Radiomics-Guided Global-Local Transformer for Weakly Supervised Pathology Localization in Chest X-Rays

Yan Han, Gregory Holste, Ying Ding, Ahmed Tewfik, *Fellow, IEEE*,
Yifan Peng, and Zhangyang Wang, *Senior Member, IEEE*

***Abstract*— Before the recent success of deep learning methods for automated medical image analysis, practitioners used handcrafted radiomic features to quantitatively describe local patches of medical images. However, extracting discriminative radiomic features relies on accurate pathology localization, which is difficult to acquire in real-world settings. Despite advances in disease classification and localization from chest X-rays, many approaches fail to incorporate clinically-informed domain-specific radiomic features. For these reasons, we propose a Radiomics-Guided Transformer (RGT) that fuses *global* image information with *local* radiomics-guided auxiliary information to provide accurate cardiopulmonary pathology localization and classification *without any bounding box annotations*. RGT consists of an image Transformer branch, a radiomics Transformer branch, and fusion layers that aggregate image and radiomics information. Using the learned self-attention of its image branch, RGT extracts a bounding box for which to compute radiomic features, which are further processed by the radiomics branch; learned image and radiomic features are then fused and mutually interact via cross-attention layers. Thus, RGT utilizes a novel end-to-end feedback loop that can bootstrap accurate pathology localization only using image-level disease labels. Experiments on the NIH ChestXRay dataset demonstrate that RGT outperforms prior works in weakly supervised disease localization (by an average margin of 3.6% over various intersection-over-union thresholds) and classification (by 1.1% in average area under the receiver operating characteristic curve). We publicly release our codes and pre-trained models at `https://github.com/VITA-Group/chext`.**

***Index Terms*— chest X-ray, deep learning, disease localization, radiomics, Transformer**

I. Introduction

In medicine, *radiomics* refers to the process of extracting quantitative and semiquantitative features from medical images, such as radiographs or computed tomography scans, for improved decision support [1]. These handcrafted radiomic features aim to describe a local "region of interest" such as a tumor with numeric features that assess qualities such as size, shape, texture, variations in pixel intensity, and relationships between neighboring pixels [2]. Given their advantages, researchers have explored the performance of radiomic features for chest X-ray analysis. For example, Shi *et al.* [3] and Saygılı [4] each extracted a set of radiomic features, which were then used to diagnose different types of pneumonia. Bai *et al.* [5] proposed a hybrid model to encode the combination of radiomic features and clinical information. Ghosh *et al.* [6] presented a new handcrafted feature to distinguish between severe and nonsevere patients. However, all of the above methods rely on *accurate pathology localization annotations* to extract radiomic features from a correct and clinically meaningful region of interest [7]. Such bounding boxes are usually expensive and time-consuming to acquire by humans and, if inaccurate, will tremendously degrade the reliability of radiomic features. There is thus an unmet need to automatically localize cardiopulmonary pathologies on chest X-rays to facilitate extraction of radiomic features.

Throughout the rapid development of deep learning approaches for medical image analysis, many researchers have made efforts utilizing convolutional neural networks (CNNs) to build automated systems for chest X-ray abnormality classification and localization [8]–[18]. However, CNN methods bear several limitations when applied to the domain of chest radiography. First, CNNs do not naturally incorporate contextual prior information, such as reason for imaging and patient history, or domain knowledge such as human anatomy and typical disease presentation on imaging. Since radiomic features are designed by humans and semantically describe local medical image regions, they represent an auxiliary modality of information embedded with domain-specific quantitative features that can enhance automated disease localization and classification. Second, chest X-rays have more subtle discriminative features compared to natural images, making their recognition more challenging. Finally, though many have studied the interpretability of deep image classifiers for other data [19]–[24], deep CNNs are often criticized for their lack of human interpretability, thus posing a major barrier to their adoption by clinicians.

With this in mind, Transformers, which have seen a surge in popularity for a variety of visual recognition tasks, provide a promising alternative to CNNs for modeling chest X-rays.

Manuscript submitted <date>.

Yan Han, Gregory Holste, Ying Ding, Ahmed Tewfik, and Zhangyang Wang are with The University of Texas at Austin, Austin, TX 78705 USA (email: yh9442@utexas.edu; gholste@utexas.edu; ying.ding@ischool.utexas.edu; tewfik@austin.utexas.edu; atlaswang@utexas.edu).

Yifan Peng is with Weill Cornell Medical College, New York, NY 10065 (email: yip4002@med.cornell.edu).

The Transformer was first introduced in the context of l
language processing [25]–[27], followed by its recent s
in computer vision [28]–[30] and multi-modal learnin
The Transformer architecture can be considered a "un
modeling tool" that can unify the feature extraction and
processes from different input modalities with a *single*
that does not require domain-specific architecture tweal
example, Arkbari *et al.* [32] demonstrated the ability tc
powerful multi-modal representations from unlabeled
audio, and text data, using a single multimodal Transf
Nagrani *et al.* design a bottleneck fusion technique that
audio- and video-derived features to interact throughou
custom Transformer architecture [33]. And Shvetsova *et a*
proposed a multi-modal, modality agnostic fusion Trans
to learn to exchange information between multiple mod
such as video, audio, and text, and integrate them
jointly multi-modal representation to obtain an embe
that aggregates multi-modal temporal information.

In the context of modeling chest X-rays, we obser[illegible] the unique potential for a Transformer-based architecture to *naturally and jointly learn from two "views" of chest X-rays*: (1) raw X-ray images that contain rich contrast details, hence benefiting from the data-driven learning capacity, and (2) radiomics that encode domain-specific quantitative features, thus guiding and regularizing the learning process with handcrafted local radiomic features. However, there exists a "chicken-and-egg" problem: extraction of useful radiomic features relies on accurate pathology localization, but the pathology localization is often absent and first needs to be learned or separately acquired.

This work presents **RGT**, a **R**adiomics-**G**uided **T**ransformer (Fig. 1). RGT consists of two Transformer-based branches, one for the raw chest X-ray and one for the radiomic features extracted from the corresponding image. Features extracted from these two "views" of the patient are then deeply fused with interaction via cross-attention layers [35]. Of note, the radiomic features need to be extracted from the learned pathology localizations, which are not readily available. The *key enabling technology* to resolve this hurdle is to construct a feedback loop, called the ***B**ring **Y**our **O**wn **A**ttention* (**BYOA**) module, which will be expanded in Sec III-B. During training, the image branch leverages its learned self-attention to estimate pathology location, which is then used to extract radiomic features from the original image for further processing by the radiomics branch. In addition to a supervised classification loss, we optimize the model with a contrastive loss that rectifies the image-derived and radiomics-derived "views" of the patient, and such an end-to-end optimization loop can bootstrap accurate pathology localizations from image data with *no bounding box annotations* used for training.

Our contributions are outlined as follows:

- We leverage radiomics as an "auxiliary input modality" that both correlates with the raw image modality and encodes domain-specific quantitative features. We then propose a novel radiomics-guided cross-attention Transformer, **RGT**, to jointly extract and fuse global image features and local radiomic features for disease localization and classification in chest X-rays.

