# SCALP - Supervised Contrastive Learning for Cardiopulmonary Disease Classification and Localization in Chest X-rays using Patient Metadata

**Ajay Jaiswal**,
UT Austin

**Tianhao Li**,
UT Austin

**Cyprian Zander**,
MIS, Germany

**Yan Han**,
UT Austin

**Justin F. Rousseau**,
UT Austin

**Yifan Peng**,
Weill Cornell Medicine

**Ying Ding**
UT Austin

## Abstract

Computer-aided diagnosis plays a salient role in more accessible and accurate cardiopulmonary diseases classification and localization on chest radiography. Millions of people get affected and die due to these diseases without an accurate and timely diagnosis. Recently proposed contrastive learning heavily relies on data augmentation, especially positive data augmentation. However, generating clinically-accurate data augmentations for medical images is extremely difficult because the common data augmentation methods in computer vision, such as sharp, blur, and crop operations, can severely alter the clinical settings of medical images. In this paper, we proposed a novel and simple data augmentation method based on patient metadata and supervised knowledge to create clinically accurate positive and negative augmentations for chest X-rays. We introduce an end-to-end framework, SCALP, which extends the self-supervised contrastive approach to a supervised setting. Specifically, SCALP pulls together chest X-rays from the same patient (positive keys) and pushes apart chest X-rays from different patients (negative keys). In addition, it uses ResNet-50 along with the triplet-attention mechanism to identify cardiopulmonary diseases, and Grad-CAM++ to highlight the abnormal regions. Our extensive experiments demonstrate that SCALP outperforms existing baselines with significant margins in both classification and localization tasks. Specifically, the average classification AUCs improve from 82.8% (SOTA using DenseNet-121) to 83.9% (SCALP using ResNet-50), while the localization results improve on average by 3.7% over different IoU thresholds.

**Keywords**

Thoracic Disorder; Contrastive Learning; Chest-Xray; Classification; Bounding Box

## I. Introduction

Chest X-rays (CXR) are one of the most common imaging tools used to examine cardiopulmonary diseases. Currently, CXRs diagnosis primarily relies on professional knowledge and meticulous observations of expert radiologists. Automated systems for medical image classification face several challenges. First, it heavily depends on manually-annotated training data which requires highly specialized radiologists to do manual annotation. Radiologists are already overloaded with their diagnosis duties and their hourly charge is costly. Second, common data augmentation methods in computer vision [1], [2] such as crop, mask, blur, and color jitter can significantly alter medical images and generate inaccurate clinical images. Third, unlike images in the general domain, there is subtle variability across medical images. In addition, a significant amount of the variance is localized in small regions. Thus, there is ***an unmet need*** for deep learning models to capture the subtle differences across diseases by attending to discriminative features present in these localized regions.

Recently, contrastive learning frameworks which heavily rely on data augmentation techniques [1]–[3] have become promising due to their ability to capture fine-grained discriminative features in the latent space. This paper proposes a novel and simple data augmentation method based on patient metadata and supervised knowledge such as disease labels to create clinically accurate positive samples for chest X-rays. The supervised classification loss helps SCALP create decision boundaries across different diseases while the patient-based contrastive loss helps SCALP learn discriminative features across different patients. Compared to other baselines [4], [5], SCALP uses simpler ResNet-50 architecture and performs significantly better. SCALP uses GradCAM++ [6] to generate activation maps that indicate the spatial location of the cardiopulmonary diseases. The highlights of our contribution are:

- Our augmentation **technique** for contrastive learning utilizes both patient metadata and supervised disease labels to generate clinically accurate positive and negative keys. *Positive keys* are generated by taking two chest radiographs of the same patient $P$ while *negative keys* are generated using radiographs from patients other than $P$ and having the same disease as $P$.

- A novel unified **framework** to simultaneously improve cardiopulmonary diseases classification and localization. We go beyond the conventional two-staged training (pre-training and fine-tuning) involved in contrastive learning. We demonstrate that single-staged end-to-end supervised contrastive learning can improve existing baselines significantly.

- We propose an innovative rectangular *Bounding Box* generation **algorithm** using pixel-thresholding and dynamic programming.

## II.　Related Work

### A.　Medical Image Diagnosis

In the past decade, machine learning and deep learning have played a vital role in analyzing medical data, primarily medical imaging data. Recognition of anomalies and their localization has been a prevalent task for image analysis. Recent surveys [7], [8] have illustrated the success of CNNs for classification and localization of several diseases in numerous medical imaging datasets varying from chest X-rays, MRIs, and CT scans. With the availability of large public chest X-rays datasets such as [9], [10], many researchers have explored the task of thoracic disease classification [4], [11]–[13]. CheXNet [4] uses 121-layer CNN trained on ChestXray14 [9] for pneumonia detection. [12] and [14] tried to improve the localization results using manually annotated localization data.

