.org/10.1109/ICCV.2019.01073>, doi:10.1109/ICCV.2019.01073.
- Ma, C., Wang, H., Hoi, S.C.H., 2019. Multi-label thoracic disease image classification with cross-attention networks, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer International Publishing, Cham. pp. 730–738. URL: http://dx.doi.org/10.1007/978-3-030-32226-7_81, doi:10.1007/978-3-030-32226-7_81.
- Majkowska, A., Mittal, S., Steiner, D.F., Reicher, J.J., McKinney, S.M., Duggan, G.E., Eswaran, K., Cameron Chen, P.H., Liu, Y., Kalidindi, S.R., Ding, A., Corrado, G.S., Tse, D., Shetty, S., 2020. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294, 421–431. URL: <http://dx.doi.org/10.1148/radiol.2019191293>, doi:10.1148/radiol.2019191293.
- Metlay, J.P., Waterer, G.W., Anzueto, A., Brozek, J., Crothers, K., Cooley, L.A., Dean, N.C., Fine, M.J., Flanders, S.A., Griffin, M.R., Lando, L., Long, A.C., Metersky, M.L., Musher, D.M., Restrepo, M.I., Whitney, C.G., 2019. Diagnosis and treatment of adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. *American Journal of Respiratory and Critical Care Medicine* 200, e45–e67. URL: <https://www.atsjournals.org/doi/full/10.1164/rccm.201908-1581ST>, doi:10.1164/rccm.201908-1581ST.
- Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T.T., Dinh, D.H., Do, C.D., Doan, L.T., Nguyen, C.N., Nguyen, B.T., Nguyen, Q.V., Hoang, A.D., Phan, H.N., Nguyen, A.T., Ho, P.H., Ngo, D.T., Nguyen, N.T., Nguyen, N.T., Dao, M., Vu, V., 2022. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data* 9, 429. URL: <https://www.nature.com/articles/s41597-022-01498-w>, doi:10.1038/s41597-022-01498-w.
- Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., Cheng, J.Z., 2021. Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. *IEEE Transactions on Medical Imaging* 40, 2698–2710. URL: [http://dx.doi.org/10.1109/tmi.2020.3042773](http://dx.doi.org/10.1109/TMI.2020.3042773), doi:10.1109/tmi.2020.3042773.

