Self-Supervision and Weak Supervision for Accurate and Interpretable Chest X-Ray Classification Models

Abhiroop Talasila, Akshaya Karthikeyan, Shanmukh Alle, Maitreya Maity, U. Deva Priyakumar
Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology Hyderabad, India
Email: deva@iiit.ac.in

***Abstract*—X-rays diagnose numerous thoracic diseases; however, accurate pathology detection requires trained radiologists. The number of available experts may impede population-level patient care, delaying medical action. State-of-the-art (SOTA) machine learning methods categorize chest X-rays across numerous diseases well but do not always account for explainability. Interpretability assessments rarely focus on the conciseness and anatomical correctness of Class Activation Mapping outcomes. These models are not accurate or dependable when tested on different datasets of the same modality. They are not used because their predicted performance is not explainable. This work introduces a self-supervised and weakly supervised pre-training pipeline with an auxiliary loss and supervised fine-tuning that retains performance across datasets. We use the Chest X-ray14 (NIH CXR) dataset for pre-training and CheXpert for fine-tuning. Our model is evaluated on chest X-ray classification tasks but is relevant to other imaging modalities and workloads. Our model outperforms supervised SOTA models on the downstream dataset in categorizing chest X-rays across 14 findings, improving Intersection over Union by 31% on the NIH CXR dataset. Our trained model offers compact, accurate, and interpretable pretrained representations to highlight anatomical sites with limited bounding box annotations.**

I. Introduction

Various pulmonary diseases are commonplace, so early diagnosis is essential for timely medical care and a better prognosis. However, this is a challenge for healthcare systems due to the lack of population-scale medical resources [1]. Chest X-rays (CXRs) play a pivotal role in diagnosing pulmonary conditions, and analyzing them can reveal crucial information about the location and severity of the pathology. Machine Learning (ML) has shown promise in accelerating the decision-making process for healthcare professionals [2]–[5]. Multiple studies have used ML methods to classify CXRs based on their conditions. Although many of these studies report the good performance of their models, they do not sufficiently address their explainability. Using the parameters of a model, we can visualize the decision boundaries of traditional ML models such as logistic regression and decision trees [6]–[8].

Deep Learning (DL) models must be employed for more complex tasks involving data from higher dimensions and achieving higher accuracy. One of the major drawbacks of DL models is that they behave like black boxes, as the neurons' weights cannot be directly understood as knowledge. DL models such as Convolutional Neural Networks (CNNs) combined with large-scale datasets have resulted in state-of-the-art (SOTA) systems in many medical imaging tasks. These models are trained in a supervised method, which requires substantial training data. Bounding box annotations for localization tasks are harder to obtain than their counterparts in classification tasks, which only require image-level labels. This issue is exacerbated for medical datasets, requiring trained radiologists for annotations. They must comply with regulatory policies like the Health Insurance Portability and Accountability Act (HIPAA) and the Institutional Review Board (IRB).

Pre-training approaches have been used to tackle the data scarcity problem in tasks by utilizing knowledge from one area to boost performance in another domain. Concurrently, unsupervised and semi-supervised learning methods have also emerged to model data distributions with limited annotations. Unsupervised pre-training has been shown to serve as a regularization method and give rise to better generalization. Recent studies have also revealed the ability of self-supervision and weak supervision to learn high-quality feature representations [9]–[11].

Early work in generative self-supervised learning (SSL) for the medical domain takes inspiration from a context encoder [12] to propose a pre-training task called context restoration [13]. The authors iteratively select two isolated patches, swap them to generate a corrupted version of the original input image, and train a generative model to restore them to their original version. By learning features from unlabeled natural photos and then unlabeled medical images, Azizi et al. [9] have investigated the efficacy of SSL as a pre-training method for medical image classification tasks. Multi-Instance Contrastive Learning (MICLe) is a minor modification of SimCLR [14] that takes advantage of multiple views of a pathology to perform contrastive learning. Their models outperform supervised baselines pretrained on ImageNet using only a limited number of labeled medical images while being robust to distribution shift.

SSL approaches can also be augmented by weak labels, potentially improving representation learning and performing better on downstream tasks. Hu et al. [15] draw inspiration from Chen et al. [13] and use a context encoder with DICOM metadata tags in a weakly-supervised manner to learn ultra-

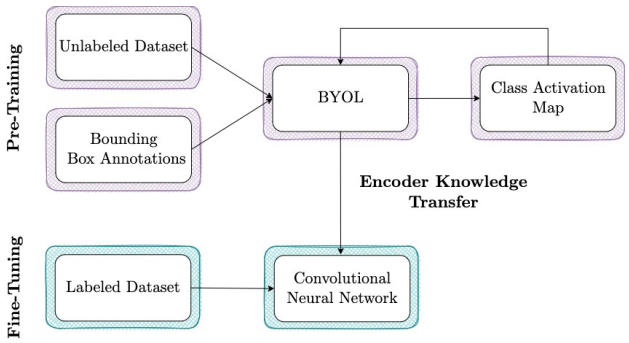

Fig. 1. Pictorial overview of our pipeline. We perform self-supervised pre-training with our auxiliary loss using unlabeled Chest X-rays and bounding box annotations from the NIH CXR dataset. The BYOL encoder is then used for downstream fine-tuning.

sound image representations and outperform approaches not using metadata across various downstream tasks. Rozenberg et al. [16] propose a technique that *learns* to localize objects using a small fraction of the dataset for which annotations are available, and the remaining images have only image-level labels. Utilizing a unique loss function and architecture incorporating shift-invariance and patch dependence, they achieve SOTA localization performance on the NIH CXR dataset.