### B.　Contrastive Learning for Chest X-rays

In the medical domain, prior work has found the performance improvement on applying contrastive learning to the chest X-rays [15], [16]. [15] presented an adaptation of MoCo [2] for chest X-ray dataset by pre-training it on [17]. The closest recent work to our knowledge [18] proposed to use patient metadata did not utilize supervised label information associated with chest X-rays. Recently, [?] has proposed supervised contrastive learning which extends the self-supervised batch contrastive approach to a fully-supervised setting, allowing us to effectively leverage label information. Inspired by [?] supervised contrastive loss, we propose to generate data augmentation by exploiting patient data and class labels together.

## III.　The Proposed Approach

Given chest X-rays with the cardiopulmonary disease labels, we aim to design a unified model that simultaneously classifies and localizes cardiopulmonary disease. We formulate both tasks in the same prediction framework and train them using a joint contrastive and supervised binary cross-entropy loss. More specifically, each image in our training data is labeled with an 8-dim vector $y = [y_1,...,y_k,...,y_K], y_k \in {0,1}, K = 8$ for each image, and $y_k$ indicates the presence or absence of 8 cardiopulmonary diseases. Our model produces a probability distribution of over 8 diseases for each image in the test set along with a heatmap with the localization information. The heatmap is passed to the BB-generation algorithm to generate rectangular bounding boxes indicating the presence of pathology.

### A.　Image Model

As shown in Figure 1, we use the residual neural network (ResNet-50) architecture considering its manageable training with limited GPU resources and popularity in numerous image classification and object detection challenges. Inspired by Triplet attention [19], we incorporated the lightweight attention module in our ResNet-50 architecture to use cross-dimension interaction between the spatial dimensions and channel dimension for better localization. Image input with shape $h \times w \times c$ produces a feature tensor with shape $h' \times w' \times c'$, where $h, w,$ and $c$ are the height, width, and the number of channels of the input image, respectively while $h' = h/32, w' = w/32,$ and $c' = 2048$. Our framework is composed of two parallel modules: the supervised classification module and the contrastive learning

module. Both modules share the same ResNet-50 encoder, i.e., the same set of parameters for encoding the chest X-ray image inputs.

**1)   Supervised Classification Module::** This module is responsible for learning the high-dimensional decision boundaries across different cardiopulmonary diseases. The encoded input chest X-ray images pass through a global average pooling layer to generate a 2048-dimensional feature vector. The feature vector is fed to a non-linear MLP layer to generate the probability distribution over 8 cardiopulmonary diseases.

We calculate $L_{Cross-Entropy}$ by summing loss from each class.

**2)   Contrastive Learning Module::** In addition to the interclass variance of abnormalities in chest X-rays (i.e., feature differences between different diseases which are captured by classification loss), chest X-rays also have a high intra-class variance (i.e., differences in the X-rays of different patients having the same disease). To capture these intra-class variances, we introduce a supervised contrastive learning module to learn discriminative intra-class features. Our Supervised Patient Metadata based Augmentation module (Section III-C) generates two augmented views $< x'_p, x''_p >$ for each image in the batch. After being encoded by a shared encoder $f(.)$, both views are fed to the global average pooling layer to generate feature embedding $f(x'_p)$ and $f(x''_p)$. These feature embeddings are then transformed through the non-linear projection head similar to [1] to generate $g(f(x'_p))$ and $g(f(x''_p))$. $L_{Contrastive}$ loss is calculated by maximizing the agreement between $g(f(x'_p))$ and $g(f(x''_p))$ [3].

## B.  Triplet Attention

To augment the quality of the localization by exploiting attention from the cross-dimension interaction in feature tensors, we integrate the Triplet Attention [19] into our architecture. Cross-Dimension Interaction involves computing attention weights for each dimension in tensor against every other dimension to capture the spatial and channel attention. In simple terms, spatial attention tells *where the channel to focus on*, while the channel attention tells *what channel to focus on*. With a minimal overload of few learnable parameters, the triplet attention mechanism successfully captures the interaction between the spatial and channel dimension of the input tensor. Following [19], the input tensor with dimension $H \times W \times C$ in SCALP uses a branching mechanism to capture dependencies between $(C,H), (C,W)$, and $(H,W)$.