- Öztürk, S., Turalı, M.Y., Çukur, T., 2025. HydraViT: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images. *Biomedical Signal Processing and Control* 100, 106959. URL: <http://dx.doi.org/10.1016/j.bspc.2024.106959>, doi:10.1016/j.bspc.2024.106959.
- Petrosian, A., Chan, H.P., Helvie, M.A., Goodsitt, M.M., Adler, D.D., 1994. Computer-aided diagnosis in mammography: classification of mass and normal tissue by texture analysis. *Physics in Medicine and Biology* 39, 2273–2288. URL: <http://dx.doi.org/10.1088/0031-9155/39/12/010>, doi:10.1088/0031-9155/39/12/010.
- Pramanik, S., Mudasser, M., Mishra, D., Savadekar, P., 2025. X-Pneumo: A hybrid deep learning framework with explainable visualizations for pneumonia detection in chest X-rays, in: 2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON), pp. 1–6. URL: <http://dx.doi.org/10.1109/NMITCON65824.2025.11187969>, doi:10.1109/NMITCON65824.2025.11187969.
- Qi, B., Zhao, G., Wei, X., Du, C., Pan, C., Yu, Y., Li, J., 2022. GREN: Graph-regularized embedding network for weakly-supervised disease localization in X-ray images. *IEEE Journal of Biomedical and Health Informatics* 26, 5142–5153. URL: <http://dx.doi.org/10.1109/JBHI.2022.3193108>, doi:10.1109/JBHI.2022.3193108.
- Qiu, Z., Rivaz, H., Xiao, Y., 2024. Is visual explanation with grad-cam more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis, in: Cao, X., Xu, X., Rekik, I., Cui, Z., Ouyang, X. (Eds.), *Machine Learning in Medical Imaging*, Springer Nature Switzerland, Cham. pp. 224–233. URL: http://dx.doi.org/10.1007/978-3-031-45676-3_23, doi:10.1007/978-3-031-45676-3_23.
- Rafferty, A., Ramaesh, R., Rajan, A., 2025. Limitations of public chest radiography datasets for artificial intelligence: Label quality, domain shift, bias and evaluation challenges. arXiv preprint. URL: <https://arxiv.org/abs/2509.15107>, doi:10.48550/ARXIV.2509.15107.
- Rahman, T., Khandakar, A., Kadir, M.A., Islam, K.R., Islam, K.F., Mazhar, R., Hamid, T., Islam, M.T., Kashem, S., Mahbub, Z.B., Ayari, M.A., Chowdhury, M.E.H., 2020. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access* 8, 191586–191601. URL: <http://dx.doi.org/10.1109/ACCESS.2020.3031384>, doi:10.1109/ACCESS.2020.3031384.
- Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., Patel, B.N., Yeom, K.W., Shpanskaya, K., Blankenberg, F.G., Seekins, J., Amrhein, T.J., Mong, D.A., Halabi, S.S., Zucker, E.J., Ng, A.Y., Lungren, M.P., 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLOS Medicine* 15, 1–17. URL: <https://doi.org/10.1371/journal.pmed.1002686>, doi:10.1371/journal.pmed.1002686.
- Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q.H., Nguyen, C.D.T., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., Lungren, M.P., Rajpurkar, P., 2022. Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence* 4, 867—878. URL: <http://dx.doi.org/10.1038/s42256-022-00536-x>, doi:10.1038/s42256-022-00536-x.
- Sedai, S., Mahapatra, D., Ge, Z., Chakravorty, R., Garnavi, R., 2018. Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in X-ray images, in: Shi, Y., Suk, H.I., Liu, M. (Eds.), *Machine Learning in Medical Imaging*, Springer International Publishing, Cham. pp. 267–275. URL: http://dx.doi.org/10.1007/978-3-030-00919-9_31, doi:10.1007/978-3-030-00919-9_31.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. URL: <http://dx.doi.org/10.1109/ICCV.2017.74>, doi:10.1109/ICCV.2017.74.
- Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M., 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* 27, 2176—2182. URL: <http://dx.doi.org/10.1038/s41591-021-01595-0>, doi:10.1038/s41591-021-01595-0.
- Sun, W., Wu, D., Luo, Y., Liu, L., Zhang, H., Wu, S., Zhang, Y., Wang, C., Zheng, H., Shen, J., Luo, C., 2022. A fully deep learning paradigm for pneumoconiosis staging on chest radiographs. *IEEE Journal of Biomedical and Health Informatics* 26, 5154–5164. URL: <http://dx.doi.org/10.1109/JBHI.2022.3190923>, doi:10.1109/JBHI.2022.3190923.
- Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M., 2018. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs, in: Shi, Y., Suk, H.I., Liu, M. (Eds.), *Machine Learning in Medical Imaging*, Springer International Publishing, Cham. pp. 249–258. URL: http://dx.doi.org/10.1007/978-3-030-00919-9_29, doi:10.1007/978-3-030-00919-9_29.
- Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., Rohban, M.H., 2022. SwinCheX: Multi-label classification on chest X-ray images with transformers. arXiv preprint. URL: <https://arxiv.org/abs/2206.04246>, doi:10.48550/ARXIV.2206.04246.