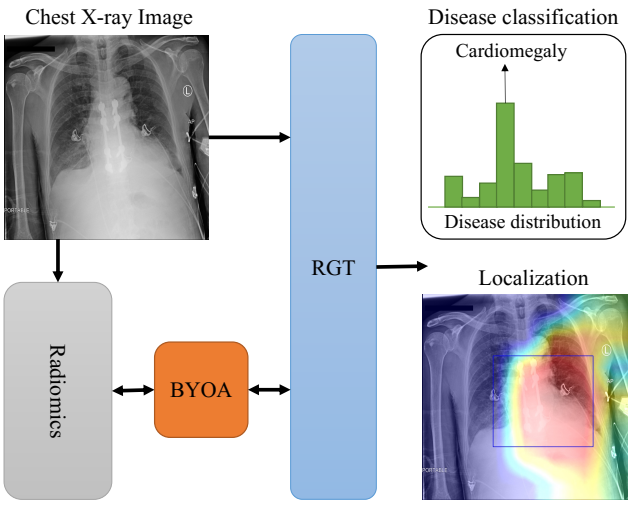

Fig. 1. General overview of our **R**adiomics-**G**uided **T**ransformer (**RGT**) framework for weakly supervised cardiopulmonary disease localization and classification from chest X-rays. RGT takes a chest X-ray as the input and produces a heatmap for pathology localization, from which a bounding box is obtained. Radiomic features are further extracted from the bounded region and fused with image-derived features to classify the pathology present. The detailed views of **RGT** framework and ***B**ring **Y**our **O**wn **A**ttention* (**BYOA**) module are given in Fig. 2 and Fig. 3, respectively.

- To resolve the key "chicken-and-egg" problem of extracting radiomic features without available cardiopulmonary pathology localization, we construct an innovative optimization loop where the learned image-level attention map is used to extract local radiomic features. Such an end-to-end loop can bootstrap accurate cardiopulmonary pathology localization from images without leveraging human-annotated bounding boxes.
- On the NIH ChestXRay benchmark [9], our approach achieves superior disease localization and classification results. RGT outperforms prior work in weakly supervised localization by an average margin of 3.6% over different intersection-over-union (IoU) thresholds.

II. Related Work

Radiomics in Medical Diagnosis. The design of radiomic features involves prior biological and medical knowledge, thus enriching the value provided by the raw pixel intensities of a medical image [37]. In the study of image-based biomarkers for cancer staging and prognostication, radiomics has shown promising predictive power [38]. Radiomics extracts quantitative features from medical images that can be used to represent tumor phenotypes, such as spatial heterogeneity of the tumor and spatial response variations. Eilaghi *et al.* [39] demonstrated that radiomic texture features, extracted from computed tomography (CT) scans, are associated with overall survival rate of pancreatic cancer. Chen *et al.* [40] revealed that the first-order radiomic features (e.g., mean, skewness, and kurtosis) are correlated with pathological responses to cancer treatment. Huang *et al.* [41] showed that radiomics could increase the positive predictive value and reduce the false-positive rate in lung cancer screening for small nodules compared to radiologists. Zhang *et al.* [42] found that radiomics

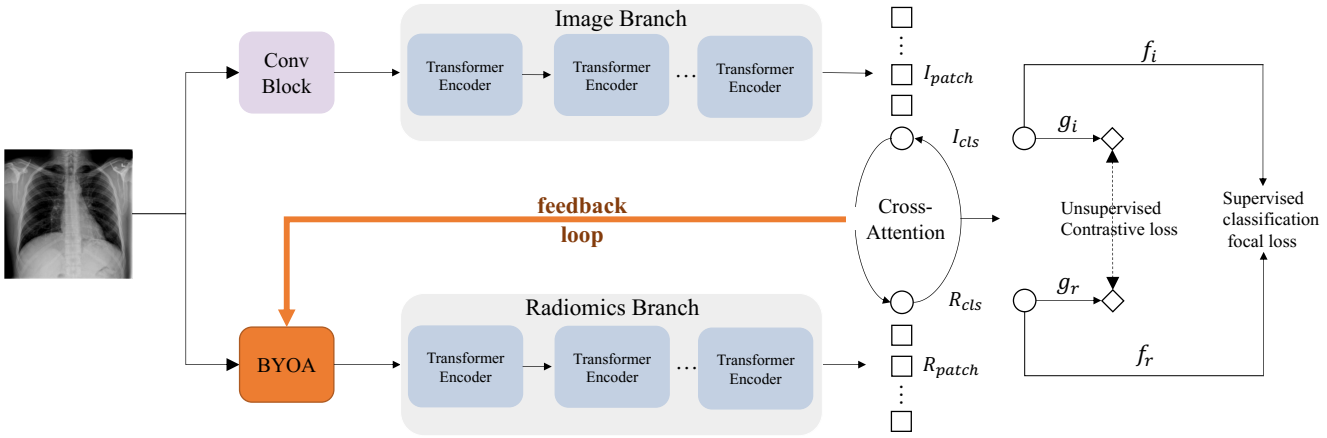

Fig. 2. Overview of our proposed model, RGT. The image branch is a Transformer that processes a chest X-ray, and the radiomics branch is a small Transformer that processes radiomic features generated by the Bootstrap Your Own Attention (BYOA) module (Fig. 3). The global image representations and local radiomics representations are then fused by an efficient cross-attention module operation on each branch's CLS tokens. Finally, the CLS tokens $\boldsymbol{I_{cls}}$ (from the image branch) and $\boldsymbol{R_{cls}}$ (from the radiomics branch) are used for disease classification. We optimize the classification error with the Focal Loss [36]. We also leverage a contrastive learning strategy that aims to rectify the global image view with the local radiomics view. Specifically, RGT generates an image view $\boldsymbol{z_i = g_i(I_{cls})}$ by a projection head $\boldsymbol{g_i}$ and radiomic view $\boldsymbol{z_r = g_r(R_{cls})}$ by projection head $\boldsymbol{g_r}$. We maximize the agreement between $\boldsymbol{z_i}$ and $\boldsymbol{z_r}$ via a contrastive loss (NT-Xent).

from multiparametric magnetic resonance imaging-based nomograms provided improved prognostic ability in advanced nasopharyngeal carcinoma. In positron emission tomography (PET)/CT imaging, Alongi *et al.* [43] leveraged radiomic features to successfully predict prostate cancer progression.

In comparison, deep learning is often criticized for being a "black box", lacking interpretability or reasoning in the form of human semantic concepts, despite achieving high predictive performance. This limitation has motivated many interpretable learning techniques including activation maximization [44], network inversion [45], GradCAM [46], and network dissection [47]. We believe that the joint utilization of radiomics and interpretable learning techniques in our framework can further advance accurate yet interpretable learning in the medical image domain.

Transformers for Medical Images. Recently, the Vision Transformer (ViT) [28] achieved state-of-the-art classification on ImageNet by directly applying Transformers with global self-attention to full-sized images. Before the advent of Transformers for visual recognition tasks, several works have augmented traditional CNNs with attention modules, seeing improved performance on various medical image analysis problems [48], [49]. Now inspired by the promising performance of ViT, researchers have begun to adapt the Transformer architecture to medical image analysis problems. For example, Chen *et al.* [50] and Hatamizadeh *et al.* [51] demonstrate that Transformers can achieve state-of-the-art medical image segmentation performance over CNN-based architectures. Similarly, Valanarasu *et al.* [52] proposed a gated axial-attention Transformer model to introduce an additional control mechanism in the self-attention module. Beyong medical image segmentation, Park *et al.* [53] utilized a hybrid CNN and Transformer framework for COVID-19 prediction. While these methods see improved performance by leveraging the Transformer architecture, none of these approaches incorporate domain-specific radiomic features. Han *et al.* [54] applied pre-extracted radiomic features to guide pneumonia detection from chest X-ray images. However, they adopted a convolutional backbone for image encoder, while using a specifically crafted radiomics encoder. Therefore, the method involves no joint interaction between image and radiomic features, requiring the use of previously acquired bounding boxes during training in order to extract radiomic features, limiting the method's usability in clinical practice.