## C.  Supervised Patient-Metadata Based Augmentation

Recently proposed Supervised Contrastive Learning [?] provides an innovative way to leverage label information for generating positive and negative keys effectively. In our case, we have acutely used patient metadata and label information simultaneously to generate positive and negative keys for the SCALP contrastive learning module. The motivation behind introducing contrastive learning in SCALP is to capture intra-class discriminative features between patients diagnosed with the same disease. Figure 2 provides an overview of our novel technique.

**1)  Positive Sampling:** Data augmentation in medical imaging is sensitive to augmentation techniques such as random crop, Gaussian blur, and color jitter proposed in prior self-supervised learning [1]–[3]. For medical images, such operations may either change the disease label or are not meaningful for grayscale X-ray images. Since our goal is to incorporate discriminative features between two different patients to SCALP, we randomly select two chest X-ray studies having the same disease label of a patient *P* using patient metadata. The first study is called *query*, and the second is called a *positive key*. We use the contrastive loss to maximize agreement between them in latent space.

**2)  Negative Sampling:** As shown in Figure 2, we select k negative keys for each patient *P* in our input batch. *Negative keys* are selected randomly from the pool of chest X-ray studies from patients except *P*, having the same diagnosis as *query*. This helps SCALP to distinguish the subtle differences between patients who are diagnosed with the same disease. Contrastive loss tries to push these *negative keys* away from *query* during training.

## D.  Loss Function

SCALP is trained using the linear combination of supervised classification and contrastive loss. For the supervised classification, we use binary cross-entropy loss. For contrastive learning, we use the extended version of NT-Xent loss [1].

**1)  Supervised classification Loss:** Our disease classification is a multi-class classification problem where we have 8 disease types. Multiple diseases can often be identified in one chest X-ray image and diseases are not mutually exclusive. We, therefore, define 8 binary classifiers for each class/disease type. Since all images in our dataset have 8 labels, the loss function for class k can be expressed as minimizing the binary cross-entry as:

$$L_k = -y_k \cdot log p(k \mid I) - (1 - y_k) \cdot log(1 - p(k \mid I)) \tag{1}$$

where $y_k$ is the ground-truth label of the $k$-th class, and $I$ is the input image. To enable end-to-end training across all the classes, we sum up the class-wise loss to calculate total supervised loss as:

$$L_{Cross-Entropy} = \sum_{n=1}^{batch-size} \sum_{k=1}^{8} L_k^n \tag{2}$$

**2)  Contrastive Loss:** Our contrastive loss extends the normalized temperature scaled cross-entropy loss (NT-Xent). Using our patient-based sampling, we randomly select a batch of $N$ chest X-ray images belonging to $N$ patients. We derive the contrastive loss on the pairs of augmented examples from the batch. Let $I'_{PD}$ be an image in batch belonging to patient $P$ with disease $D$, $sim(x,y)$ denotes similarity between $x$ and $y$, and $f(.)$ denotes ResNet-50 encoder, $g(.)$ denotes the projection head. The loss function $l(I'_{PD})$ for a positive pair of example $< I'_{PD}, I''_{Pd} >$ is defined as:

$$l(I'_{PD}) = -log \frac{e^{(sim(g(f(I'_{PD})), g(f(I''_{PD})))/\tau)}}{\sum_k \unicode{x1D7D9}_{[p \neq P, d = D]} e^{(sim(g(f(I'_{PD})), g(f(I''_{pd})))/\tau)}} \qquad (3)$$

where $\unicode{x1D7D9}_{[p \neq P, d = D]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $p \quad P$ and $d = D$, $\tau$ is the temperature parameter. The final contrastive loss is calculated as the sum over all instances in batch:

$$L_{Contrastive} = \sum_{k=1}^{batch-size} l(I^k_{PD}) \qquad (4)$$

Eventually, we treat SCALP learning as the optimization of both contrastive and supervised cross-entropy loss together. Total loss for SCALP is defined as:

$$L_{Total} = \lambda \times L_{Cross-Entropy} + (1 - \lambda) \times L_{Contrastive} \qquad (5)$$

### E. Bounding Box Generation Algorithm

We propose an innovative and time-efficient approach $\mathcal{O}(n^2)$ to generate regular-shaped rectangular *bounding boxes* on chest X-rays indicating the approximate spatial location of the predicted cardiopulmonary disease. As shown in Figure 1, we feed the k-th layer of our image encoder (ResNet-50) to Gradient-weighted Class Activation Mapping (GradCAM++) [6] to extract the attention maps/heatmaps. Due to the simplicity of intensity distributions in these heatmaps, we first scale heatmaps to the range [0, 255] and apply an ad-hoc threshold to convert these heatmaps into a binary matrix. Pixel values are converted to 1 if its intensity is greater than the threshold and 0 otherwise. Many previous works [9], [23] use only intensity threshold to generate bounding boxes which lead to many false positives. We use dynamic programming to generate a set of k candidate rectangles for bounding boxes and eliminate false positives by selecting the candidate which has the highest average intensity per pixel.