Our work takes inspiration from Hu et al. [15] and Rozenberg et al. [16] and improves the SOTA using the model proposed in Bootstrap Your Own Latent (BYOL). BYOL is an SSL approach for image representation learning without using negative pairs [17]. We develop a self-supervised and weakly supervised pre-training pipeline with an auxiliary loss followed by supervised fine-tuning to enable superior representation learning. In the pre-training phase, we train our model on the NIH CXR [11] dataset in a self-supervised manner with an auxiliary loss, which is calculated using the probability distributions of the Gradient-weighted Class Activation Mapping (Grad-CAM) heat-map and its corresponding bounding box annotation(s) (if available) [18]. Finally, in a supervised fashion, we fine-tune this model on the CheXpert dataset [19] and output multi-label classification probabilities. The model pipeline is briefly visualized in Fig. 1 and is explained further in the Methods section. Through experiments, we aim to provide evidence that our proposed model:

1) generates more localized and accurate interpretability outputs than SOTA models
2) achieves classification performance comparable to or better than SOTA-supervised models
3) outperforms existing explainability outputs using only a limited number of annotations ($< 1\%$ of the dataset)
4) is generalizable and retains performance across datasets

II. Methodology

A. Explainability

Any explainability technique in healthcare would roughly come under one of two groups: visual or textual. In this study, we focus on visual explainability. Class Activation Mapping (CAM) is one of several methods to visualize which parts of an image a CNN is *looking* at to make a decision. CAM generates heat maps by leveraging global average pooling layers to activate class-specific semantic regions in images [20]. Some popular variations of CAM are listed below:

- Grad-CAM takes into account the weights and gradients going into the final convolution layer to emphasize the regions that contribute more to the final prediction [18]
- Grad-CAM++ is an expansion of Grad-CAM, with enhanced localization of multiple instances of a class [21]
- Score-CAM is a gradient-free method and uses a linear mix of activation maps and weights [22]

Grad-CAM generates an approximate localization map of the *activated* regions in the prediction images using the class-specific gradients flowing into the final convolutional layer. To obtain the activation maps, we calculate the gradient of y^c (score for class c) w.r.t feature maps of the convolutional layer. The gradients from backpropagation are globally average-pooled to obtain the importance weights α_k^c of feature map k. As indicated in (2), the final heat map is created by taking the weighted sum of the feature maps, A^k, with weights α_k^c, followed by a ReLU.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients from backprop}} \tag{1}$$

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{2}$$

Our model compares the CAM output of an input image to its corresponding bounding box mask and calculates what we refer to as CAM loss.

B. Pre-training Techniques

Supervised pre-training techniques are useful in tasks with limited data and annotations. One popular pre-training technique is transfer learning, where a model is fine-tuned from pre-trained weights rather than trained from random initialization. Although pre-training techniques have boosted the growth of DL applications, this technique is not always effective when the downstream task and dataset differ considerably from the original training data.

TorchXRayVision (XRV) is an open-source library designed to interact with CXR datasets, their pre-trained models, and other pre-processing tools [23]. The models are trained on some of the most extensive public CXR datasets such as CheXpert, Chest X-ray14 (NIH CXR), and MIMIC CXR [11], [19], [24]. We use the pre-trained models available in the XRV library to act as supervised SOTA.

C. Self-Supervised and Weakly-Supervised Learning

SSL is a training method that allows learning robust features without human annotations. This makes it a good choice for situations where data is scarce, like in healthcare. Unlabeled domain-specific images are used in SSL approaches during pre-training by generating labels from the data to learn more relevant representations. For example, during the pre-training phase, BERT, a revolutionary NLP model, learns from text samples with some missing words. The model is then trained to extract supervisory signals from the input data to predict missing words [25].

Weakly supervised learning is a blanket term that covers three types of weak supervision: inaccurate supervision, where the labels are erroneous; inexact supervision, where the data has weak or coarse labels; and incomplete supervision, where labels exist only for a subset of training data [26]. This work uses the bounding box annotations available in NIH CXR to provide weak supervision during self-supervised pre-training.

1) Bootstrap Your Own Latent: BYOL is a new method for learning self-supervised image representations. It employs *online* and *target* neural networks that learn from one another. The online network is trained using an augmented view of an input image to predict the representation of another enhanced version of the same image in the target network. We use BYOL because it nearly equals the best-supervised baseline in terms of top-1 accuracy on ImageNet and surpasses other self-supervised baselines such as SimCLR while requiring 30% fewer parameters [17].

