Received: 19 October 2020 | Revised: 12 July 2021 | Accepted: 27 July 2021

DOI: 10.1002/mp.15185

RESEARCH ARTICLE

MEDICAL PHYSICS

Weakly supervised pneumonia localization in chest X-rays using generative adversarial networks

Krishna Nand Keshavamurthy[1,2] | Carsten Eickhoff[1] | Krishna Juluru[2]

[1] Brown University, Providence, Rhode Island, USA

[2] Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, New York, USA

Correspondence
Krishna Nand Keshavamurthy, Memorial Sloan Kettering Cancer Center Radiology, 321 E 61St Street, G26, New York, NY, 10065 USA.
Email: keshavak@mskcc.org

Carsten Eickhoff and Krishna Juluru should be considered joint senior authors.

Funding information
NIH/NCI Cancer Center Support, Grant/Award Number: P30 CA008748

Abstract

Purpose: Automatic localization of pneumonia on chest X-rays (CXRs) is highly desirable both as an interpretive aid to the radiologist and for timely diagnosis of the disease. However, pneumonia's amorphous appearance on CXRs and complexity of normal anatomy in the chest present key challenges that hinder accurate localization. Existing studies in this area are either not optimized to preserve spatial information of abnormality or depend on expensive expert-annotated bounding boxes. We present a novel generative adversarial network (GAN)-based machine learning approach for this problem, which is weakly supervised (does not require any location annotations), was trained to retain spatial information, and can produce pixel-wise abnormality maps highlighting regions of abnormality (as opposed to bounding boxes around abnormality).

Methods: Our method is based on the Wasserstein GAN framework and, to the best of our knowledge, the first application of GANs to this problem. Specifically, from an abnormal CXR as input, we generated the corresponding *pseudo* normal CXR image as output. The *pseudo* normal CXR is the "hypothetical" normal, if the same abnormal CXR were not to have any abnormalities. We surmise that the difference between the *pseudo* normal and the abnormal CXR highlights the pixels suspected to have pneumonia and hence is our output abnormality map. We trained our algorithm on an "unpaired" data set of abnormal and normal CXRs and did not require any location annotations such as bounding boxes/segmentations of abnormal regions. Furthermore, we incorporated additional prior knowledge/constraints into the model and showed that they help improve localization performance. We validated the model on a data set consisting of 14 184 CXRs from the Radiological Society of North America pneumonia detection challenge.

Results: We evaluated our methods by comparing the generated abnormality maps with radiologist annotated bounding boxes using receiver operating characteristic (ROC) analysis, image similarity metrics such as normalized cross-correlation/mutual information, and abnormality detection rate. We also present visual examples of the abnormality maps, covering various scenarios of abnormality occurrence. Results demonstrate the ability to highlight regions of abnormality with the best method achieving an ROC area under the curve (AUC) of 0.77 and a detection rate of 85%. The GAN tended to perform better as prior knowledge/constraints were incorporated into the model.

Conclusions: We presented a novel GAN based approach for localizing pneumonia on CXRs that (1) does not require expensive hand annotated location ground truth; and (2) was trained to produce abnormality maps at the pixel level as opposed to bounding boxes. We demonstrated the efficacy of our methods via quantitative and qualitative results.

KEYWORDS
Generative adversarial networks GAN, pneumonia localization, weakly supervised

1 | INTRODUCTION

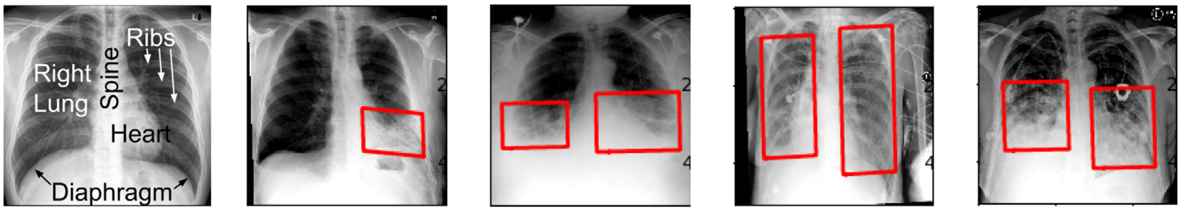
Pneumonia is a form of acute lung infection that affects millions of people worldwide annually, and is responsible for over 50 000 deaths every year in the U.S. alone.[1] Chest X-rays (CXRs) are one of the most common and effective methods for diagnosing pneumonia.[2] While easy to acquire, CXRs are difficult to interpret due to the amorphous presentation of infection and the complexity of adjacent normal anatomy.[3] In a busy practice, radiologists may be asked to interpret hundreds of CXRs every day. The high volume of examinations causes delays in result reporting, and fatigue that increases the risk of missing subtle findings.[4–6] Automated tools that can help in triaging examinations and in highlighting regions of abnormality within these examinations can help improve the timeliness and accuracy of the CXR reports.[7–9]

Pneumonia usually manifests as one or more areas of increased opacity on a CXR.[10] However, automated localization is made difficult by a number of factors including (1) hazy/amorphous appearance; (2) lack of shape priors; (3) complexity of lung anatomy and subtleties of the findings; (4) variations in the appearance of the CXRs due to positioning of the patient and depth of inspiration;[11] (5) superimposition of the anatomy introduced by planar technique; and (6) presence of various other kinds of lung abnormalities. Figure 1 presents some examples to illustrate the problem. Several studies have reported interobserver disagreement[12–15] between radiologists and interpretation error[6,16,17] in detection of pneumonia from CXRs, which can potentially result in delayed or missed diagnosis.[18] In practice, in addition to the CXRs, radiologists may also review the clinical history, vital signs, and laboratory exams of the patients to make an accurate assessment. Hence, automated localization of pneumonia from CXR alone is a challenging task.

Related literature: The last decade has witnessed tremendous success in the application of deep learning (DL) methods to image analysis problems. Although there are several prior studies that report neural network (NN)-based classification of CXRs with pneumonia,[19] there are few that can localize the disease at the pixel level. These can broadly be classified into ones that (1) use the final feature maps of an NN classifier to derive a heat map that highlights the regions/pixels of abnormality, and ones that (2) predict a bounding box (BB) around the abnormality. We summarize these approaches in the next few paragraphs.