## IV.    Experiments

### A.    Dataset and Preprocessing

NIH Chest X-ray dataset [9] consists of 112,120 chest X-rays collected from 30,805 patients, and each image is labeled with 8 cardiopulmonary disease labels. The NIH dataset also includes high-quality bounding box annotations for 880 images by radiologists. We separate these 880 images from our entire dataset, and they are used only to evaluate disease localization. Our method does not require any training data related to bounding boxes which is a significant difference compared to other existing baseline methods [12], [13] which use some percentage of these images for training. We follow the same protocol as [12], to shuffle our dataset (excluding images with BB annotations) into three subsets: 70% for training, 10% for validation, and 20% for testing. In order to prevent data leakage across patients, we make sure that there is no patient overlap between our subsets.

## B.    Implementation Details

We used the ResNet-50 model with triplet attention and initialized it with pre-trained weights provided by [19]. Our MLP layer is a two-layered fully-connected network with RELU non-linearity, while the projection head is defined similarly to [1]. Both our projection head and MLP layer are randomly initialized. We have used 0.01 learning rate and weight decay of $10^{-6}$ and $10^{-4}$ for contrastive and classification loss, respectively. SCALP uses an SGD optimizer and learning rate scheduler with step size and gamma value of 10 and 0.1, respectively.

**Algorithm 1:**

Bounding Box Generation Algorithm

| | |
|---|---|
| 1 | **Input:** k-th layer attention map/heatmap from ResNet-50 |
| 2 | **Output:** coordinates (x1, y1, x2, y2) of the bounding box |
| 3 | **S**cale heatmap intensities to [0, 255] and create a mask matrix with the same dimension as the heatmap. |
| 4 | **if** *pixel > 180* **then** |
| 5 | L mask[pixel] = 1 |
| 6 | **else** |
| 7 | L mask[pixel] = 0 |
| 8 | **U**sing dynamic programming [24], generate k maximum area rectangles as candidate BB. |
| 9 | **E**xpand candidate rectangles uniformly across the edge till newly added *ratio (0s count, 1s count) > 1* |
| 10 | **S**elect the rectangle with the maximum average pixel intensity mapped in the heatmap and return its coordinates. |

## C.    Disease Identification

SCALP classification is a multi-label classification problem. It assigns one or more labels among 8 cardiopulmonary diseases. We conduct a 3-fold cross-validation (Table I). We compare SCALP with reference models, which have published state-of-the-art performance of disease classification on the NIH dataset. We have used Area under the Receiver Operating Characteristics (AUC) to estimate the performance of our model in Table I. Our results also present the 3-fold cross-validation to show the robustness of our model. Compared to other baselines, SCALP achieves a mean AUROC score of 0.839 using ResNet-50 across the 8 different classes, which is 0.011 higher than the SOTA (uses DenseNet-121) on disease classification.

To understand the importance of contrastive module for disease classification, we trained SCALP with and without the contrastive loss and evaluated performance on the test set of NIH data. Table III presents the significant drop of > 9 % AUC when we exclude the contrastive module from the SCALP pipeline. This demonstrates the importance of our innovative association of contrastive modules with the classification pipeline. Our experiments in Table IV prove our hypothesis that both contrastive and cross-entropy loss is important. In a calculated ratio, they help SCALP learn both disease-level and patient-level discriminative visual features.

### D. Disease Localization

The NIH dataset has 880 images labeled by radiologists with the bounding box information. We have used this dataset to evaluate the performance of SCALP for disease localization. Many prior works [12], [13] have used a fraction of ground truth (GT) bounding boxes for training and evaluated their system on the remaining. To ensure a robust evaluation, we do not use any GT for training, and Table II presents our evaluation results on all 880 images. For localization, we evaluated our detected regular rectangular regions against the annotated ground truth (GT) bounding boxes, using intersection over union ratio (IoU). The localization is defined as correct only if $IoU > T(IoU)$. We evaluate SCALP for different thresholds ranging from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7} as shown in Table II. A higher IoU threshold is preferred for disease localization because clinical usage requires high accuracy. Note that SCALP mean performance for 8 diseases is significantly better than the baseline under all IoU thresholds. When the IoU is set to 0.1, SCALP outperforms the baseline in terms of Cardiomegaly, Infiltration, Mass, and Pneumonia.