III. Method

An overview of RGT is illustrated in Fig. 2. In the following subsections, we will first present Cross-Attention Vision Transformer (CrossViT), a recent two-branch ViT backbone on which RGT is built, and then describe the methodological innovations required to naturally incorporate domain-specific quantitative features in the form of radiomics for improved cardiopulmonary pathology localization and classification.

A. Preliminary: ViT and Cross-Attention

ViT first converts an image into a sequence of patch tokens by dividing the image into fixed-size patches and linearly projecting each patch into so-called "tokens". A special CLS (class) token is prepended to the sequence of image patches, as in the original BERT [26]. Then, all tokens are passed through stacked Transformer encoder layers. Finally, the hidden state corresponding to the CLS token is used as the aggregate sequence representation used for image classification.

A Transformer encoder is composed of a sequence of blocks, where each block consists of (1) a multi-headed self-attention and (2) a feed-forward neural network. Layer normalization and residual shortcuts are, respectively, applied before and after every block. The granularity of the patch size affects the accuracy and complexity of ViT. Therefore, ViT was observed to reach greater performance with smaller (more fine-grained) patch sizes, but at the cost of higher floating-point operations (FLOPS) and memory consumption [35]. To

relieve this problem, CrossViT [35] proposed a dual-branch ViT where each branch operates at a different patch size, as its own "view" of the image. The cross-attention module is then used to fuse information between the branches in order to balance the patch sizes and complexity. Similar to ViT, the final hidden vector obtained from the `CLS` tokens from the two branches are then used for image classification.

B. Our Proposed RGT Model

CrossViT supplies a graceful framework to simultaneously process and fuse two different "views" from the same input data (e.g., different-size image patches in the original paper) [35]. In RGT, we extend this idea by treating the raw image itself as one "view" and the radiomic feature extracted from this image as another "view" (Fig. 2). The global image representation and local radiomics representations are then fused by interacting through cross-attention. Here, a Transformer serves as the modality-agnostic backbone for both views.

Specifically, we introduce a dual-branch cross-attention Transformer where the first (primary) branch operates on the image, while the second (auxiliary) branch handles the radiomic features. To resolve the "chicken-and-egg" dilemma in extracting reliable radiomic features without bounding boxes, we have designed a novel ***B**ootstrap **Y**our **O**wn **A**ttention* (**BYOA**) module, using a feedback loop to learn pathology localization for radiomic feature extraction. A simple yet effective module is also utilized to fuse information between the branches. In the subsequent sections, we will describe the two branches, the BOYA module, and the fusion module.

Image Branch. The primary image branch uses a Progressive-Sampling ViT (PS-ViT) [55] as its backbone. Unlike the vanilla ViT that splits images into fixed-size tokens, PS-ViT utilizes an iterative and progressive sampling strategy to locate discriminative regions and avoid over-partitioning object structures. We experimentally observed PS-ViT outperforms ViT and other variants in our framework because it generates higher-quality and more structure-aware attention maps, which are crucial for estimating the pathology localization during training.

Radiomics Branch. The complementary radiomics branch is used to process and learn deep representations of radiomic features. Handcrafted features can encompass a wide range of categories, such as first-order (basic intensity and shaped-based features), second-order (texture features extracted from various matrices), and more advanced features including those calculated from Fourier and wavelet transforms. Specifically, the 107 radiomic features utilized in this work come from the following categories described below:

- First-order statistics measure the distribution of pixel intensities within the region of interest. Such features include energy (the measurement of the magnitude of pixel values), entropy (the measurement of uncertainty in the image values), and max/mean/median gray level intensity. In total, we extract 18 first-order radiomic features.
- Shape-based features – such as mesh surface, pixel surface, and perimeter – describe the two-dimensional size and shape of the region of interest. While RGT can only produce rectangular bounding boxes for radiomics extraction, shape-based features can still be useful to quantify the size and aspect ratio of the extracted region. A total of 14 shape features are used in this work.
- Gray-level features describe statistical patterns in the pixel intensity values, drawn from the Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM). In particular, we compute 24 GLCM features, 16 GLSZM features, 16 GLRLM features, 5 NGTDM features, and 14 GLDM features.

For this branch, we use a vanilla Transformer [56] as the radiomics encoder. Please note that the only difference is that the positional encoding module is discarded, since there does not exist any positional relationship between individual radiomic features.

Bootstrap Your Own Attention (BYOA): A Feedback Loop Module. Our main *roadblock* concerns how to generate robust radiomic features without pathology localization. On one hand, radiomic features are highly sensitive to the choice of local region of interest, for which we have no bounding box annotation. On the other hand, image features would benefit from the guidance of radiomics that encode important domain-specific quantitative features. The learning of image and radiomic features thus mutually depend on each other, forming a challenging chicken-and-egg problem.

To address this issue, we design **BYOA** to constitute an end-to-end feedback loop that can bootstrap accurate pathology localization from image data without any bounding box annotations (Fig. 3). BYOA contains two components: attention map generation and radiomic feature extraction.

- *Attention Map Generation.* Similar to the approach in Caron *et al.* [57], we extract self-attention of the `CLS` token from the heads of the last layer. RGT produces two `CLS` tokens from two branches, but the attention maps *only* come from the image branch. To generate bounding boxes for radiomic features extraction, we first apply a threshold on the learned self-attention maps. This threshold, controlling the percentage of most responsive pixels kept for further processing, will influence the size of the resulting bounding box and thus the quality of radiomic features. After thresholding the attention map, image processing steps including a maximum filter and five consecutive binary dilations are used to "grow" the region of interest and smooth boundaries. Then, connected-components labeling is performed, after which we find the "center of mass" of each component. If this center of mass pixel is in the top decile of intensity values, a bounding box is drawn around it according to the mean height and width of the known bounding box annotations for the given disease class of interest. Here, we utilize one kind of prior knowledge of different diseases, e.g. Cardiomegaly usually occurs in the heart area, and localized Pneumonia usually occurs in the lung area, and the information of the average bounding boxes of these diseases could be seen as one kind of free-available prior knowledge, which

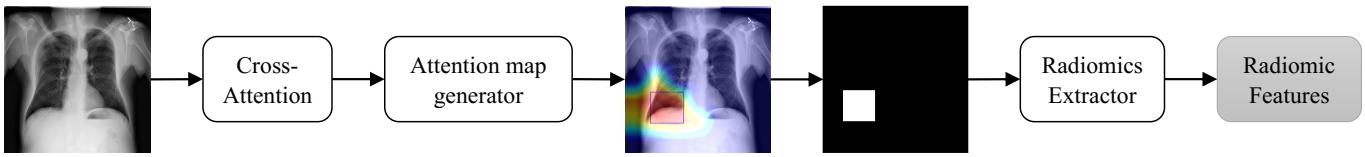

Fig. 3. Overview of our Bootstrap Your Own Attention (BYOA) module. For the input chest X-rays, we look at the self-attention of the CLS token of the Image branch on the heads of the final output of the cross attention module. Then we apply a threshold of 0.1, meaning we only keep the top 10% of pixels in the generated attention map, to produce bounding boxes. Then with the generated bounding boxes, we use the *Pyradiomics* tool to extract radiomic features from the region of interest.

could improve the accuracy of our model. And as one limitation of our method, for stable training, our method will generate per-class identically-sized bounding boxes. But during testing, we relaxed the setting of the bounding box generation.