The first class of methods is weakly supervised and uses the information in the final layers of a neural network (NN) classifier to localize regions of interest. A popular subset of such approaches is based on the so-called class activation maps (CAM).[20] In CAM, the second-to-last-layer feature maps of an NN classifier are used to construct class-specific activation maps, with the expectation that they capture the pixel-level differences between the classes. Irvin et al[21] use this approach on a convolutional NN (CNN) classification model (which they call the *CheXNet*) and present heat maps to visualize the regions most indicative of the disease. Wang et al[22] take a similar approach and use the global average pooling layer of a classification CNN model for localizing pneumonia. While easier to implement, CAM and related approaches are optimized for classification rather than location preserving spatial information. As a result, CAM-based approaches may perform suboptimally on localization tasks and may miss spatial regions of abnormality.[23]

The second class of methods is supervised, relying on BBs marked around the regions of abnormality for training. In these approaches, classification and BB prediction is carried out simultaneously using models trained on both class labels and hand-annotated BBs.[24–26] It is important to note that these approaches predict a BB around the abnormality rather than a detailed pixel-level abnormality map. Li et al[24] present such an approach, where they additionally incorporate multiple instance learning (MIL)[27] for abnormality localization. Taghanaki et al[25] also use a similar approach with added latent space modeling and attention



(a) Normal CXR

(b) Single-site abnormality

(c) Multisite abnormality

(d) Diffuse abnormality

(e) Nodular abnormality

FIGURE 1 Exemplary normal (a) and abnormal CXRs with pneumonia (b)–(e). Anatomical structures are labeled in the normal CXR; abnormal regions are marked by red bounding boxes in the abnormal CXRs. Notice pneumonia's amorphous appearance, varying in the number of sites of occurrence, size, shape, intensity, and textural patterns in the abnormal CXRs. Also notice the lack of a clear boundary between the abnormal region and adjacent normal anatomy such as the heart (b,c), diaphragm (b,c,e), and lung walls (b,c). A key challenge is to discern such subtle differences

mechanism. A key drawback of these approaches is their dependence on expensive radiologist annotated BB labels. Furthermore, known variability between radiologists in the interpretation of chest radiographs[12,13] presents additional challenges for ground truth preparation, algorithm validation, and model generalization.

As a final note on prior work, we briefly describe the Radiological Society of North America (RSNA) pneumonia detection challenge, which was organized recognizing the importance and difficulty of the problem.[28] The goal of the challenge was to build an algorithm to automatically predict BBs around lung opacities using radiologist-annotated BB labels for training. Although the competing groups experimented with a multitude of NN-based algorithms, most of the top performing models were variants of the classification and BB localization approach described above and hence have the same drawbacks. It is also interesting to note that the algorithm designs were mostly driven by iterative empirical evidence and depended on heavy data-/model-specific optimizations. For example, the winning entry used an ensemble of 50 neural network models in addition to custom data pre-/postprocessing.[29]

Summary of the problem: There is a need for pneumonia localization algorithms that are (1) optimized for location-preserving abnormality information unlike CAM-based approaches; (2) do not depend on expensive ground truth location annotations unlike BB approaches; (3) can produce detailed pixel-wise abnormality maps, as opposed to a BB around the abnormality; and finally (4) do not rely on heavy data-/model-centric optimizations, which can pose substantial challenges for reproducibility and generalizability.

Our contribution: In this work, we present a novel generative adversarial network (GAN)-based machine learning approach for localizing pneumonia on CXRs. Our method is inspired by the work of Zhang et al[30] and Baumgartner et al[23] and, to the best of our knowledge, the first application of GANs to this problem. Specifically, we pose the problem as an image-to-image translation task and predict a *pseudo* normal CXR from an abnormal CXR using a Wasserstein GAN-based framework. The *pseudo* normal CXR is the "hypothetical" normal, if the same abnormal CXR were not to have any abnormalities. We surmise that the difference between the *pseudo* normal and the abnormal CXR captures the characteristics of abnormalities, and hence, is our output abnormality map. Furthermore, we explore sequentially incorporating additional prior knowledge/constraints into the GAN model, such as (1) being able to preserve the normal images "*as is*" when fed as input to the model and (2) registering all the images to a common reference space to constrain anatomical variability.

We would like to note some of the key features of our method: (1) it is weakly supervised, in that it does not depend on any location annotations; (2) is trained on an "unpaired" data set of abnormal and normal CXRs (i.e., there is no normal CXR corresponding to an abnormal CXR and vice versa); and (3) is trained to preserve abnormality location information (unlike CAM) and can produce detailed pixel-wise abnormality maps (unlike BB approaches).

Finally, we evaluated our methods by measuring the overlap/similarity between our abnormality maps and radiologist annotated ground truth BBs. We compare our approaches to CAM[20] as well as to adding BB supervision to our methods. Quantitative metrics for comparison included receiver operating characteristic (ROC) analysis, scalar performance measures at the optimal ROC threshold, image similarity measures such as normalized cross-correlation (NCC)/mutual information (MI), and abnormality detection rate (DR). We also present visual examples of the predicted abnormality maps, covering various scenarios of abnormality occurrence (e.g., those listed in Figure 1) and some special cases such as metal device detection.

2 | METHODS

Let A denote the domain of CXRs with pneumonia and let N denote the domain of CXRs that are normal. Given only an unpaired training data set of images from both domains $a_1, a_2, a_3 \dots a_m$ and $n_1, n_2, n_3 \dots n_n$ (no other ground truth information like abnormality location/BB annotations), our goal is to estimate a pixel-wise abnormality map that highlights the regions of pneumonia in a novel abnormal image. By unpaired, we mean that there is no overlap between the patient cohorts of images in A and N. We take a GAN-inspired image-to-image translation approach to tackle this problem. We briefly describe GANs and their variants in the next section to motivate our algorithm.

2.1 | Relevant background on GANs and their variants

GANs are neural network models that can learn and sample data from high-dimensional probability distributions such as images. The classical GAN for images consists of an image generator and a discriminator, which compete with each other in a zero-sum game.[31] The goal of the generator is to fool the discriminator by generating synthetic images that are indistinguishable from real images. The goal of the discriminator is to distinguish between the generated synthetic images and the real images. Typically, the generator and discriminator are neural network models, trained simultaneously. The output of the discriminator is used to train the generator to improve image generation quality.