Note that our innovative bounding box generation algorithm successfully eliminates dispersed attention, and identifies regions where maximum attention is concentrated. For example, in "Effusion", the generated heatmap has dispersed attention on both sides of the lungs. However, attention intensity is concentrated in the left side of the lung. Our algorithm is able to generate a bounding box on the left side of the lung and have a high overlap with the ground-truth. Similarly, for "Infiltration" and "Nodule", many undesirable patches of attention have been eliminated which is helpful in improving the IoU evaluation of SCALP. The attention maps generated by SCALP are sharp and focused compared to our reference baseline [9]. Overall, our results show that the predicted disease localizations have significant alignment with the ground truth and can serve as interpretable cues for the disease classification.

## V. Conclusion

In this work, we propose a simple and effective end-to-end framework SCALP using supervised contrastive learning to identify cardiopulmonary diseases in chest X-ray. We go beyond two-stage training (pre-training and fine-tuning), and demonstrate that an end-to-end supervised contrastive training using two images from the same patient as a positive pair, can significantly outperform SOTA on disease classification. SCALP can jointly model disease identification and localization using the linear combination of contrastive and classification loss. We also propose a time-efficient Bounding Box generation algorithm that generates bounding boxes from the attention map of SCALP. Our extensive qualitative and quantitative results demonstrate the effectiveness of SCALP and its state-of-the-art performance.

## Acknowledgment

## References

[1]. Chen T, Kornblith S, Norouzi M, and Hinton G, "A simple framework for contrastive learning of visual representations," in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[2]. He K, Fan H, Wu Y, Xie S, and Girshick R, "Momentum contrast for unsupervised visual representation learning," in CVPR, 2020, pp. 9729–9738.

[3]. Chen X, Fan H, Girshick R, and He K, "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:2003.04297, 2020.

[4]. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K et al. , "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[5]. Ma C, Wang H, and Hoi SCH, "Multi-label Thoracic Disease Image Classification with Cross-Attention Networks," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Cham, 2019, vol. 11769, pp. 730–738.

[6]. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, and Batra D, "Grad-cam: Why did you say that?" arXiv preprint arXiv:1611.07450, 2016.

[7]. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, and Khan MK, "Medical image analysis using convolutional neural networks: A review," Journal of Medical Systems, vol. 42, no. 11, 2018.

[8]. Rezaei M, Yang H, and Meinel C, "Deep learning for medical image analysis," arXiv preprint arXiv:1708.08987, 2017.

[9]. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, and Summers RM, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in CVPR, 2017, pp. 2097–2106.

[10]. Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng C.-y., Peng Y, Lu Z, Mark RG, Berkowitz SJ, and Horng S, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," arXiv preprint arXiv:1901.07042, 2019.

[11]. Seyyed-Kalantari L, Liu G, McDermott M, and Ghassemi M, "Chexclusion: Fairness gaps in deep chest x-ray classifiers," arXiv preprint arXiv:2003.00827, 2020.

[12]. Li Z, Wang C, Han M, Xue Y, Wei W, Li L-J, and Fei-Fei L, "Thoracic disease identification and localization with limited supervision," in CVPR, 2018, pp. 8290–8299.

[13]. Liu J, Zhao G, Fei Y, Zhang M, Wang Y, and Yu Y, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," in IC/CV, October 2019.

[14]. Chaitanya K, Erdil E, Karani N, and Konukoglu E, "Contrastive learning of global and local features for medical image segmentation with limited annotations," arXiv preprint arXiv:2006.10511, 2020.

[15]. Sowrirajan H, Yang J, Ng AY, and Rajpurkar P, "Moco pretraining improves representation and transferability of chest x-ray models," in Medical Imaging with Deep Learning. PMLR, 2021, pp. 728–744.

[16]. Sriram A, Muckley M, Sinha K, Shamout F, Pineau J, Geras KJ, Azour L, Aphinyanaphongs Y, Yakubova N, and Moore W, "COVID-19 Prognosis via Self-Supervised Representation Learning and MultiImage Prediction," arXiv:2101.04909 [cs], Jan. 2021.

[17]. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, and Marklund H, "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 590–597.