The online network has weights θ and consists of three stages: an *encoder* f_θ, a *projector*, g_θ and a *predictor* q_θ, as shown in Fig. 2. The target and online networks have the same architecture, but the former has different weights, δ. The target network weights are an exponential moving average of the online parameters, θ. Given an input image, x, BYOL produces two augmented views, v and v', by applying augmentations t and t', respectively. From their respective augmented views, the online network outputs a *representation* y_θ and a *projection* z_θ, whereas the target network outputs y'_δ and a projection z'_δ. The output of the final pooling layer is given by the representation y, which is projected to a smaller space using a multi-layer perceptron g_θ.

BYOL minimizes the mean squared error $\mathcal{L}_{\theta,\delta}$ between the normalized prediction outputs from the online network, $q_\theta(z_\theta)$, and the target projections, $sg(z'_\delta)$, where sg means stop-gradient. It is symmetrized by alternatively feeding v' and v to the online and target networks, respectively, to compute $\widetilde{\mathcal{L}}_{\theta,\delta}$. After each training step, the optimization step is performed to minimize $\mathcal{L}^{\text{BYOL}}_{\theta,\delta} = \mathcal{L}_{\theta,\delta} + \widetilde{\mathcal{L}}_{\theta,\delta}$ w.r.t θ only, and given a target decay rate $\tau \in [0, 1]$, δ is updated as given in (5). At the end of the training, only the encoder f_θ is retained for further fine-tuning.

$$\mathcal{L}_{\theta,\delta} = \|\overline{q_\theta}(z_\theta) - \overline{z}'_\delta\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\delta \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\delta\|_2} \quad (3)$$

$$\theta \leftarrow \text{optimizer}(\theta, \nabla_\theta \mathcal{L}^{\text{BYOL}}_{\theta,\delta}, \eta) \quad (4)$$

TABLE I
DATASET CHARACTERISTICS

Task	Dataset	# of images		
		Train	*Validation*	*Test*
Pre-training	NIH CXR	86,524	-	25,596
Fine-tuning	CheXpert	178,731	44,683	234

$$\delta \leftarrow \tau\delta + (1 - \tau)\theta \quad (5)$$

2) BYOL with CAM loss: If an input image x has an associated bounding box annotation in NIH CXR, we calculate a weighted sum of the feature maps y_θ as given in (1)-(2) to obtain the Grad-CAM heat-map $\mathcal{L}_{\text{GC}}(y_\theta)$. All annotations of an input image are merged into one binary mask ($BB\ Mask$), and the Binary Cross Entropy loss, $\mathcal{L}_{CAM}$, is calculated using the Grad-CAM heat-map, $\mathcal{L}_{\text{GC}}$. In (6)-(7), y refers to the binary mask, σ denotes the Sigmoid function, $l_{1,c}$ is the loss of the first sample in a batch of size N, and c is the class of sample numbered n.

$$\mathcal{L}_{CAM}(\mathcal{L}_{\text{GC}}, y) = mean(\{l_{1,c}, \cdots, l_{N,c}\}^\top) \quad (6)$$

$$l_{n,c} = -y_{n,c} \cdot \log \sigma(\mathcal{L}^{n,c}_{\text{GC}}) - (1 - y_{n,c}) \cdot \log(1 - \sigma(\mathcal{L}^{n,c}_{\text{GC}})) \quad (7)$$

Finally, $\mathcal{L}^{\text{BYOL}}_{\theta,\delta}$ is updated with our loss as shown in (8). λ controls how much we penalize incorrect Grad-CAM outputs.

$$\mathcal{L}^{\text{BYOL}}_{\theta,\delta} = \mathcal{L}_{\theta,\delta} + \widetilde{\mathcal{L}}_{\theta,\delta} + \lambda \cdot \mathcal{L}_{CAM} \quad (8)$$

III. EXPERIMENTS

A. Datasets

The NIH CXR dataset comprises of 112,120 CXR images from 30,805 individual patients, while CheXpert has 224,316 CXR images from 65,240 unique patients. The former is used for the pre-training task, while the latter is used for the downstream classification task. The labels for both datasets were extracted using Natural Language Processing from the corresponding radiological reports. Each report was labeled for the occurrence of 14 pathologies as either *positive* or *negative*, and CheXpert contains an extra *uncertain* label. We regard the unreliable labels in CheXpert to be *negative*. Table I provides a summary of the dataset.

The datasets were divided by patient ID to prevent data leakage, and we followed the standard training and testing splits. The NIH CXR dataset contains 984 manually annotated bounding box annotations, divided 80/20 between training and test splits. CheXpert offers an official validation split that we use for testing, and we divide the training dataset into training (80%) and validation (20%). The 234 images in the CheXpert test set were annotated by three board-certified radiologists separately after examining the images. In their annotations, all present/uncertain likely cases are treated as positive, and all absent/uncertain/unlikely cases are treated as negative. The final label is generated using a majority vote.