- *Radiomic Features Extraction.* Given the original images and generated bounding boxes, we used Pyradiomics [7] to extract a variety of radiomic features, including 18 first-order features, 14 shape-based features, and 73 gray-level features (see Appendix for full list). For feature extraction, we adopt the default settings of PyRadiomics version 3.0.1, which includes no spatial resampling, discretization, rescaling, or normalization; this is not necessary, as input radiographs have already been min-max normalized as part of model preprocessing. All features are derived from the original image (no wavelet, Laplacian of Gaussian, or other filters are applied before feature extraction).

Cross-Attention Fusion Module. To aggregate global image information with local radiomics information, this fusion step involves the CLS token of the image branch and patch tokens of the radiomics branch, similarly, it also involves the CLS token of the radiomics branch and patch tokens of the image branch. As the CLS token is the aggregate representation of the branch, this interaction helps include information from multiple scales. Please refer to Chen *et al.* [35] for more details about the cross-attention mechanism.

C. Semi-Supervised Loss Function

In our framework, we aim to make the learned image features from the CLS token similar to the learned radiomic features in order to localize pathologies in the chest X-rays. As shown in Fig. 2, RGT is trained using the linear combination of the supervised classification and unsupervised contrastive losses. For the supervised classification, considering that the chest X-ray dataset is usually highly imbalanced, we adopt the Focal Loss [36]. For unsupervised contrastive learning, we use the cross-view contrastive loss [58].

Supervised Classification Focal Loss. We feed the output of the CLS tokens I_{cls} (from the image branch) and R_{cls} (from the radiomics branch) to a simple linear classifier. The supervised classification focal loss $\mathcal{L}_{fl}$ is defined as

$$\mathcal{L}_{fl} = \begin{cases} -\alpha\,(1-y')^{\gamma}\log y', & y=1 \\ -(1-\alpha)y'^{\gamma}\log\,(1-y')\,, & y=0 \end{cases} \tag{1}$$

The hyperparameter α allows us to give different importance to positive and negative examples, whereas γ is used to distinguish easy and hard samples , forcing the model to place more emphasis on difficult examples.

Unsupervised Cross-View Contrastive Loss. Our contrastive loss extends the normalized temperature scaled cross-entropy loss (NT-Xent). The difference is that we maximize agreement between two feature views extracted from different input formats, one from the image and the other from radiomic features.

Given an anchor chest X-ray in a minibatch, the positive sample will be its radiomic feature view, and the negative samples will be other chest X-rays (both image and radiomics views). Since the CLS token can be regarded as the representation of the input modality, we only need to maximize the agreement between each modality's CLS tokens. Suppose $I_{cls,k}$ and $R_{cls,k}$ are the $k-$th image features and radiomic features in the minibatch, respectively, and $sim(\cdot)$ the cosine similarity. Then the contrastive loss function $\mathcal{L}_{cl}$ is defined as

$$\mathcal{L}_{cl} = -\log \frac{\exp(sim(g_i(I_{cls,k}), g_r(R_{cls,k}))/\tau)}{\sum_{k=1}^{N}\exp(sim(g_i(I_{cls,k}), g_r(R_{cls,k}))/\tau)} \tag{2}$$

where τ is the temperature. The final contrastive loss is summed over all instances in the minibatch.

Overall, we treat RGT training as a weakly-supervised multi-task learning problem. In our chest X-ray setting, there exist two types of labels: disease class labels and pathology bounding box annotations. In our case, we *only* use the disease labels for training, even though the ultimate goal is to accurately localize those pathologies. Here, when we say "weakly-supervised" localization, we mean that we are able to localize pathologies only using supervision from whole-image disease labels. The combined loss function for supervised disease classification and unsupervised cross-view contrastive learning is as follows:

$$\mathcal{L} = (1-\lambda)\times\mathcal{L}_{cl} + \lambda\times\mathcal{L}_{fl} \tag{3}$$

IV. Experiments

A. Dataset and Protocol Setting

The NIH ChestXRay dataset [9] consists of 112,120 chest X-rays collected from 30,805 patients, where each image is labeled with one or more of 14 cardiopulmonary diseases. The labels are extracted from the associated radiology report using an automatic labeler [59] with a reported accuracy of 90%. For a subset of 880 images, the NIH dataset also provides bounding box localizations associated with eight disease classes: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. The remaining six diseases are diffuse in nature, meaning it is not clinically meaningful

to provide a "localization" for these pathologies. Since this study aims to develop a model for weakly supervised disease localization, we only proceed with the eight diseases which have ground truth bounding box annotations. Specifically, we only use the image-level disease labels for these eight focal diseases to train RGT, binning all other classes into the already provided "No Findings" category. A significant difference between our method and existing baseline methods for pathology localization [10], [60] is that our method does not require any training data related to the bounding box while others use some percentage of these images for training.

In our experiments, we followed the same protocol as in related studies [9], [10], randomly partitioning the dataset (excluding images with bounding box 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 train, validation, and test set.

B. Implementation Details

We build our image branch encoder based on PS-ViT [55], and apply their default hyperparameters for training. We use a shallower image encoder than the original PS-ViT, using 6 layers. For the radiomic branch encoder, since the radiomic features are already informative features, we use a small standard Transformer (2 layers) to learn representations of the radiomic features. We then add one more cross-attention layer to fuse the learned image features with the learned radiomic features. We set the batch size to 128 and train the model for 50 epochs. We used a cosine linear-rate scheduler with a linear warm-up of 5 epochs, an initial learning rate of 0.004, and a weight decay of 0.05. We downscale the images to 224×224 and normalize based on the mean and standard deviation of images in the ImageNet training set. We also augment the training data with random horizontal flipping. During the evaluation, we resize the image to 256×256 and take the center crop 224×224 as the input.

C. Pathology Localization

The NIH Chest X-ray dataset contains 880 images labeled by radiologists with bounding box information, which we use to evaluate the performance of RGT for pathology localization. Many prior works [10], [60] have used a fraction of ground truth (GT) bounding boxes for training and evaluated their system on the remaining examples. Unlike these approaches, RGT *uses no bounding box annotations during training*, only using the subset of bounding box-annotated images for evaluation. Table I presents our evaluation results on all 880 images. We used [9] as our baseline to compare our localization results since it follows the same experimental setting of weakly supervised training on only disease labels.

1) Evaluation Metric: For localization, we evaluated our detected regular rectangular regions against the annotated bounding boxes, using a thresholded **IoU accuracy**, following Wang *et al.* [9]. Our localization results are only calculated for the 880 images that have ground truth annotation for 8 diseases. To compute IoU accuracy, the localization is defined as "correct" only if the observed IoU between the predicted and ground truth localization exceeds a fixed IoU threshold, T(IoU). We evaluated RGT for different thresholds ranging from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7} as shown in Table I.

2) Comparison with Prior Works: We compared disease localization accuracy under varying IoU with baselines following the same training setting as RGT (Table I). Unlike other baselines [10], [60] that use a portion of 880 images for evaluation (because they need the remaining data for training), we used all 880 annotated images for evaluation. Therefore, no k-fold cross-validation for localization was performed. RGT average localization performance across 8 diseases is considerably higher than the baseline under all IoU thresholds. When the IoU threshold is set to 0.1, RGT outperforms the baseline [9] in the Cardiomegaly, Infiltration, Mass, and Pneumonia classes. Even with higher thresholds, our model is superior to the baseline. For example, when evaluated at T(IoU) = 0.5, our "Cardiomegaly" accuracy is 32%, while the reference model achieves 18%. Similarly, our "Pneumonia" accuracy is 12%, while the reference model reaches 3% accuracy. Note that some diseases can appear in multiple locations, but the ground truth might have mentioned only one such location. This can significantly impact the accuracy at high thresholds.