[18]. Vu YNT, Wang R, Balachandar N, Liu C, Ng AY, and Rajpurkar P, "Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation," in Proceedings of the 6th Machine Learning for Healthcare Conference, ser. Proceedings of Machine Learning Research, vol. 149. PMLR, 06–07 Aug 2021, pp. 755–769.

[19]. Misra D, Nalamada T, Arasanipalai AU, and Hou Q, "Rotate to attend: Convolutional triplet attention module," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3139–3148.

[20]. Wang X, Peng Y, Lu L, Lu Z, and Summers RM, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in CVPR, 2018, pp. 9049–9058.

[21]. Yao L, Poblenz E, Dagunts D, Covington B, Bernard D, and Lyman K, "Learning to diagnose from scratch by exploiting dependencies among labels," arXiv preprint arXiv:1710.10501, 2017.

[22]. Kumar P, Grewal M, and Srivastava MM, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in International Conference Image Analysis and Recognition. Springer, 2018, pp. 546–552.

[23]. Han Y, Chen C, Tang L, Lin M, Jaiswal A, Wang S, Tewfik A, Shih G, Ding Y, and Peng Y, "Using Radiomics as Prior Knowledge for Thorax Disease Classification and Localization in Chest X-rays," AMIA … Annual Symposium proceedings. AMIA Symposium, vol. 2021, pp. 546–555, 2021.

[24]. GeeksforGeeks, "Maximum size Rectangle in Binary Matrix," https://www.geeksforgeeks.org/maximum-size-rectangle-binary-submatrix-1s/
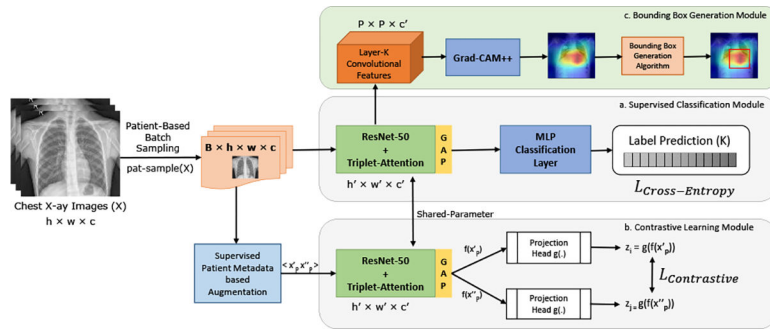
**Fig. 1.**

Model Overview. The input images are sampled in batches with a constraint that no two images in a batch are from the same patient. Learning is performed using a shared encoder (Resent-50) with triplet attention and joint loss from the supervised classification and contrastive learning module.
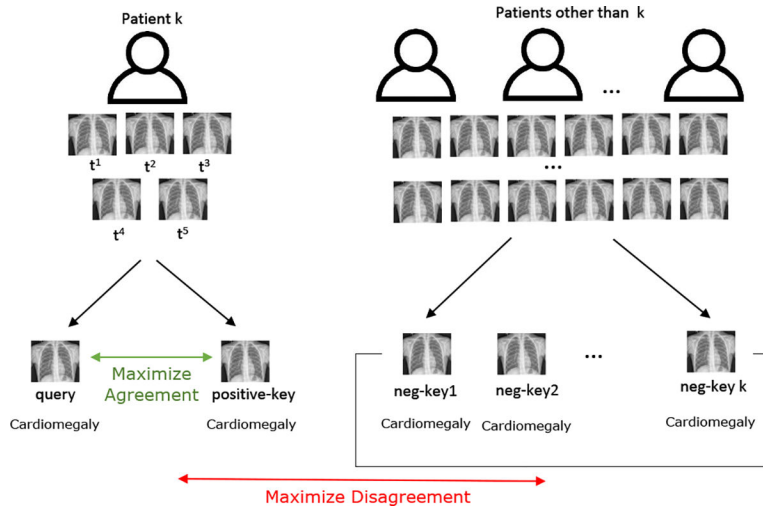
**Fig. 2.**
Overview of our Supervised Patient-Metadata Based Contrastive Augmentation. Positive and Negative Keys are constructed using Patient Metadata (association of patient id with chest X-rays) and Supervised disease labels of chest X-rays.
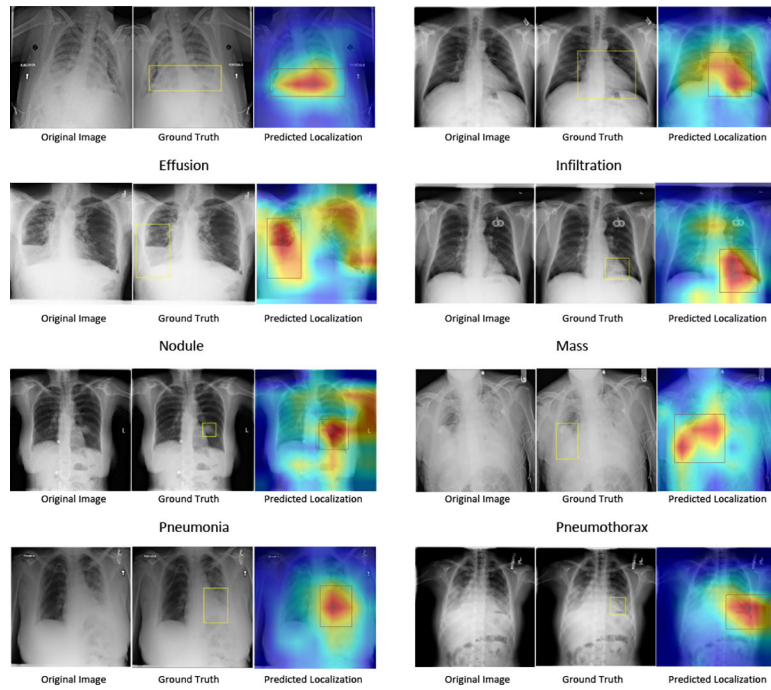
**Fig. 3.**