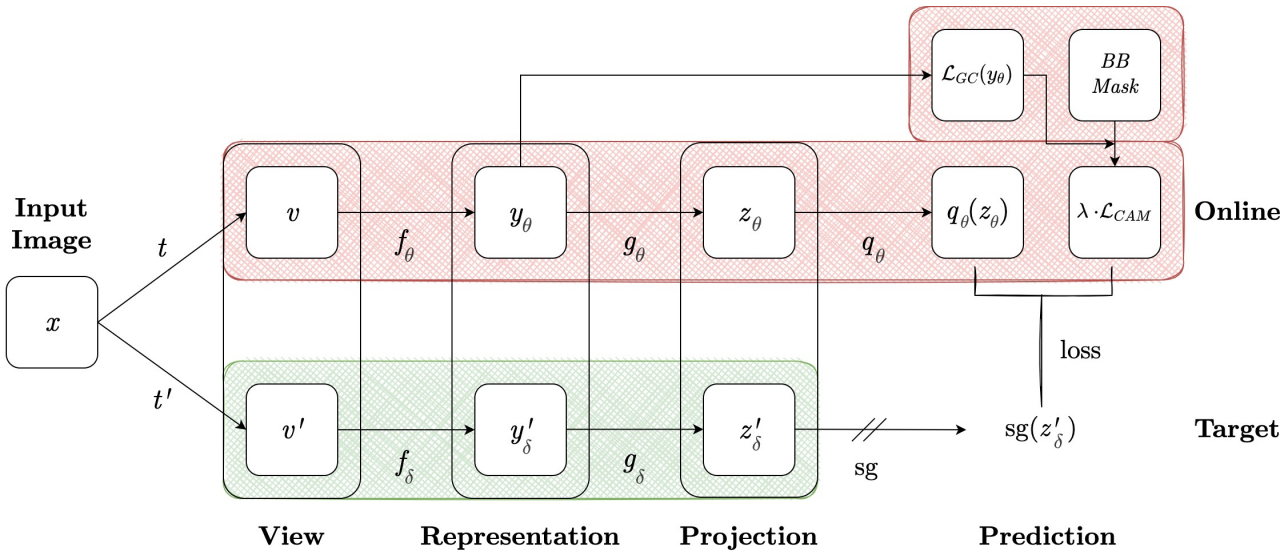

Fig. 2. BYOL with CAM loss. θ are trained weights, δ are an exponential moving average of θ, and sg is stop-gradient. CAM loss is calculated only if an input image has a corresponding bounding box (BB) annotation. Model minimizes similarity and CAM loss between $q_\theta(z_\theta)$, $\mathcal{L}_{CAM}$ and $\text{sg}(z'_\delta)$. Encoder f_θ is retained for downstream fine-tuning

B. Evaluation Metrics

Intersection over Union (IoU) is a metric used to quantify the amount of intersection between two bounding boxes. We generate bounding boxes from the Grad-CAM outputs to compare them quantitatively with bounding box annotations from NIH CXR. For this, we first threshold pixels having a value less than 127 in the Grad-CAM output, find corresponding contours, and then draw an approximate rectangle around the region of interest. In an ideal situation, models would achieve IoU scores close to 1.0, which signifies perfect overlap. We report the area under the receiver operating characteristic (AUROC) and precision-recall (AUPRC) curves to measure classification performance. We also report the model's output values as confidence scores for the ground truth label.

C. Architecture configurations

We use BYOL during the pre-training phase and reuse the encoder backbone for downstream fine-tuning. We train the model using three popular CAM variations - Grad-CAM, Grad-CAM++, and Score-CAM to choose the optimal CAM algorithm for this task.

We also train the model using ResNets of varying depths (18/34/50) to confirm that this task's encoder is not over-parameterized. The ResNet encoders are not pre-trained and are modified to accept single-channel (grayscale) input images. For BYOL, we use a projection size of 256 and a hidden projection size of 512. While fine-tuning, we replace the encoder's fully-connected layer with another having an *input* dimension of 2048 for ResNet-50, 512 for ResNet-34/ResNet-18, and an *output* dimension of 14 corresponding to the number of classes in CheXpert.