3) Discussion of Visualization: More importantly, we also include our own trained ViT as an additional baseline here. The quantitative results above demonstrate that, compared to the standard ViT, the additional radiomics branch and BYOA module enable RGT to learn more accurate and fine-grained pathology localizations. Example visualizations of localization results of both ViT and RGT can be seen in Fig. 4. We can observe that RGT produces qualtitatively more accurate localizations than ViT for all diseases, but particularly Atelectasis, Cardiomegaly, Infiltration, Pneumonia, and Pneumothorax. Visualizations for most diseases reveal that both models often attend to regions outside the clinically relevant region of interest. However, RGT consistently attends to a smaller number of "extraneous" pixels than the standard ViT. Further, RGT always contains a significant portion of the ground truth localized region, while the ViT attention map does not – for example, see Nodule, Pneumonia, and Pneumothorax.

D. Pathology Classification

Pathology classification for chest X-rays is a multi-label classification problem. The objective is to assign one or more labels (among 8 cardiopulmonary diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax) to each input image at inference time. We compared RGT with related reference approaches, which represent state-of-the-art disease classifcation performance on the NIH ChestXRay dataset. For RGT, we report the average AUC of 3 runs to show the robustness of our model.

1) Evaluation Metric: We used Area under the Receiver Operating Characteristic Curve (AUC) to estimate the performance of our model [66]. A higher AUC score implies a model that is more capable of discriminating between classes. We also provide mean AUC across all the classes to highlight the overall performance of our model.

TABLE I

WEAKLY SUPERVISED PATHOLOGY LOCALIZATION RESULTS ON THE NIH CHESTXRAY DATASET AS MEASURED BY IOU ACCURACY AT A FIXED THRESHOLD. PLEASE NOTE THAT SINCE RGT WAS SOLELY SUPERVISED BY DISEASE CLASS LABELS (NOT PATHOLOGY LOCALIZATIONS), WE ONLY COMPARE LOCALIZATION PERFORMANCE WITH PREVIOUS METHODS FOLLOWING THE SAME SETTING FOR FAIR EVALUATION.

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
0.1	Wang *et al.* [9]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.569
	ViT	0.58	0.91	0.61	0.77	0.44	0.11	0.75	0.25	0.553
	RGT	0.61	0.95	0.65	0.82	0.50	0.13	0.79	0.28	**0.591**
0.2	Wang *et al.* [9]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.371
	ViT	0.38	0.85	0.39	0.55	0.24	0.01	0.51	0.15	0.385
	RGT	0.41	0.91	0.41	0.59	0.26	0.05	0.57	0.19	**0.424**
0.3	Wang *et al.* [9]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.221
	ViT	0.20	0.45	0.19	0.32	0.06	0.00	0.21	0.02	0.181
	RGT	0.28	0.79	0.22	0.38	0.12	0.01	0.41	0.05	**0.283**
0.4	Wang *et al.* [9]	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.115
	ViT	0.10	0.21	0.03	0.05	0.02	0.00	0.04	0.00	0.056
	RGT	0.17	0.54	0.13	0.18	0.07	0.01	0.26	0.02	**0.173**
0.5	Wang *et al.* [9]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.061
	ViT	0.05	0.15	0.01	0.04	0.02	0.00	0.03	0.00	0.034
	RGT	0.08	0.32	0.05	0.09	0.05	0.00	0.12	0.01	**0.090**
0.6	Wang *et al.* [9]	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.029
	ViT	0.01	0.03	0.01	0.01	0.01	0.00	0.01	0.00	0.010
	RGT	0.02	0.15	0.03	0.04	0.03	0.00	0.06	0.00	**0.041**
0.7	Wang *et al.* [9]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.011
	ViT	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.001
	RGT	0.01	0.04	0.01	0.02	0.01	0.00	0.03	0.00	**0.015**

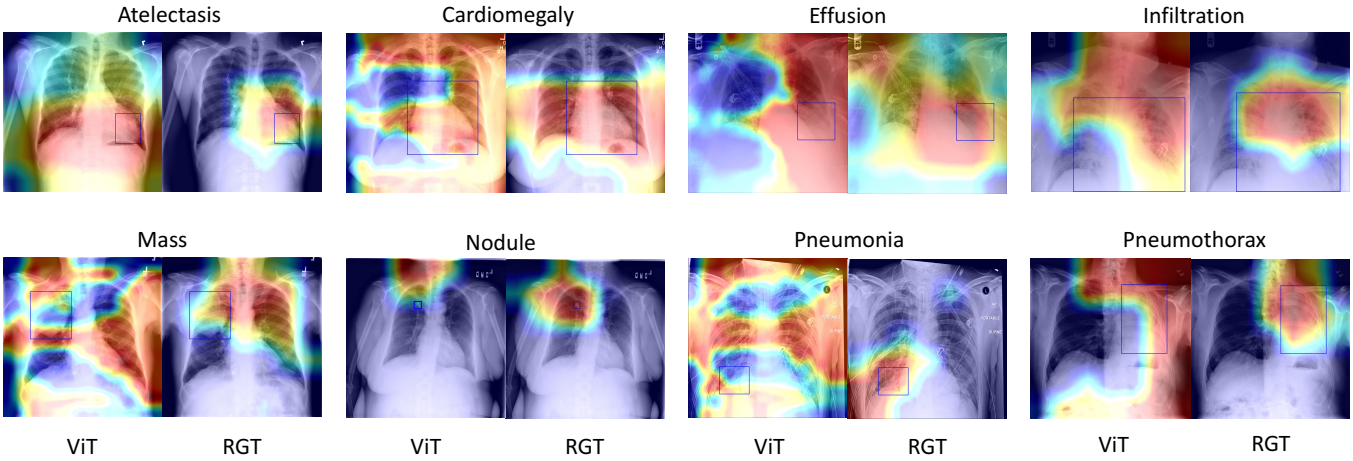

Fig. 4. Example visualizations of pathology localization when evaluated on the 880 NIH ChestXRay images with bounding box annotations. The attention maps are generated from the self-attention maps of the `CLS` token. The ground-truth bounding boxes are shown in blue. The left image in each pair is the localization result of ViT [28], and the right one is our localization results obtained by RGT. All examples are positive for the corresponding disease labels. Best viewed in color.