The traditional GAN takes in a random vector as input and generates data from the learned probability distribution as output. A GAN can also be trained conditioned on additional information such as class labels or other images to generate data with specific properties. Such training imposes restrictions on the modes of the learned data distribution and are called conditional GANs (cGAN). Conditional GANs are well suited for image-to-image translation,[32,33] where the goal is to take images from one domain and transform them to have the style/characteristics of images from another domain. cGANs have been applied to various problems in medical imaging including low-dose computed tomography (CT) denoising,[34] superresolution in retinal fundus images,[35] magnetic resonance (MR) reconstruction,[36] CT image synthesis from MR,[37,38] and brain image segmentation.[39]

Although many impressive results have been reported using the basic GAN/cGAN setup described above, they suffer from training instability and mode collapse problems. Mode collapse refers to the phenomenon where the generator learns to generate samples from a few modes of the data distribution while missing many others, even though the samples from the missing modes occur in the data set. These problems exacerbate the difficulty of training these models and potentially prevent the learned distribution from converging to the real data distribution.[40] Arjovsky et al[40] proposed the Wasserstein GAN, a variation on the traditional GAN that addresses these drawbacks. In Wasserstein GANs, the discriminator is replaced with a critic function based on the Wasserstein distance metric. The Wasserstein metric is a meaningful and robust way to measure the distance between the distributions of the generated and real images. This metric is minimized iteratively during GAN training to improve image generation performance. Section 2.2 provides some intuitions on the working of the metric. Wasserstein GANs have been shown to have better optimization/convergence properties,[40] owing to its favorable continuity (everywhere[1]) and differentiability (almost everywhere[2]) properties under mild assumptions. A key advantage of the method is that it allows the critic to be trained till optimality, providing better gradients to the generator and hence helping to produce images with higher quality and diversity.[40] With this background on GANs, we next describe our methodology in detail.

2.2 | Our methodology

In this work, we take a conditional Wasserstein GAN-based image-to-image translation approach (abnormal CXR to abnormality map) to localize abnormalities on CXRs. Specifically, we train a cGAN that generates a pixel-wise abnormality map conditioned on the input abnormal CXR. We leverage the Wasserstein critic function to take advantage of its favorable properties for GAN training. Let $p_d(a)$ and $p_d(n)$ denote the distributions of images contained in A (abnormal CXRs) and N (normal CXRs) in our data set, respectively. We model the translation of images from domain A to domain N using the following additive relationship:[23,30]

$$\hat{n}_i = G(a_i) + a_i. \tag{1}$$

Here, a_i is an abnormal image, $\hat{n}_i$ is the translated *pseudo* normal image, and G is the GAN generator. The *pseudo* normal CXR $\hat{n}_i$ is the "hypothetical" normal if the abnormal CXR a_i were not to have any abnormalities. $G(a_i)$ is an image generated by G with a_i as input. We call $G(a_i)$ the "difference map" because it represents the difference between $\hat{n}_i$ and a_i. Using this formulation, the goal is for the difference map $G(a_i)$ to capture (or for G to learn) all the relevant information that makes an abnormal CXR different from a normal CXR. This information manifests as differences in intensities of pixel/regions belonging to abnormalities. Hence, the difference map $G(a_i)$ is our predicted abnormality map. To train the generator, we compare the distribution of the *pseudo* normal images (obtained by adding $G(a_i)$ to a_i) to the distribution of the real normal images using the Wasserstein distance metric. This measure is backpropagated through the generator network to optimize its weights and improve image generation performance. A block schematic of our method is shown in Figure 2. We next describe our generator, critic, and loss functions (L_{gan}, L_{sim}, L_{iden}) in detail.

2.2.1 | Generator

Generating the difference map is a dense prediction task unlike classification, that is, as opposed to a scalar output, the generator needs to produce a spatially structured pixel-wise map at the same resolution as the input image. Other examples of dense image prediction tasks in computer vision include semantic segmentation[41] and optical flow.[42] For example, in semantic segmentation, the goal is to label each pixel of an input image as belonging to a particular object class, producing a pixel-wise segmentation map. In recent years, fully convolutional neural network (FCN) models have gained popularity for dense image prediction tasks.[41] FCNs consist of convolution, pooling, nonlinearity, and upsampling operations; they do not have dense connection layers. This reduces the number of parameters in the network. Also, FCNs do not require the input images to be of a predetermined specific size because there are no fixed number of units in any layer (e.g., as required by dense layers).[43] Building on the advantages of FCNs

[1] Continuous everywhere refers to a function being continuous everywhere in its domain. In the case of the Wasserstein metric, continuous everywhere implies continuity everywhere over the space of all the generator network parameters.
[2] A function is said to be almost everywhere differentiable if the set of points in its domain where it is not differentiable is contained in a set that has measure zero.

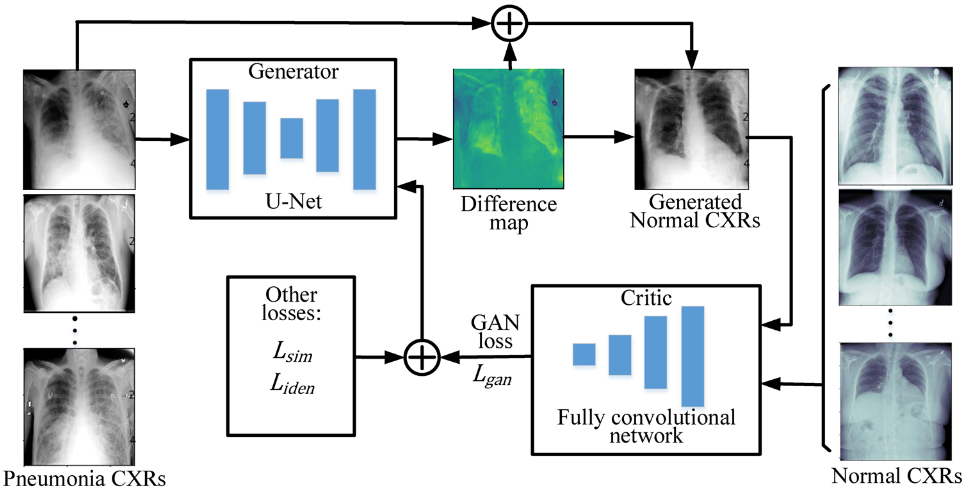


FIGURE 2 Block schematic of our approach for pneumonia localization on CXRs

for dense prediction, we model our generator based on a popular FCN architecture called the U-Net.[44] We next describe the U-Net and our model implementation.