D. Training

a) Pre-training: The PyTorch framework was used for all training, validation, and testing purposes. Before training, all images were resized to 256×256 pixels and normalized with the dataset mean and standard deviation. Data augmentation was performed using transformations like gaussian blur, random resized crop, and random affine. The target network's augmentations have a higher probability of occurrence and are slightly stronger to enable better representation learning. We update BYOL's weights using Adam optimizer, minimizing $\mathcal{L}^{\text{BYOL}}_{\theta,\delta}$ in (8), with hyper-parameters set to batch size $= 128$, $\lambda = 2$, learning rate $= 0.0001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The exponential moving average parameter τ for the target network is set to 0.99. Increasing the value of λ has a negative effect on the model performance. The models were trained for 500 epochs over 4 Nvidia 3080 Ti cards.

b) Fine-tuning: We augment the data using only random affine transformations with values similar to those of the target network and $p = 0.5$. We fine-tune the pre-trained BYOL encoder using Adam optimizer with all other hyper-parameters similar to those during pre-training. As our loss function, we use Binary Cross Entropy ((6)-(7)) with a weight for positive examples (different for training and validation splits). After four consecutive epochs with no improvement in training loss, we implement a learning rate scheduler that decreases the learning rate by a factor of 0.1 up to a minimum value of 0.000001. The maximum number of epochs the model was trained for was 200, and the model with the greatest validation AUROC was retained.

E. Comparative Study

We choose the best-performing configuration to carry out ablation and comparative studies. We compare our model

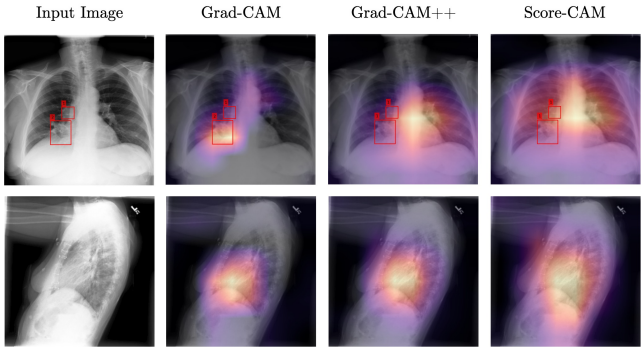

Fig. 3. Comparison of various CAM outputs. Chest X-rays in the first row are from NIH CXR, and others are from CheXpert. The bounding boxes in red show the ground truth annotation. The bright regions correspond to the highest level of activation.

with current supervised models trained using image-level labels on visual explainability and multi-label classification performance. The different configurations that we train are:

- Baseline: A ResNet-50 model trained from scratch on either of the datasets mentioned in Table I in a supervised manner to establish baseline performance
- XRV (**SOTA**): A DenseNet-121 pre-trained model from the XRV library trained on multiple large CXR datasets
- XRV + CAM Loss: The same model as XRV but fine-tuned on NIH CXR with our proposed CAM Loss
- SSL + FT (**Ablation**): A ResNet backbone pre-trained on NIH CXR and fine-tuned on CheXpert
- SSL + CAM Loss + FT (**Proposed**): A ResNet backbone pre-trained on NIH CXR with our proposed CAM loss and then fine-tuned on CheXpert

IV. Results

A. Architecture Configurations

To choose the optimal CAM algorithm for the visual explainability of our model, we compare the outputs of Grad-CAM, Grad-CAM++, and Score-CAM, with ground truth as seen in Fig. 3. We choose Grad-CAM as it produces precise activation maps resulting in better localization performance.

Supplementary Fig. 1 shows the corresponding Grad-CAM outputs in which the localization ability degrades as smaller ResNets are used, indicating that they cannot capture the complex features of CXRs. Hence, we choose ResNet-50 as the backbone encoder for our model.

B. Visual Explainability

To assess the visual explainability of our models, we generate heat maps using Grad-CAM for samples containing various pathologies from the NIH CXR dataset, as shown in Fig. 4. The heat maps follow the Magma color map in OpenCV, with the bright regions corresponding to the highest activation level while classifying CXRs. The red bounding boxes show the ground truth annotation in the NIH CXR dataset. Grad-CAM outputs from the XRV + CAM Loss model don't improve over the standard XRV outputs and hence are not included.

Our SSL + CAM Loss + FT model performs the best when compared with baselines and XRV's models. Further, there is a noticeable improvement over the SSL + FT model, which was not trained with our CAM Loss. This advantage is apparent in the lateral view image in the third row of Fig. 4, which is positive for *Effusion*. The SSL + FT model activates incorrectly, while the CAM loss variant shows much better localization. Our model performs reliably across all pathologies, imaging views, and datasets, demonstrating knowledge retention even after downstream fine-tuning. For example, the scan in the second row is positive for the *Infiltrate* class, which is not present in CheXpert. It correctly activates inside the bounding box annotations and performs better than other models.

Another example is shown in the fourth row with ground truth labels *Atelectasis* and *Mass*. CheXpert doesn't have a *Mass* class (confidence scores are only for *Atelectasis*), but NIH CXR does. Our model activates for the corresponding region well, which means that it was able to retain anatomical information from the pre-training phase. Our model performs inadequately with certain hard-to-differentiate pathologies like *Atelectasis*, which refers to the lung's partial collapse or incomplete inflation. It fails to identify the presence of *Atelectasis* in the scan in the second last row. But in certain cases, provided the pathology region-of-interest is large and discriminative enough, the model successfully predicts *Atelectasis*, as shown in the first row. All models show good activation around the enlarged heart for the scan in the last row, with the CAM loss variant being the most confident and concise. *Cardiomegaly* refers to an enlarged heart and is one of the easiest pathologies to differentiate due to its being visually distinctive. Our model outputs are attributed with high IoU and confidence scores compared to other models. For the NIH CXR test set, the XRV model achieves an average IoU of 0.61 compared to our SSL + CAM Loss + FT model's score of 0.92, signifying an improvement of 31%.