2) Comparison with Prior Works: AUC scores for each disease and mean AUC across eight diseases are presented in Table II. We not only compared RGT with previous CNN-based state-of-the-art (SOTA) models, but also several Transformer-based models. We find that RGT outperformed all baseline approaches with respect to mean AUC across all diseases; specifically, RGT reached 0.839 mean AUC, outperforming the previous SOTA for disease classification [8] by a margin of 0.011. When considering classification performance on individual disease classes, RGT also achieved best performance on four of the eight classes. Our proposed model outperformed the next-best baseline by a margin 0.13 AUC for Infiltration, 0.09 for Nodule, and 0.02 for Mass. Compared to the Transformer-based models, the key difference is that we utilize the extracted radiomic features for disease prediction, improving the classification accuracy and enriching the model's interpretability due to the utilization of handcrafted radiomic features. Please note that Liu *et al.* [11] used 5-fold cross-validation in their model evaluation. While the problem settings are very similar, the evaluation schemes are so

TABLE II

PATHOLOGY CLASSIFICATION RESULTS FOR CNN- AND TRANSFORMER-BASED METHODS ON THE NIH CHESTXRAY DATASET, AS MEASURED BY AUC. FOR EACH COLUMN, **BOLD** VALUES DENOTE THE BEST RESULTS FOR THE GIVEN DISEASE CLASS. FOR RGT, THE AVERAGE AUC PER CLASS IS PRESENTED, WITH THE STANDARD DEVIATION IN PARENTHESES, ACROSS THREE TRAINING RUNS WITH DIFFERENT RANDOM INITIALIZATIONS.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
CNN									
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.* [61]	0.73	0.84	0.79	0.67	0.73	0.69	0.72	0.85	0.753
Yao *et al.* [62]	0.77	0.90	0.86	0.70	0.79	0.72	0.71	0.84	0.786
Rajpurkar *et al.* [8]	0.82	0.91	**0.88**	0.72	0.86	0.78	0.76	**0.89**	0.828
Kumar *et al.* [63]	0.76	0.91	0.86	0.69	0.75	0.67	0.72	0.86	0.778
Liu *et al.* [11]	0.79	0.87	**0.88**	0.69	0.81	0.73	0.75	**0.89**	0.801
Seyyed *et al.* [64]	0.81	**0.92**	0.87	0.72	0.83	0.78	0.76	0.88	0.821
Han *et al.* [65]	**0.83**	**0.92**	0.87	0.76	0.85	0.76	0.77	0.86	0.828
Transformer									
ViT	0.74	0.78	0.81	0.72	0.70	0.66	0.65	0.76	0.728
CrossViT	0.69	0.71	0.72	0.72	0.74	0.79	**0.82**	0.88	0.759
PS-ViT	0.75	0.81	0.82	0.73	0.79	0.73	0.69	0.81	0.766
RGT (ours)	0.80	**0.92**	0.78	**0.86**	**0.88**	**0.88**	0.79	0.81	**0.839**
	(±0.02)	(±0.00)	(±0.01)	(±0.01)	(±0.02)	(±0.00)	(±0.01)	(±0.02)	–

different that a direct comparison of this work to RGT would be inappropriate.

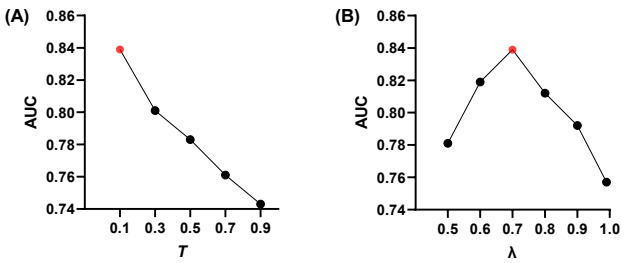

Fig. 5. Effect of (A) varying T in attention map generation and (B) varying λ in Equation (3) on pathology classification for the NIH ChestXRay dataset.

3) Effect of Attention Map Threshold: We investigate the impact of the threshold (T), used in the process of attention map generation, on the performance of RGT for disease classification. Fig. 5A summarizes the AUC comparison of RGT for different values of T. Higher values of T imply smaller bounding boxes from which to extract radiomic features. During our experiments, we found that RGT performs better on the disease classification task when larger bounding boxes are generated. Since radiomic features are typically computed for highly localized – often small – regions of interest, this was originally an unintuitive finding. There appears to be a tradeoff between bounding box size and the resulting performance on the disease classification and localization tasks. Specifically, extracting smaller boxes that are accurately localized and ignore as much background signal as possible should lead to more robust and useful radiomic features. However, attending to smaller regions of the image comes at the expense of decreasing the "receptive field" of learned global image features, thus degrading the quality of the classification task. This observation emphasizes the difference between disease classification and localization tasks: global information aids classification while rich local information aids localization.

4) Effect of Contrastive Learning: We also investigated the impact of the unsupervised contrastive loss on RGT's disease classification ability. Specifically, we evaluate the performance of RGT for disease classification by varying λ in equation (3). Fig. 5B summarizes the AUC comparison of RGT for different values of λ. Higher values of λ implies lower weight to contrastive loss. During our experiments, we found that RGT performs worse when small weight (1%) is given to the contrastive loss. RGT's performance improves when we increase contrastive loss weight, but after a certain point ($\lambda = 0.7$), performance considerably decreases. This confirms our hypothesis that both contrastive and focal losses are important, but that care must be taken to properly balance the objectives. The supervised classification and unsupervised constrastive losses enable RGT to learn both disease-level and patient-level discriminative visual features.

E. System Usability Study

We hired two radiologist experts to validate the usefulness of RGT's disease localizations; one expert has 3.5 years of experience, while the other has 5. For this additional study, we randomly selected 10 images from the NIH ChestXRay dataset that did not have ground truth bounding box annotations. We then used RGT to predict the disease classification and localization visualization. Finally, we asked two radiologists to provide their own pathology localization for each image. Each radiologist was instructed to draw a rectangular bounding box around the clinically relevant region of interest within 90 seconds. Results can be seen in Fig. 6. Each image contains two human-annotated bounding boxes (red is Expert 1, and blue is Expert 2) and the extracted attention map from our RGT.

For disease classification, the two radiologists agreed with RGT's prediction for all ten cases. For the localization task, we

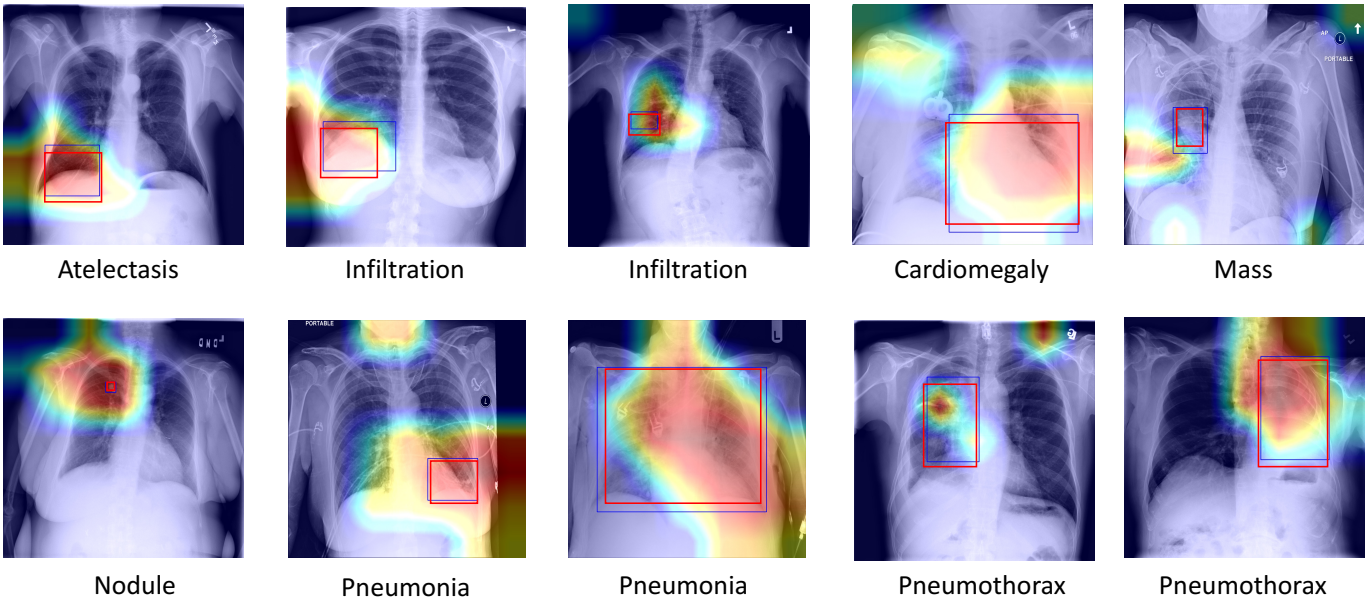

Fig. 6. System Usability Study. Visualizations of pathology localization come from 10 randomly selected chest X-rays from the NIH ChestXRay dataset that do not have ground truth localization annotations. Saliency heat maps are generated from the self-attention maps of the `CLS` token from our trained RGT model. The red and blue represent the pathology localizations provided by two radiologists, who were instructed to draw a rectangular bounding box around the most salient image region in 90 seconds.