U-Nets[44] are FCN architectures that have been successfully applied to semantic segmentation and other dense prediction tasks. The U-Net consists of the following parts: (1) a contracting path consisting of a series of convolution, nonlinearity (e.g., rectified linear units [ReLUs]), and pooling operations that reduce the dimensionality of the input image; (2) a symmetric expanding path where the pooling operations are now replaced by upconvolutions (transposed convolutions); and (3) skip connections between the contracting and expanding paths that bypass one or more layers of the network. The upconvolution operations increase the resolution of the output of the network layers in the expanding path. The skip connections enable the lower resolution feature maps from the expanding path to be combined with the higher resolution feature maps from the contracting path. The motivation behind this is that the contracting path interprets the image and its context (e.g., what is in the image), while the expanding path combined with the higher resolution features from the contracting path enables localization (where in the image). Merging features from various resolutions via skip connections aids in combining spatial information with contextual information.[43] Our U-Net implementation is shown in Figure A.1 in Online Appendix A and is based on Baumgartner et al.[23] The contracting path consists of 3×3 pixel convolutions with stride 1, 2×2 max pooling operations with stride 2, and ReLUs for nonlinearity. Starting with 16 convolutional filters in the first layer, the number of filters is doubled after each max pooling operation, reaching a maximum of 128 filters. The expanding path consists of up/transposed convolutions with stride 2, concatenation with contracting path features, 3×3 pixel convolutions, and a 1×1 convolution with no nonlinearity. The number of filters are halved after every upconvolution operation. All layers except for the last one use batch normalization. We next describe our critic function.

2.2.2 | Critic and Wasserstein distance

Let the critic function be denoted by C. The role of C is to compute a measure of dissimilarity between the distributions of the generated and real images. We achieve this by leveraging the Wasserstein metric, also called the earth mover's distance (EMD).[45] Informally, one distribution can be imagined to be a mass of earth spread over some space and the other distribution to be a group of holes in the same space. Then the Wasserstein distance is a special case of the transportation problem[46] and computes the minimum cost/amount of work required to transport earth to fill the holes. A unit of earth moved by a unit of ground distance is considered one unit of cost here.[47] This problem is intractable in the original form. Using Kantorovich–Rubinstein duality[48] and 1-Lipschitz continuity constraints,[49] Arjovsky et al[40] showed that the Wasserstein distance can be approximated using neural network functions. Kantorovich–Rubinstein duality converts the optimal transport problem, which is a special case of linear programming problems, to its dual, a maximization problem over 1-Lipschitz functions.[50] The solution of the dual is identical to the solution of the primal and allows to approximate the Wasserstein distance using parameterized functions.[40] Here, we take this approach and model the critic function C using an FCN architecture.[23] Our network is shown in Figure A.2 in Online Appendix A. It takes an image as input (either the *pseudo* normal or the real normal image) and produces a scalar output. The network is composed of 3×3 pixel convolutions with stride 1, 2×2 max pooling operations with stride 2, ReLUs for nonlinearity, a 1 × 1 convolution layer with no nonlinearity, and a final global average pooling layer. As with the generator network, the numbers of filters are doubled after each max pooling operation, reaching a maximum of 256 filters. Batch normalization was not applied to the layers as it prevented the critic from learning during the initial stages of training. This was consistent with similar observations reported by other studies in the literature.[23,51] Finally, the Wasserstein distance is computed by optimizing the neural network parameters of C according to the following equation:

$$W(p_d(n), p_d(\hat{n})) = L_{gan}(G, C) = \max_{C \in \mathcal{C}} [\mathbb{E}_{n \sim p_d(n)}[C(n)] - \mathbb{E}_{a \sim p_d(a)}[C(a + G(a))]], \quad (2)$$

where W is the Wasserstein distance, n is a normal image, a is an abnormal image, $a + G(a)$ is a *pseudo* normal image, $p_d(\hat{n})$ is the distribution of *pseudo* normal images, $\mathcal{C}$ is the set of 1-Lipschitz functions, and $\mathbb{E}[.]$ is the expectation operator. C is ensured to be 1-Lipschitz continuous when the norm of the gradients of C (with respect to its input) are at most 1 everywhere. A soft version of this constraint is incorporated as an additional term to the loss function, a penalty on the norm of the gradients of C.[51] Intuitively, the optimized critic C at each step would be relatively large for normal images and small for pneumonia images or *pseudo* normal images that have not converged to normal image distribution. The Wasserstein distance W would be relatively large for pneumonia or *pseudo* normal images compared to that of normal images because it measures how dissimilar the images are to normal images. From here on, we also denote the Wasserstein distance by GAN loss $L_{gan}(G, C)$, which is backpropagated iteratively through the generator network during training.

2.2.3 | Other losses

Training with the GAN loss alone may not be sufficient to generate meaningful abnormality maps/*pseudo* normal CXRs. For any patient, the abnormal CXR and the corresponding *pseudo* normal CXR would have many pixels in common (e.g., in healthy regions, i.e., those outside the abnormal regions), which need to be preserved during the image translation process. The GAN loss itself does not account for this patient-specific constraint. Hence, in addition to the GAN loss, we impose a minimality regularizer, L_{sim}, on the difference map by minimizing its norm[23]

$$L_{sim}(G) = \|G(a)\|_{L_1}, \quad (3)$$

where $\|.\|_{L_1}$ is the L1-norm.