Although CheXpert does not offer bounding box annotations, we generate the CAM outputs for its samples to assess our model's robustness. Supplementary Fig. 2 presents the Grad-CAM outputs for samples containing various pathologies from the CheXpert dataset. We observe that the predictions from our SSL + CAM Loss + FT model are more localized. For some visually discriminative classes like *Support Devices* and instances of *No Finding* in the last two rows, respectively, our model correctly activates for the former class and does not activate as expected for *No Finding*. For the last row, even though the CXR does not contain any pathology, all models except ours incorrectly activate regions in the sample.

C. Multi-label Classification

Even if a CXR classification model can *look* at the right place, it is paramount to accurately classify the pathologies present in the CXRs. In this part, we compare the classification performance of all models included in the comparative analysis. As our model was fine-tuned using CheXpert for the downstream classification task, we first present the evaluation scores calculated on CheXpert's official validation set (our

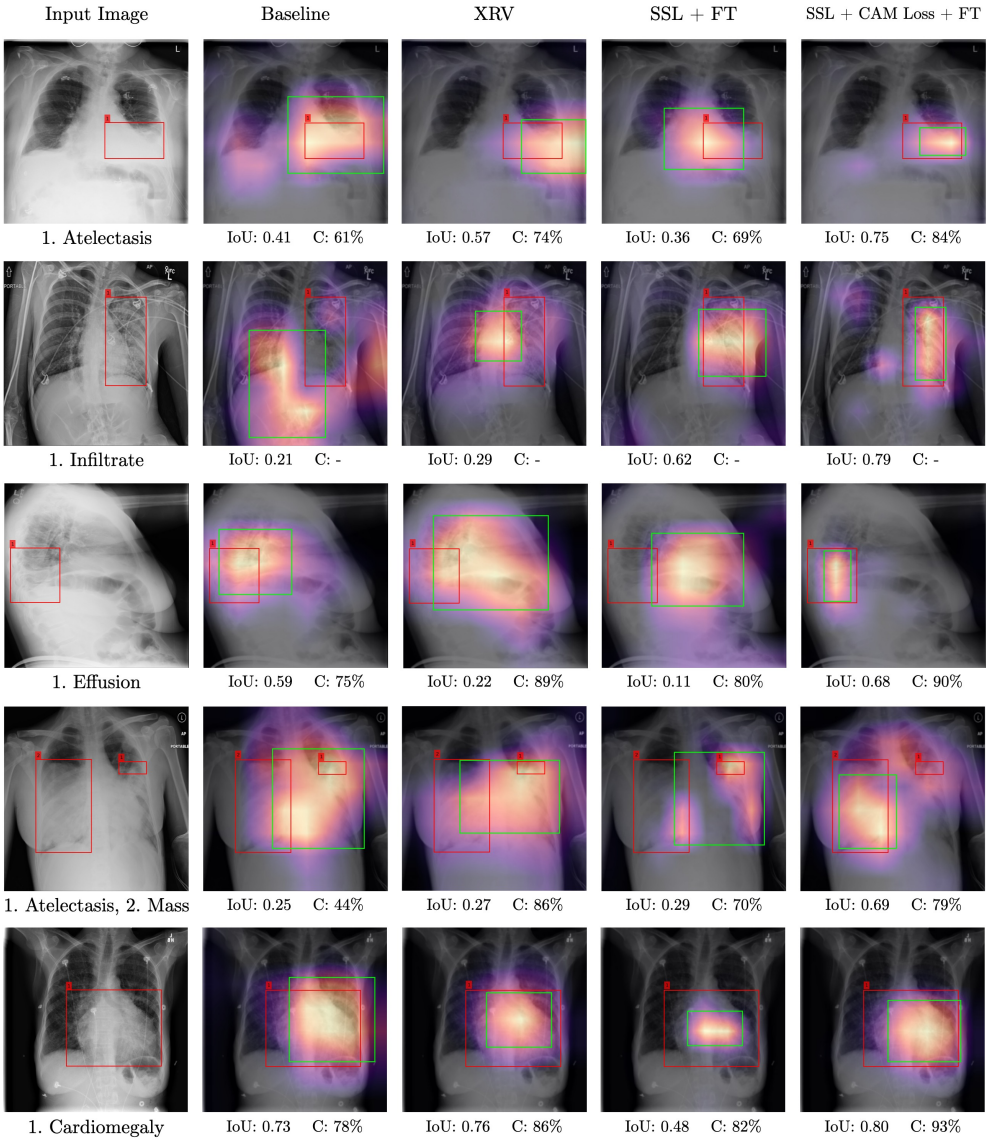

Fig. 4. Grad-CAM outputs of NIH CXR test set. The first column contains ground truth; the others are outputs from various models. IoU and model's confidence (C) measures are detailed in Evaluation Metrics. The bounding boxes in red show ground truth annotation and the ones in green are obtained from Grad-CAM outputs.

test set) in Table II. Our proposed SSL + CAM Loss + FT model achieves the highest AUROC scores for 8 out of 13 classes and comparable AUROC scores for the rest. Our model outperforms the other baseline and SOTA models. The XRV models weren't trained on the labels for *No Finding* and *Pleural Other*, and the CheXpert validation set doesn't have any samples for *Fracture*, which is why the corresponding cells are empty in Table II.