observe that the inter-rater consistency between two radiologists is very high, suggesting that they clearly agreed on the most salient image region. Overall, the radiologists found the RGT attention maps to significantly overlap with their own localizations, demonstrating the usefulness of our approach. With the exception of the "Mass" example, there is a strong agreement between the most responsive pixels in RGT's heat map and the radiologists' annotations.

F. Limitations and Discussion

There exist two main limitations to this approach. One limitation is the fact that self-attention provides only a coarse approximation of salient regions unless trained on extremely large amounts of data (e.g., see DINO [57]). Without this scale of chest radiography data available, other principled methods for saliency visualization may provide more fine-grained localizations for radiomics extraction than the native self-attention of our proposed RGT architecture. For example, the Anchors approach of Ribeiro *et al.* [21] or other input space visualization methods like LIME [20], SHAP [24], and deep Taylor decomposition [22] could be used in place of our proposed heatmap generation process. Future work will consider adapting such approaches to generate more accurate localizations for improved radiomics feature extraction, and thus better downstream disease classification and localization.

Another limitation of this approach is the fact that fixed-sized bounding boxes per target disease class are generated during RGT training. This would, for example, make it difficult to distinguish the visual presentation of diffuse vs. localized pneumonia. However, this can be alleviated with finer granularity in the image-level disease labels used to train RGT; for instance, if "diffuse" and "localized" pneumonia were distinct class labels, then RGT would be able to provide visually distinct localizations of the two conditions. Future work may involve a module that learns the optimal bounding box dimensions for each disease in an unsupervised manner. Alternatively, the adoption of other saliency visualization methods instead of RGT's self-attention – as explained in the previous paragraph – may resolve this limitation of fixed-size bounding boxes per disease class.

V. Conclusion

In this paper, we propose a radiomics-guided cross-attention Transformer, RGT, that can jointly localize and classify abnormalities in chest X-rays without supervision from bounding box annotations. Our approach differs from previous related studies in the choice of a unified Transformer architecture, the use of radiomic features, and a feedback loop for image and radiomic features to mutually interact during the training process. This work aims to bring the field of computer-aided diagnosis closer to clinical practice by making domain-specific quantitative features (in the form of radiomics) more accessible to automated medical image analysis tools, with the hope that this will increase the model's interpretability. Experimental results demonstrate that our method outperforms state-of-the-art algorithms in this weakly supervised setting, particularly for disease localization, where our method can generate more accurate and clinically useful bounding boxes.

References

[1] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *arXiv preprint arXiv:1612.07003*, 2016.
[2] V. Parekh and M. A. Jacobs, "Radiomics: a new application from established techniques," *Expert Review of Precision Medicine and Drug Development*, vol. 1, no. 2, pp. 207–226, 2016.
[3] F. Shi, L. Xia, F. Shan *et al.*, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. arxiv e-prints [preprint] 2020."

[4] A. Saygılı, "A new approach for computer-aided detection of coronavirus (covid-19) from ct and x-ray images using machine learning methods," *Applied Soft Computing*, vol. 105, p. 107323, 2021.

[5] X. Bai, C. Fang, Y. Zhou, S. Bai, Z. Liu, L. Xia, Q. Chen, Y. Xu, T. Xia, S. Gong *et al.*, "Predicting covid-19 malignant progression with ai techniques," 2020.

[6] B. Ghosh, N. Kumar, N. Singh, A. K. Sadhu, N. Ghosh, P. Mitra, and J. Chatterjee, "A quantitative lung computed tomography image feature for multi-center severity assessment of covid-19," *medRxiv*, 2020.

[7] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[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] 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.

[10] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8290–8299.

[11] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 632–10 641.

[12] 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.

[13] Y. Wang, K. Zheng, C.-T. Cheng, X.-Y. Zhou, Z. Zheng, J. Xiao, L. Lu, C.-H. Liao, and S. Miao, "Knowledge distillation with adaptive asymmetric label sharpening for semi-supervised fracture detection in chest x-rays," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 599–610.

[14] K. Yu, S. Ghosh, Z. Liu, C. Deible, and K. Batmanghelich, "Anatomy-guided weakly-supervised abnormality localization in chest x-rays," *arXiv preprint arXiv:2206.12704*, 2022.

[15] T. B. Chandra, B. K. Singh, and D. Jain, "Disease localization and severity assessment in chest x-ray images using multi-stage superpixels classification," *Computer Methods and Programs in Biomedicine*, p. 106947, 2022.

[16] C. Fernando, S. Kolonne, H. Kumarasinghe, and D. Meedeniya, "Chest radiographs classification using multi-model deep learning: A comparative study," in *2022 2nd International Conference on Advanced Research in Computing (ICARC)*. IEEE, 2022, pp. 165–170.

[17] K. Kumarasinghe, S. Kolonne, K. Fernando, and D. Meedeniya, "U-net based chest x-ray segmentation with ensemble classification for covid-19 and pneumonia." *International Journal of Online & Biomedical Engineering*, vol. 18, no. 7, 2022.

[18] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. De la Torre Díez, and G. Marques, "Chest x-ray analysis empowered with deep learning: A systematic review," *Applied Soft Computing*, p. 109319, 2022.

[19] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Transactions on Image Processing*, vol. 29, pp. 225–236, 2019.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[21] ——, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[22] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern recognition*, vol. 65, pp. 211–222, 2017.

[23] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[30] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[31] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, "Do transformers really perform bad for graph representation?" *arXiv preprint arXiv:2106.05234*, 2021.

[32] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," 2021.

[33] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[34] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once–multi-modal fusion transformer for video retrieval," *arXiv preprint arXiv:2112.04446*, 2021.

[35] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *arXiv preprint arXiv:2103.14899*, 2021.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[37] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[38] H. Nasief, C. Zheng, D. Schott, W. Hall, S. Tsai, B. Erickson, and X. A. Li, "A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer," *NPJ precision oncology*, vol. 3, no. 1, pp. 1–10, 2019.

[39] A. Eilaghi, S. Baig, Y. Zhang, J. Zhang, P. Karanicolas, S. Gallinger, F. Khalvati, and M. A. Haider, "Ct texture features are associated with overall survival in pancreatic ductal adenocarcinoma–a quantitative analysis," *BMC medical imaging*, vol. 17, no. 1, pp. 1–7, 2017.

[40] X. Chen, K. Oshima, D. Schott, H. Wu, W. Hall, Y. Song, Y. Tao, D. Li, C. Zheng, P. Knechtges *et al.*, "Assessment of treatment response during chemoradiation therapy for pancreatic cancer based on quantitative radiomic analysis of daily cts: An exploratory study," *PLoS One*, vol. 12, no. 6, p. e0178961, 2017.