In addition to the above losses, we introduce additional prior information/constraints into the model that can help improve the generation performance. We describe two such extensions here:

1. Image registration: As we noted in the introduction, normal anatomical variability in a patient population increases the complexity of abnormality localization. The GAN would require to learn this normal variability and be able to distinguish it from the variability caused due to abnormality. One way to mitigate the effect of normal anatomical variability is by registering all images to a common reference space. The reference space acts as an anatomical prior, constraining anatomical variability and helping the GAN focus on the features of interest in each abnormal image, that is, regions of abnormality. For the purposes of this study, we picked an exemplary normal CXR as our reference image with the help of a radiologist and registered all other images to it. Our image registration method was intensity based with an affine transformation function and MI similarity metric.[52]
2. Identity loss: In addition to producing realistic looking *pseudo* images from abnormal images, the GAN should retain a normal image "*as is*" when fed as input to the network.[33] This can help the GAN better understand the semantics of normal images/anatomy and as a result help perform better abnormal-to-normal image translation. We refer to this constraint as the identity loss, L_{iden}, and compute it as

$$L_{iden}(G) = \mathbb{E}_{n \sim p_d(n)}[\|n - G(n)\|_{L_1}], \quad (4)$$

where $\mathbb{E}[.]$ is the expectation operator, n is a real normal image, and $\|.\|_{L_1}$ is the L1-norm.

2.2.4 | GAN optimization and training

Putting everything together, our final optimization function for GAN training is given by

$$G^* = \arg\min_{G} \left[L_{gan}(G, C) + \lambda_1 L_{sim}(G) + \lambda_2 L_{iden}(G)\right], \quad (5)$$

where G^* is the optimal generator G and λ_1 and λ_2 are the weights assigned to each of the loss functions. We note here that the optimization also included the gradient penalty and image registration was performed as prestep, as discussed above.

Lastly, we describe our training procedure. The generator and the discriminator were trained simultaneously in an alternating fashion.[40] Multiple critic weight updates were performed for every generator weight update to ensure accurate computation of the Wasserstein distance. We leveraged the ADAM algorithm[53] for optimizing the weights of the network. We experimented with various combinations of the loss functions, setting the corresponding weight parameters λ_1 and λ_2 to 0 when the losses were not included during training. This concludes the description of our methodology. We next describe our data, experiments and validation procedure.

2.3 | Data, experiments, and validation

2.3.1 | Data

Our data set consisted of a total of 14 184 frontal-view CXRs, with 5659 labeled positive for pneumonia and the rest normal. The data are publicly available as part of the RSNA pneumonia detection challenge,[28] and are a subset of the bigger US National Institute of Health (NIH) CXR data set.[22,54] The BBs on the abnormal CXRs were annotated by six different board-certified radiologists from multiple institutions using a web-based commercial annotation tool. Any given CXR could contain multiple BBs if more than one area was suspected to have pneumonia. More details on the data and the annotation process are available at the RSNA challenge website[55] and described in detail by Shih et al.[54]

2.3.2 | Experiments

All the images were of size 1024×1024 pixels with intensities ranging from 0 to 255. We resampled the images to a resolution of 512×512 pixels for efficient training while retaining sufficient pixel-level anatomical information required for the problem.[29,56] We split the data into 65% training, 15% validation, and 20% test sets using random stratified sampling and used the same splits for all our experiments. The training set was used to optimize the weights of the network, whereas the validation set was used for hyper parameter tuning (e.g., determining the number of training epochs using early stopping[57]). The test set was held out during training and was used to measure algorithm generalization performance. All our results are presented on the test set.

For the ADAM optimizer, we used the following parameter settings: $\beta_1 = 0$, $\beta_2 = 0.9$, and learning rate = 0.001. The loss weights λ_1 and λ_2 were set to 100 and 10, respectively, when the corresponding loss functions were included during training. These values were found to be optimal empirically. We accumulated gradients from four batches, each of size 8, before performing a step of gradient descent. This effectively allowed for training with a batch size of 32. All training was performed on an Nvidia Tesla P100 GPU with 16 GB of RAM and took approximately 40 h per session.

The affine image registrations were performed using the Insight Toolkit (ITK) library.[58] The MI similarity metric was maximized using the gradient descent optimizer (learning rate = 1.0, maximal number of iterations = 200, minimum value for convergence = 1e-6). Linear interpolation was used for resampling. ITKs multiresolution registration framework was utilized to perform coarse to fine image registrations with three levels of image pyramid (factors by which the images were shrunk at each level = [4, 2, 1], smoothing Gaussian sigmas at each level = [2, 1, 0]). The multiresolution registration approach is widely popular and aids in improving accuracy, robustness, and speed.[58] For our methods that used image registration, the ground truth BBs were also transformed appropriately using the obtained affine transformation.

We present results for different combinations of our methods, such as the plain Wasserstein GAN, Wasserstein GAN with identity loss, Wasserstein GAN with image registration and identity loss, and so on. For comparison, we implemented the CAM methodology from Zhou et al,[20] which is based on a neural network image classifier. As noted in the introduction, CAM produces class-specific output maps that are surmised to capture the visual differences between two image classes. In our case, the two image classes are the normal and abnormal CXRs, with the resulting output map highlighting regions of abnormality. The CAM classifier architecture was similar to our critic function with a few changes in the last layer to produce a classification output. Further, we also compare to two supervised approaches: (1) a supervised version of our method by additionally incorporating BB information and (2) RetinaNet,[59] a state-of-the-art deep CNN-based object detection model. For (1), a BB loss L_{bb} was added to the GAN optimization (Equation (5)), computed as the negative dice coefficient[60] between a sigmoid transformed $G(a)$ and a BB mask image. The dice coefficient measures the ratio of twice the intersection between the predicted mask and the ground truth mask, to the sum of the masks. The BB mask image itself was obtained by setting the pixels inside the BBs to 1s and those outside to 0s. This loss was weighted by parameter λ_3, set to an optimal value of 200 found empirically. For (2), we trained a RetinaNet model using the detectron2 library,[61] which takes in ground truth object BBs during training and predicts BBs at locations likely to contain the object of interest in the test images. We would like to note that the RetinaNet model was also used by the top performers in the RSNA pneumonia detection challenge.

2.3.3 | Validation

We evaluate the methods by measuring the overlap/similarity between the generated abnormality maps and the ground truth BBs using multiple quantitative metrics:

(1) **ROC curve analysis**: We threshold the [0,1]-normalized abnormality maps with thresholds ranging from 0 to 1. Each such thresholding results in a binary abnormality map, with 1s inside the predicted abnormal region and 0s outside it. We then compute the pixel-level true positive rate (TPR) and false positive rate (FPR) of our prediction as follows: TPR = #pixels in the intersection between the predicted abnormal region and the ground truth BB masks over the total #pixels inside the ground truth BB masks; FPR = #pixels that are predicted as abnormal but are outside the ground truth BB masks over the total #pixels outside the ground truth BB masks. Lastly, we average the TPR and FPR values at each threshold across all cases in our test set to compute summary TPR and FPR values at that threshold. These values are used to plot the ROC curve.