Figure 5's precision-recall curves illustrate the trade-off between precision and recall for various thresholds. We obtain a high area under the curve when recall and precision values are high. A high recall correlates with a low rate of false negatives, while a high precision correlates with a low percentage of false positives. For each model, we give the Average precision (AP), which summarises the above plot as a weighted mean of precisions acquired at each threshold, where the weight is determined as the increase in recall from the preceding threshold. We also exhibit iso-f1 curves that include all points in the precision/recall space with identical F1 scores. The SSL + CAM Loss + FT model achieves the greatest AP score (0.68), which is the best model.

We also evaluate and compare the classification performance on the NIH CXR dataset's official test set to verify robustness. We calculate the AUROC scores for the subset of pathologies NIH CXR has in common with CheXpert (Table III) and observe that performance on NIH CXR is similar to that on CheXpert. Our model achieves the highest AUROC scores for 2 out of 7 classes present and comparable AUROC scores for the rest.

TABLE II
MULTI-LABEL CLASSIFICATION AUROC ON CHEXPERT TEST SET (VALUES IN PARENTHESIS REPRESENT THE NUMBER OF SAMPLES POSITIVE FOR THE CLASS)

Classes	**Baseline**	**XRV** (SOTA)	**XRV + CAM Loss**	**SSL + FT**	**SSL + CAM Loss + FT** (ours)
No Finding (38)	0.72	-	-	0.82	**0.84**
Enlarged Cardiomediastinum (109)	0.54	0.67	0.64	0.62	**0.69**
Cardiomegaly (68)	0.75	**0.87**	0.84	0.83	0.85
Lung Opacity (126)	0.75	0.84	0.80	0.86	**0.87**
Lung Lesion (1)	0.67	**0.88**	0.68	0.81	0.84
Edema (45)	0.81	**0.92**	0.91	0.90	**0.92**
Consolidation (33)	0.85	0.87	0.88	0.83	**0.89**
Pneumonia (8)	0.66	**0.74**	0.69	0.60	0.65
Atelectasis (80)	0.72	0.85	0.85	0.85	**0.87**
Pneumothorax (8)	0.58	**0.73**	0.69	0.69	0.70
Pleural Effusion (67)	0.70	**0.86**	0.82	0.80	0.83
Pleural Other (1)	0.81	-	-	0.95	**0.97**
Fracture (0)	-	-	-	-	-
Support Devices (107)	0.64	0.84	0.80	0.80	**0.85**
Average AUROC	0.71	0.82	0.78	0.79	**0.83**

TABLE III
MULTI-LABEL CLASSIFICATION AUROC ON NIH CXR TEST SET. ONLY CLASSES COMMON WITH CHEXPERT WERE TESTED (VALUES IN PARENTHESIS REPRESENT THE NUMBER OF SAMPLES POSITIVE FOR THE CLASS)

Classes	**Baseline**	**XRV** (SOTA)	**XRV + CAM Loss**	**SSL + FT**	**SSL + CAM Loss + FT** (ours)
Atelectasis (3279)	0.75	0.85	0.86	0.84	**0.87**
Cardiomegaly (1069)	0.80	**0.89**	0.86	0.85	0.86
Consolidation (1815)	0.78	**0.90**	0.88	0.83	0.85
Edema (925)	0.82	0.91	0.89	0.90	**0.92**
Pneumonia (555)	0.65	**0.87**	0.83	0.86	0.86
Effusion (4658)	0.76	**0.90**	0.85	0.81	0.87
Pneumothorax (2665)	0.69	0.79	0.74	**0.83**	0.82
Average AUROC	0.75	**0.87**	0.84	0.84	0.86

V. CONCLUSION

Chest X-rays are an important tool for identifying a variety of thoracic disorders. However, healthcare institutions face pressure to find enough qualified radiologists to provide prompt and appropriate patient care. Machine learning models can help with this, but they need to be interpretable and trustworthy to be used in the real world. Clinical applications of explainable DL algorithms will likely be a human-in-the-loop hybrid in which medical experts, like radiologists, control the decision-making process [27]. Current state-of-the-art models for chest X-ray classification do not always validate the accuracy of their outputs and can lose performance when applied to other datasets. This paper suggests a new approach to addressing these issues in settings with limited resources.