Examples of visualization of localization on the test images. We plot the results of diseases near the thoracic. The attention maps are generated from the fourth layer of SCALP's encoder and overlapped with its corresponding original radiology image. The ground-truth and the predicted bounding boxes are shown in yellow and red color respectively.
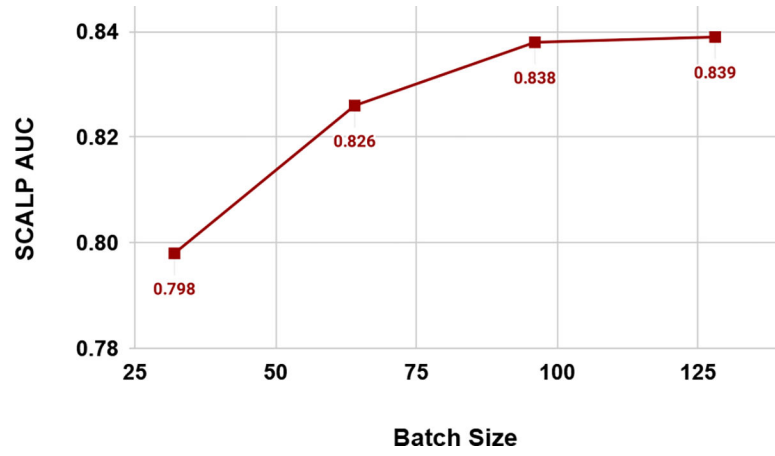
**Fig. 4.**
Effect of Batch Size on SCALP performance for Disease Classification. Prior works [1], [2] have verified that contrastive learning benefits from a larger batch size. SCALP shows a similar trend with increasing batch size.

**TABLE I**

Comparison with the baseline models for AUC of each class and average AUC.

| Method | Atelectasis | Cardiomegaly | Effusion | Infiltration | Mass | Nodule | Pneumonia | Pneumothorax | Mean |
|---|---|---|---|---|---|---|---|---|---|
| *Wang et. al.* [9] | 0.72 | 0.81 | 0.78 | 0.61 | 0.71 | 0.67 | 0.63 | 0.81 | 0.718 |
| *Wang et. al.* [20] | 0.73 | 0.84 | 0.79 | 0.67 | 0.73 | 0.69 | 0.72 | 0.85 | 0.753 |
| *Yao et. al.* [21] | 0.77 | 0.90 | 0.86 | 0.70 | 0.79 | 0.72 | 0.71 | 0.84 | 0.786 |
| *Raj. et. al.* [4] | 0.82 | 0.91 | 0.88 | 0.72 | 0.86 | 0.78 | 0.76 | 0.89 | 0.828 |
| *Kum. et. al.* [22] | 0.76 | 0.91 | 0.86 | 0.69 | 0.75 | 0.67 | 0.72 | 0.86 | 0.778 |
| *Liu et. al.* [13] | 0.79 | 0.87 | 0.88 | 0.69 | 0.81 | 0.73 | 0.75 | 0.89 | 0.801 |
| *Seyed. et. al.* [11] | 0.81 | 0.92 | 0.87 | 0.72 | 0.83 | 0.78 | 0.76 | 0.88 | 0.821 |
| Our model (std) | 0.79±0.01 | 0.92±0.00 | 0.79±0.01 | 0.89±0.01 | 0.88±0.02 | 0.87±0.00 | 0.77±0.01 | 0.81±0.02 | 0.839 |

For each column, red values denote the best results. Note that the best baseline with mean AUC 0.828 uses the DenseNet-121 architecture, while our model is trained using a comparatively simple and lignt-weight ResNet-50 architecture.