(2) **Scalar performance metrics at optimal ROC threshold**: We present scalar performance metrics such as sensitivity, specificity, geometric mean, accuracy, and F_1 score at the maximal Youden's index[62] point. The Youden's index is a statistic that summarizes the performance of a diagnostic test, given by sensitivity + specificity − 1.

(3) **Similarity between the abnormality maps and the ground truth BB images**: We compute image similarity scores between the [0,1]-normalized abnormality maps and the ground truth BB images (the BB masks converted to images by setting the pixels inside the BB masks to 1s and those outside to 0s). We chose NCC and MI as our image similarity metrics because they compute a similarity score that is not sensitive to the absolute image intensity values in the two images.

(4) **Pneumonia DR**: We compute DR as the rate of successful detection of abnormality at the level of whole images. A detection is deemed successful if the prediction has an overlap of at least 50% with the union of ground truth BBs of the image.[63–66] The motivation for this metric is to evaluate how well our methods "detect" the regions of abnormality as opposed to how accurately they overlap with the ground truth pixels. We compute DR at different intensity thresholds and plot them against the corresponding pixel-level FPR defined above. The DR curves for all the methods and their AUCs are presented for pixel FPR $< 25\%$.

The RetinaNet model is also evaluated in a similar way by thresholding the [0,1] confidence scores associated with the BB predictions with thresholds ranging from 0 to 1.

Finally, we present visual examples of CXRs and the corresponding abnormality maps generated using all the methods. These examples cover various scenarios of abnormality occurrence such as single-site localized lung opacity, multisite localized lung opacity, diffuse opacity covering the entire lung(s), mild intensity opacity, nodular/patchy opacity, and so on. Additionally, we also show some interesting cases such as response to metal objects and normal images. We conclude with some examples where our method differs from the ground truth.

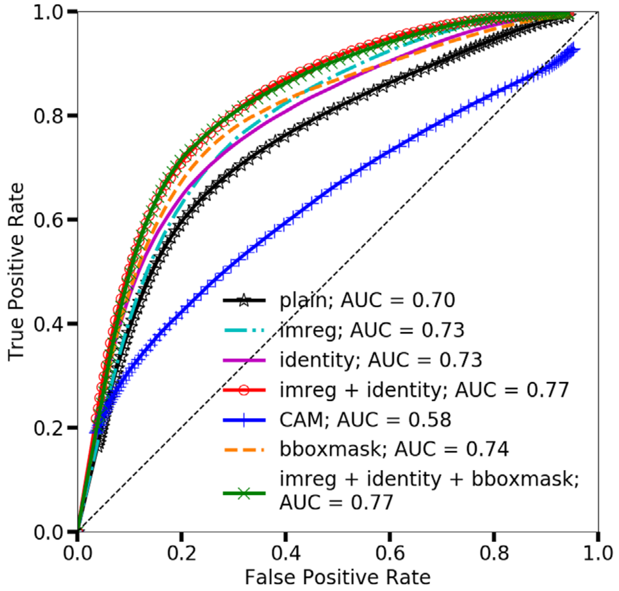


FIGURE 3 ROC curves of all methods. Note that all our methods perform better than CAM. The combination of Wasserstein GAN with minimality regularizer, image registration, and identity loss, `imreg + identity`, achieves the best ROC with an AUC of 0.77. `imreg + identity + bboxmask` achieves a similar performance

3 | RESULTS

In this section, we present quantitative and qualitative results for all the methods. In the description below, we refer to the Wasserstein GAN with only minimality regularizer L_{sim} as "`plain`." All the other methods are additions on top of this basic condition, that is, `plain` with image registration is referred to as "`imreg`," `plain` with identity loss as "`identity`," `plain` with BB loss as "`bboxmask`," `plain` with image registration, identity and BB loss as "`imreg + identity + bboxmask`," and finally, `plain` with image registration and identity loss as "`imreg + identity`." All methods except `bboxmask` and `imreg + identity + bboxmask`, are weakly supervised and are the main contributions of this paper. `bboxmask` and `imreg + identity + bboxmask` use BB supervision (via BB loss L_{bb} described in the previous section) and are included only for comparison.

Figure 3 shows the ROC curves for pixel overlap between the predicted abnormal region and the ground truth BBs. We see that all our methods achieve

TABLE 1 Scalar performance metrics at maximal Youden's index

Method	Sensitivity	Specificity	Geometric mean	Accuracy	F_1 score
plain	0.66	0.75	0.70	0.74	0.36
imreg	0.72	0.72	0.72	0.72	0.38
identity	0.69	0.75	0.72	0.75	0.38
CAM	0.40	**0.82**	0.57	0.76	0.32
bboxmask	0.73	0.74	0.74	0.74	0.40
imreg + identity + bboxmask	0.74	0.78	0.75	**0.77**	0.43
imreg + identity	**0.74**	0.77	**0.75**	0.76	**0.43**

TABLE 2 Image similarity scores between abnormality maps and ground truth bounding box images

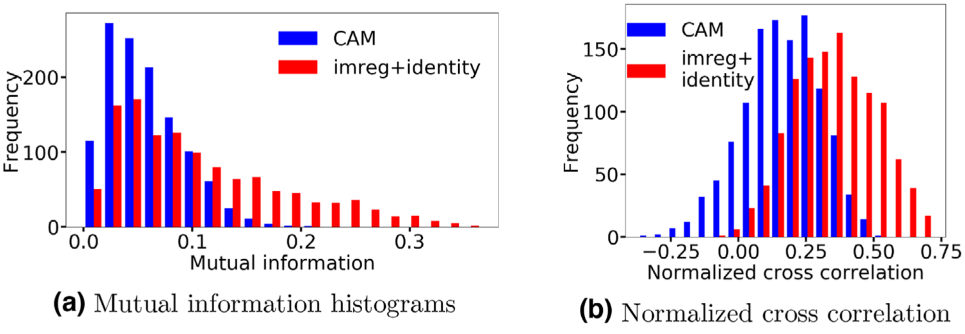
Method	Median NCC	Median MI
plain	0.26	0.05
imreg	0.30	0.07
identity	0.27	0.06
CAM	0.16	0.04
bboxmask	0.30	0.07
imreg + identity + bboxmask	0.35	0.09
imreg + identity	**0.35**	**0.09**

a superior performance compared to CAM, with an AUC greater than 0.70. Our top performing model is `imreg + identity` with the highest AUC of 0.77. `imreg` and `identity` perform better than `plain` with an AUC of 0.73. We also note that `bboxmask` performs better than `plain` (AUC 0.74), whereas `imreg + identity + bboxmask` has the same performance as that of `imreg + identity`.