We propose a self-supervised method that uses bounding boxes as weak labels to improve the interpretability of medical image classification algorithms. We create a pre-training pipeline that uses a modified BYOL network, and the NIH CXR dataset's constrained bounding boxes, followed by fine-tuning on the CheXpert dataset. Using annotations from a small fraction of a large unlabeled dataset during the pre-training phase significantly increases localization and certainty metrics while maintaining performance across datasets. We also evaluate various model configurations and choose ResNet-50 as the backbone encoder and Grad-CAM as the optimal CAM algorithm.

The experiments show that the proposed model generates more localized and accurate Grad-CAM outputs than current state-of-the-art models, achieves superior classification performance and is more generalizable and transferable to other datasets. The use of annotations during the pre-training phase provides feedback to the model on where to *look* during the representation learning phase. This improves performance compared to current state-of-the-art techniques that only use image-level labeling.

Overall, this research shows that a self-supervised approach using bounding boxes as weak labels can improve the interpretability and generalizability of medical image classification algorithms, and has the potential to be applied to other medical imaging modalities. We believe that further research could be done with more annotated data and higher input resolution to further improve performance.

ACKNOWLEDGEMENTS

We thank IHub-Data, International Institute of Information Technology, Hyderabad, for their support.

REFERENCES

[1] R. M. Scheffler, J. X. Liu, Y. Kinfu, and M. R. Dal Poz, "Forecasting the global shortage of physicians: an economic-and needs-based approach," *Bulletin of the World Health Organization*, vol. 86, pp. 516–523B, 2008.

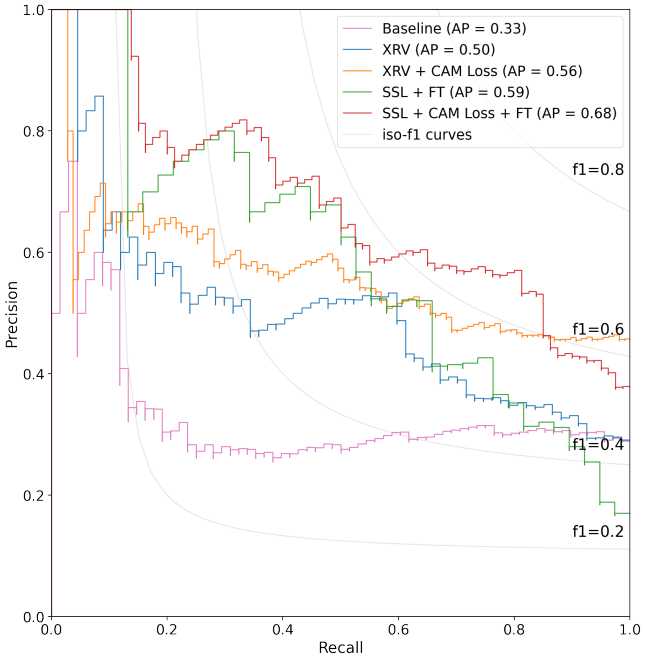

Fig. 5. Precision-Recall plots for each model tested on the CheXpert test set. The Average Precision (AP) of each model is a weighted mean of precisions at each threshold. An iso-F1 curve contains all points in the precision/recall space whose F1 scores are the same.

[2] A. J. Sweatt, H. K. Hedlin, V. Balasubramanian, A. Hsi, L. K. Blum, W. H. Robinson, F. Haddad, P. M. Hickey, R. Condliffe, A. Lawrie *et al.*, "Discovery of distinct immune phenotypes using machine learning in pulmonary arterial hypertension," *Circulation research*, vol. 124, no. 6, pp. 904–919, 2019.

[3] D. Spathis and P. Vlamos, "Diagnosing asthma and chronic obstructive pulmonary disease with machine learning," *Health informatics journal*, vol. 25, no. 3, pp. 811–827, 2019.

[4] S. V. Razavi-Termeh, A. Sadeghi-Niaraki, and S.-M. Choi, "Asthma-prone areas modeling using a machine learning model," *Scientific Reports*, vol. 11, no. 1, pp. 1–16, 2021.

[5] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[6] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, pp. 1–10, 2021.

[7] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 103–110.

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

[9] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3478–3488.

[10] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 201–208.

[11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[12] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[13] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.

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

[15] S.-Y. Hu, S. Wang, W.-H. Weng, J. Wang, X. Wang, A. Ozturk, Q. Li, V. Kumar, and A. E. Samir, "Self-supervised pretraining with dicom metadata in ultrasound imaging," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 732–749.

[16] E. Rozenberg, D. Freedman, and A. Bronstein, "Localization with limited annotation for chest x-rays," in *Machine Learning for Health Workshop*. PMLR, 2020, pp. 52–65.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[19] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "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.

[20] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." *CVPR*, 2016.

[21] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[22] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[23] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir *et al.*, "Torchxrayvision: A library of chest x-ray datasets and models," *arXiv preprint arXiv:2111.00595*, 2021.

[24] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[26] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[27] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.