**TABLE II**

Disease localization under varying IoU on the NIH Chest X-ray dataset.

| T(IoU) | Model | Atelectasis | Cardiomegaly | Effusion | Infiltration | Mass | Nodule | Pneumonia | Pneumothorax | Mean |
|--------|-------|-------------|--------------|----------|--------------|------|--------|-----------|--------------|------|
| 0.1 | Baseline [9] | 0.69 | 0.94 | 0.66 | 0.71 | 0.40 | 0.14 | 0.63 | 0.38 | 0.569 |
|      | Our Model    | 0.62 | 0.97 | 0.64 | 0.81 | 0.51 | 0.12 | 0.8  | 0.29 | 0.595 |
| 0.2 | Baseline [9] | 0.47 | 0.68 | 0.45 | 0.48 | 0.26 | 0.05 | 0.35 | 0.23 | 0.371 |
|      | Our Model    | 0.42 | 0.92 | 0.42 | 0.6  | 0.25 | 0.04 | 0.56 | 0.18 | 0.434 |
| 0.3 | Baseline [9] | 0.24 | 0.46 | 0.30 | 0.28 | 0.15 | 0.04 | 0.17 | 0.13 | 0.221 |
|      | Our Model    | 0.29 | 0.78 | 0.23 | 0.37 | 0.13 | 0.01 | 0.4  | 0.05 | 0.283 |
| 0.4 | Baseline [9] | 0.09 | 0.28 | 0.20 | 0.12 | 0.07 | 0.01 | 0.08 | 0.07 | 0.115 |
|      | Our Model    | 0.18 | 0.55 | 0.12 | 0.19 | 0.09 | 0.01 | 0.25 | 0.02 | 0.176 |
| 0.5 | Baseline [9] | 0.05 | 0.18 | 0.11 | 0.07 | 0.01 | 0.01 | 0.03 | 0.03 | 0.061 |
|      | Our Model    | 0.07 | 0.33 | 0.04 | 0.10 | 0.04 | 0.0  | 0.14 | 0.10 | 0.102 |
| 0.6 | Baseline [9] | 0.02 | 0.08 | 0.05 | 0.02 | 0.00 | 0.01 | 0.02 | 0.03 | 0.029 |
|      | Our Model    | 0.02 | 0.14 | 0.02 | 0.04 | 0.03 | 0.00 | 0.07 | 0.00 | 0.040 |
| 0.7 | Baseline [9] | 0.01 | 0.03 | 0.02 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.011 |
|      | Our Model    | 0.01 | 0.04 | 0.01 | 0.03 | 0.01 | 0.00 | 0.02 | 0.00 | 0.015 |

Note that since our model doesn't use any ground truth bounding box information, to fairly evaluate the performance of our model, we only consider the previous methods' results with the same settings as SCALP.

**TABLE III**

AUC comparison of SCALP with and without contrastive learning module.

|  | SCALP w/o Contrastive | SCALP |
| --- | --- | --- |
| Atelectasis | 0.751 | 0.79 |
| Cardiomegaly | 0.850 | 0.92 |
| Effusion | 0.833 | 0.79 |
| Infiltration | 0.670 | 0.89 |
| Mass | 0.694 | 0.88 |
| Nodule | 0.640 | 0.87 |
| Pneumonia | 0.700 | 0.77 |
| Pneumothorax | 0.792 | 0.81 |
| **Mean** | 0.7413 | 0.839 |

**TABLE IV**

AUC comparison of SCALP for varying λ in Equation 5.

| λ | 0.99 | 0.90 | 0.85 | 0.80 | 0.75 | 0.70 |
|---|------|------|------|------|------|------|
| **AUC** | 0.766 | 0.794 | 0.822 | **0.839** | 0.818 | 0.785 |

A higher value of λ implies lower weight to the contrastive loss. Scalp achieves the best performance when 80% weight is given to classification loss and 20% weight is given to contrastive loss.