Table 1 shows scalar performance metrics at the maximal Youden's index ROC point. We see a similar trend for sensitivity, geometric mean, and F_1 score, where including the additional losses over `plain` improves prediction performance. `imreg + identity` achieves the highest sensitivity, geometric mean, and F_1 score. CAM and `imreg + identity + bboxmask` achieve the highest specificity and accuracy, respectively, followed closely by `imreg + identity`. We also observe that all the methods suffer from a low F_1 score with the highest value of 0.43.

Table 2 lists the MI and NCC image similarity metrics, comparing the predicted abnormality maps with the ground truth BB mask images. Median similarity scores across all test images are presented with higher values, indicating higher similarity to ground truth. We again observe a similar trend where all our methods achieve a higher score compared to CAM. Including additional losses over `plain` aids performance with `imreg + identity` achieving the highest scores. This is also illustrated in Figure 4, where we see that the distribution of `imreg + identity` scores are shifted to the right toward higher values compared to the distribution of CAM scores.

Figure 5 shows the abnormality DR plotted against false positive rate (FPR) of pixel overlap with ground truth. We see that all our methods have a superior DR curve compared to CAM and achieve a DR > 80% at a maximum pixel-level FPR of 0.23. `imreg + identity` and `imreg + identity + bboxmask` achieve the best performance with a DR > 80% at a lower pixel FPR of 0.18 and a DR of 85% at a maximum pixel FPR of 0.25. We note that the DR curves of all our methods increase rapidly between an FPR of 0.05 and 0.15 and start to plateau post 0.2. We also performed free-response ROC (FROC) analysis to evaluate the localization performance at the level of individual abnormalities (as opposed to DR analysis above, which is an image-level measure of detection success). The FROC methods were based on connected components



(a) Mutual information histograms

(b) Normalized cross correlation

FIGURE 4 Histograms of mutual information and normalized cross-correlation scores for our method `imreg+identity` and CAM

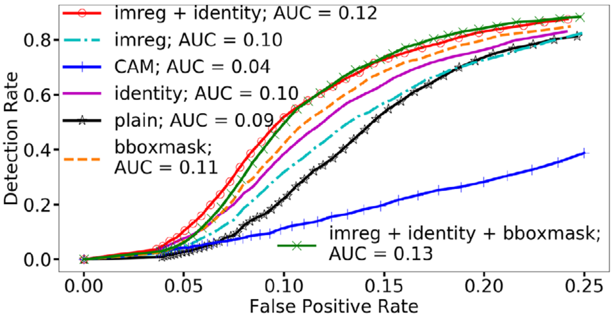
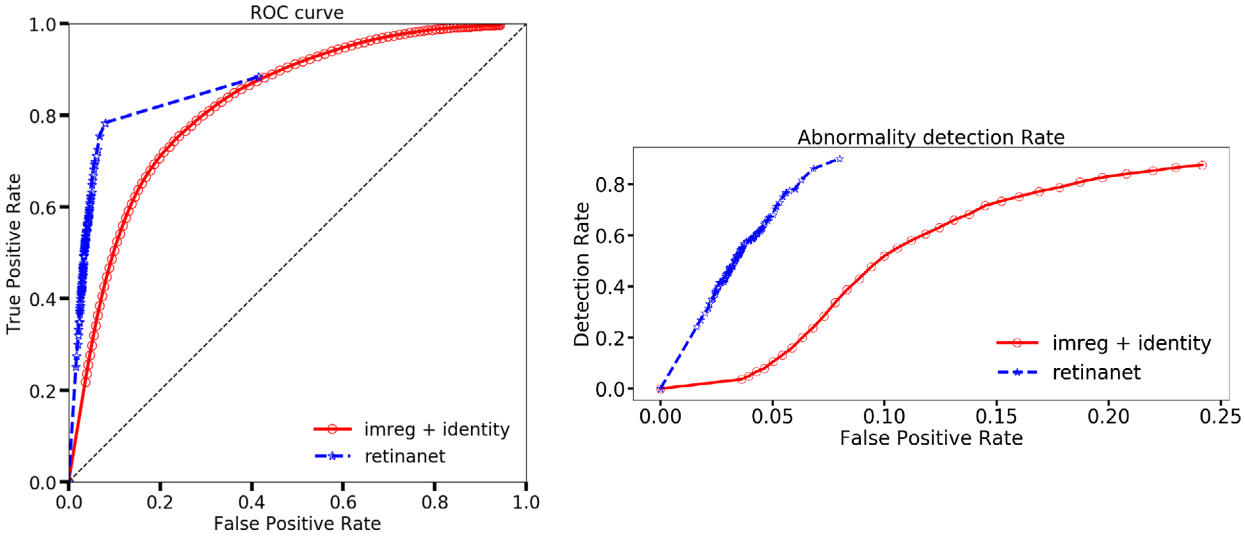


FIGURE 5 Abnormality detection rate curves for all methods. Note that `imreg + identity` and `imreg + identity + bboxmask` achieve the highest scores

analysis[67,68] and intensity peaks of the abnormality maps.[69] An abnormality-level detection TPR was computed for successful detection of individual abnormalities (as opposed to whole image as in DR analysis or pixel-level TPR as in the ROC analysis) across all images and measured against the average number of false positive detections per image, at each intensity threshold of our abnormality map. Our algorithm `imreg + identity` attained a maximum abnormality-level detection TPR of 0.75 with a corresponding average false positive detection value of 0.5 per image. The TPR was lower for CAM with maximum value <0.5, and corresponding average false positives per image being higher in the 2–3 range. More details on our FROC methodology and results can be found in Online Appendix B.