- UNSCEAR, 2010. Sources and Effects of Ionizing Radiation: United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008 Report to the General Assembly, with Scientific Annexes. volume 1. United Nations, New York. URL: https://www.unscear.org/unscear/uploads/documents/publications/UNSCEAR_2008_Annex-A-CORR.pdf.
- Viniavskyi, O., Dobko, M., Dobosevych, O., 2020. Weakly-supervised segmentation for disease localization in chest X-ray images, in: Michałowski, M., Moskowitch, R. (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham. pp. 249–259. URL: http://dx.doi.org/10.1007/978-3-030-59137-3_23, doi:[10.1007/978-3-030-59137-3_23](https://doi.org/10.1007/978-3-030-59137-3_23).
- Wang, H., Jia, H., Lu, L., Xia, Y., 2020. Thorax-Net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography. IEEE Journal of Biomedical and Health Informatics 24, 475–485. URL: <http://dx.doi.org/10.1109/JBHI.2019.2928369>, doi:[10.1109/JBHI.2019.2928369](https://doi.org/10.1109/JBHI.2019.2928369).
- Wang, T., Huang, K., Xu, M., Huang, J., 2024. Weakly supervised chest X-ray abnormality localization with non-linear modulation and foreground control. Scientific Reports 14. URL: <http://dx.doi.org/10.1038/s41598-024-79701-8>, doi:[10.1038/s41598-024-79701-8](https://doi.org/10.1038/s41598-024-79701-8).
- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases [dataset], in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 3462–3471. URL: <http://dx.doi.org/10.1109/CVPR.2017.369>, doi:[10.1109/cvpr.2017.369](https://doi.org/10.1109/CVPR.2017.369).
- Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9049–9058. URL: <http://dx.doi.org/10.1109/CVPR.2018.00943>, doi:[10.1109/CVPR.2018.00943](https://doi.org/10.1109/CVPR.2018.00943).
- WHO, 2016. Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches. WHO/HTM/TB/2016.20, World Health Organization, Geneva. URL: <https://www.who.int/publications/i/item/9789241511506>.
- WHO, 2022. WHO guidelines on the use of chest imaging in COVID-19. World Health Organization, Geneva. URL: <https://www.ncbi.nlm.nih.gov/books/NBK586657/>.
- Wollek, A., Graf, R., Čečatka, S., Fink, N., Willem, T., Sabel, B.O., Lasser, T., 2023. Attention-based saliency maps improve interpretability of pneumothorax classification. Radiology: Artificial Intelligence 5, e220187. URL: <https://doi.org/10.1148/ryai.220187>, doi:[10.1148/ryai.220187](https://doi.org/10.1148/ryai.220187), arXiv:<https://doi.org/10.1148/ryai.220187>. PMID: 37035429.
- Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L., 2022. TinyViT: Fast pretraining distillation for small vision transformers, in: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham. pp. 68–85. URL: http://dx.doi.org/10.1007/978-3-031-19803-8_5, doi:[10.1007/978-3-031-19803-8_5](https://doi.org/10.1007/978-3-031-19803-8_5).
- Xiong, Y., Liu, H., Gupta, S., Akin, B., Bender, G., Wang, Y., Kindermans, P.J., Tan, M., Singh, V., Chen, B., 2021. MobileDets: Searching for object detection architectures for mobile accelerators, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3824–3833. URL: <http://dx.doi.org/10.1109/CVPR46437.2021.00382>, doi:[10.1109/CVPR46437.2021.00382](https://doi.org/10.1109/CVPR46437.2021.00382).
- Xu, L., Bennamoun, M., Boussaid, F., Laga, H., Ouyang, W., Xu, D., 2024. MCTformer+: Multi-class token transformer for weakly supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 8380–8395. URL: <http://dx.doi.org/10.1109/TPAMI.2024.3404422>, doi:[10.1109/TPAMI.2024.3404422](https://doi.org/10.1109/TPAMI.2024.3404422).
- Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D., 2022. Multi-class token transformer for weakly supervised semantic segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4300–4309. URL: <http://dx.doi.org/10.1109/CVPR52688.2022.00427>, doi:[10.1109/CVPR52688.2022.00427](https://doi.org/10.1109/CVPR52688.2022.00427).
- Xu, Q., Duan, W., 2024. DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays. Computers in Biology and Medicine 168, 107742. URL: <https://www.sciencedirect.com/science/article/pii/S0010482523012076>, doi:<https://doi.org/10.1016/j.combiomed.2023.107742>.
- Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K., 2017. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint. URL: <https://arxiv.org/abs/1710.10501>, doi:[10.48550/ARXIV.1710.10501](https://doi.org/10.48550/ARXIV.1710.10501).
- Yao, L., Proskey, J., Poblenz, E., Covington, B., Lyman, K., 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv preprint. URL: <https://arxiv.org/abs/1803.07703>, doi:[10.48550/ARXIV.1803.07703](https://doi.org/10.48550/ARXIV.1803.07703).

Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Coffee, M., Zeng, Z., Li, S., Huang, T.H., Wang, Z., Rehg, J.M., Kautz, H., Zhang, A., 2024. The Clever Hans mirage: A comprehensive survey on spurious correlations in machine learning. arXiv preprint. URL: <https://arxiv.org/abs/2402.12715>, doi:10.48550/ARXIV.2402.12715.

Zhai, W., Wu, P., Zhu, K., Cao, Y., Wu, F., Zha, Z.J., 2023. Background activation suppression for weakly supervised object localization and semantic segmentation. International Journal of Computer Vision 132, 750–775. URL: <http://dx.doi.org/10.1007/s11263-023-01919-2>, doi:10.1007/s11263-023-01919-2.

Zhao, G., 2021. Cross chest graph for disease diagnosis with structural relational reasoning, in: Proceedings of the 29th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. pp. 612—620. URL: <https://doi.org/10.1145/3474085.3475221>, doi:10.1145/3474085.3475221.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929. URL: <http://dx.doi.org/10.1109/CVPR.2016.319>, doi:10.1109/CVPR.2016.319.

Zhu, X., Pang, S., Zhang, X., Huang, J., Zhao, L., Tang, K., Feng, Q., 2022. PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization. Computerized Medical Imaging and Graphics 102, 102137. URL: <http://dx.doi.org/10.1016/j.compmedimag.2022.102137>, doi:10.1016/j.compmedimag.2022.102137.

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K., 2021. Deep learning for chest X-ray analysis: A survey. Medical Image Analysis 72, 102125. URL: <https://www.sciencedirect.com/science/article/pii/S1361841521001717>, doi:<https://doi.org/10.1016/j.media.2021.102125>.