[41] P. Huang, S. Park, R. Yan, J. Lee, L. C. Chu, C. T. Lin, A. Hussien, J. Rathmell, B. Thomas, C. Chen *et al.*, "Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study," *Radiology*, vol. 286, no. 1, pp. 286–295, 2018.

[42] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," *arXiv preprint arXiv:1809.04698*, 2018.

[43] P. Alongi, A. Stefano, A. Comelli, R. Laudicella, S. Scalisi, G. Arnone, S. Barone, M. Spada, P. Purpura, T. V. Bartolotta *et al.*, "Radiomics analysis of 18f-choline pet/ct in the prediction of disease outcome in high-risk prostate cancer: An explorative study on machine learning feature classification in 94 patients," *European Radiology*, vol. 31, no. 7, pp. 4595–4605, 2021.

[44] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[45] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

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

[47] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.
[48] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
[49] X. Wang, S. Han, Y. Chen, D. Gao, and N. Vasconcelos, "Volumetric attention for 3d medical image segmentation and detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 175–184.
[50] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[51] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.
[52] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *arXiv preprint arXiv:2102.10662*, 2021.
[53] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, "Vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification," *arXiv preprint arXiv:2104.07235*, 2021.
[54] Y. Han, C. Chen, A. H. Tewfik, Y. Ding, and Y. Peng, "Pneumonia detection on chest x-ray using radiomic features and contrastive learning," *arXiv preprint arXiv:2101.04269*, 2021.
[55] X. Yue, S. Sun, Z. Kuang, M. Wei, P. Torr, W. Zhang, and D. Lin, "Vision transformer with progressive sampling," *arXiv preprint arXiv:2108.01684*, 2021.
[56] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
[57] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.
[58] 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.
[59] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports." in *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2017, 2018, pp. 188–196.
[60] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
[61] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.
[62] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.
[63] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 546–552.
[64] L. Seyyed-Kalantari, G. Liu, M. McDermott, and M. Ghassemi, "Chexclusion: Fairness gaps in deep chest x-ray classifiers," *arXiv preprint arXiv:2003.00827*, 2020.
[65] Y. Han, C. Chen, L. Tang, M. Lin, A. Jaiswal, S. Wang, A. Tewfik, G. Shih, Y. Ding, and Y. Peng, "Using radiomics as prior knowledge for thorax disease classification and localization in chest x-rays," *arXiv preprint arXiv:2011.12506*, 2020.
[66] FawcettTom, "An introduction to roc analysis," *Pattern Recognition Letters*, 2006.

APPENDIX

TABLE III

NAME AND TYPE OF ALL 107 RADIOMIC FEATURES USED IN THIS STUDY. FEATURE NAMES ARE DEFINED BY PYRADIOMICS, AND DETAILED DESCRIPTIONS OF EACH FEATURE CAN BE FOUND IN THE PYRADIOMICS DOCUMENTATION: HTTPS://PYRADIOMICS.READTHEDOCS.IO/EN/LATEST/FEATURES.HTML. GLCM = GRAY LEVEL CO-OCCURRENCE MATRIX; GLDM = GRAY LEVEL DEPENDENCE MATRIX, GLRLM = GRAY LEVEL RUN-LENGTH MATRIX; GLSZM = GRAY LEVEL SIZE ZONE MATRIX; NGTDM = NEIGHBORING GRAY TONE DISTANCE MATRIX.

PyRadiomics Feature	Feature Type	PyRadiomics Feature	Feature Type
Elongation	Shape	SumEntropy	GLCM
Flatness	Shape	SumSquares	GLCM
LeastAxisLength	Shape	DependenceEntropy	GLDM
MajorAxisLength	Shape	DependenceNonUniformity	GLDM
Maximum2DDiameterColumn	Shape	DependenceNonUniformityNormalized	GLDM
Maximum2DDiameterRow	Shape	DependenceVariance	GLDM
Maximum2DDiameterSlice	Shape	GrayLevelNonUniformity	GLDM
Maximum3DDiameter	Shape	GrayLevelVariance	GLDM
MeshVolume	Shape	HighGrayLevelEmphasis	GLDM
MinorAxisLength	Shape	LargeDependenceEmphasis	GLDM
Sphericity	Shape	LargeDependenceHighGrayLevelEmphasis	GLDM
SurfaceArea	Shape	LargeDependenceLowGrayLevelEmphasis	GLDM
SurfaceVolumeRatio	Shape	LowGrayLevelEmphasis	GLDM
VoxelVolume	Shape	SmallDependenceEmphasis	GLDM
10Percentile	First-Order	SmallDependenceHighGrayLevelEmphasis	GLDM
90Percentile	First-Order	SmallDependenceLowGrayLevelEmphasis	GLDM
Energy	First-Order	GrayLevelNonUniformity	GLRLM
Entropy	First-Order	GrayLevelNonUniformityNormalized	GLRLM
InterquartileRange	First-Order	GrayLevelVariance	GLRLM
Kurtosis	First-Order	HighGrayLevelRunEmphasis	GLRLM
Maximum	First-Order	LongRunEmphasis	GLRLM
MeanAbsoluteDeviation	First-Order	LongRunHighGrayLevelEmphasis	GLRLM
Mean	First-Order	LongRunLowGrayLevelEmphasis	GLRLM
Median	First-Order	LowGrayLevelRunEmphasis	GLRLM
Minimum	First-Order	RunEntropy	GLRLM
Range	First-Order	RunLengthNonUniformity	GLRLM
RobustMeanAbsoluteDeviation	First-Order	RunLengthNonUniformityNormalized	GLRLM
RootMeanSquared	First-Order	RunPercentage	GLRLM
Skewness	First-Order	RunVariance	GLRLM
TotalEnergy	First-Order	ShortRunEmphasis	GLRLM
Uniformity	First-Order	ShortRunHighGrayLevelEmphasis	GLRLM
Variance	First-Order	ShortRunLowGrayLevelEmphasis	GLRLM
Autocorrelation	GLCM	GrayLevelNonUniformity	GLSZM
ClusterProminence	GLCM	GrayLevelNonUniformityNormalized	GLSZM
ClusterShade	GLCM	GrayLevelVariance	GLSZM
ClusterTendency	GLCM	HighGrayLevelZoneEmphasis	GLSZM
Contrast	GLCM	LargeAreaEmphasis	GLSZM
Correlation	GLCM	LargeAreaHighGrayLevelEmphasis	GLSZM
DifferenceAverage	GLCM	LargeAreaLowGrayLevelEmphasis	GLSZM
DifferenceEntropy	GLCM	LowGrayLevelZoneEmphasis	GLSZM
DifferenceVariance	GLCM	SizeZoneNonUniformity	GLSZM
Id	GLCM	SizeZoneNonUniformityNormalized	GLSZM
Idm	GLCM	SmallAreaEmphasis	GLSZM
Idmn	GLCM	SmallAreaHighGrayLevelEmphasis	GLSZM
Idn	GLCM	SmallAreaLowGrayLevelEmphasis	GLSZM
Imc1	GLCM	ZoneEntropy	GLSZM
Imc2	GLCM	ZonePercentage	GLSZM
InverseVariance	GLCM	ZoneVariance	GLSZM
JointAverage	GLCM	Busyness	NGTDM
JointEnergy	GLCM	Coarseness	NGTDM
JointEntropy	GLCM	Complexity	NGTDM
MCC	GLCM	Contrast	NGTDM
MaximumProbability	GLCM	Strength	NGTDM
SumAverage	GLCM		