Figure 6 shows a comparison of the performance of our method and the BB-supervised RetinaNet model. The RetinaNet is closer to the ideal ROC point (top left) compared to our method, as seen in Figure 6(a). We note that the ROC of RetinaNet does not extend toward pixel-level TPR, FPR value of 1 (top right) and discuss this in Online Appendix C. From the DR curves in Figure 6(b), we observe that the RetinaNet model achieves a higher DR at a much lower pixel FPR compared to our algorithm. Further details of the RetinaNet comparison and visual examples of BB prediction can be found in Online Appendix C. This concludes our quantitative results.

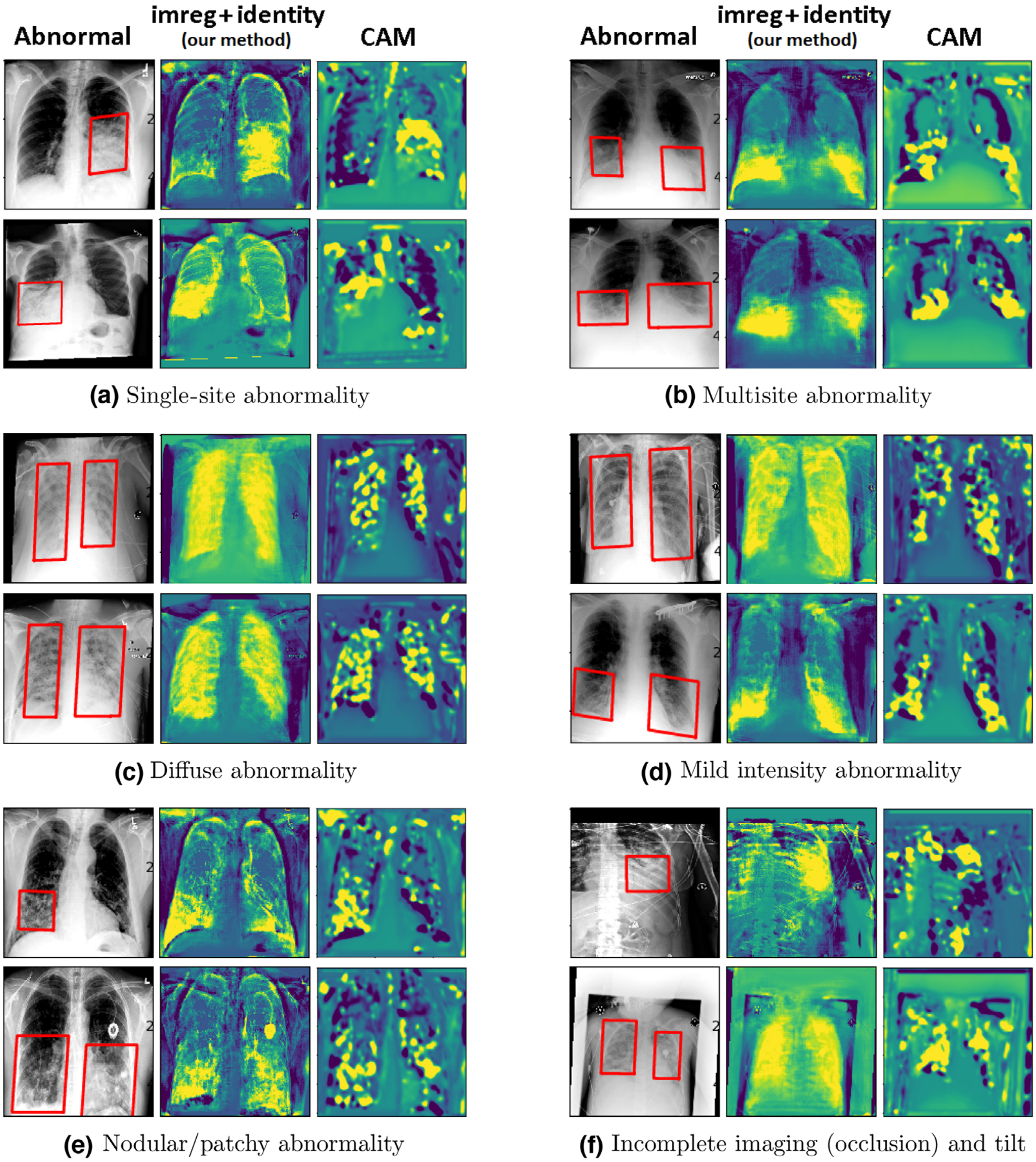
We next present exemplary results of our methods covering various scenarios of abnormality occurrence. The scenarios include abnormalities varying in size, number, shape, location, orientation, intensity, and textural patterns. In each case, the abnormal CXRs are shown along with the ground truth BB annotations marked around the abnormality in red. The corresponding abnormality maps are presented as heat maps in the viridis scale,[70] where the colors range from purple to green to yellow, representing increasing intensities. The higher temperature/intensity regions on the heat maps (yellow) indicate presence of abnormality and the lower temperature/intensity regions (green/blue/purple) indicate the lack thereof. For ease of interpretation, the regions on the CXR suspicious for pneumonia can be roughly described as those that appear hazy and lack sharp boundaries. The abnormality localization would be considered a success if the corresponding regions on the abnormality map appear bright with high contrast compared to the background. By medical convention, right and left sides are always in reference to the patient. In the chest radiographs shown, the patient is facing forward. Therefore, the lung on the left side of the image is the patient's right lung, and vice versa. Throughout this manuscript, right and left lungs will always be in refer-



(a) ROC curves of our method and RetinaNet.

(b) Detection rates of our method and RetinaNet.

FIGURE 6 Comparison of performance of our method and state-of-the-art bounding-box-supervised method, RetinaNet



(a) Single-site abnormality

(b) Multisite abnormality

(c) Diffuse abnormality

(d) Mild intensity abnormality

(e) Nodular/patchy abnormality

(f) Incomplete imaging (occlusion) and tilt

FIGURE 7 Examples of abnormality detection on CXRs covering different scenarios of abnormality occurrence. Two examples are presented for each scenario, with the original CXRs shown on the left (abnormalities marked by bounding boxes shown in red), our method `imreg + identity` shown at the center and CAM shown on the right. Note the fulsome response produced by our method compared to CAM, which is incomplete with many holes

ence to the patient. Finally, it is also important to note that normal regions such as the heart (oval structure in the lower chest, left of midline, overlapping with the lower edge of the left lung), ribs and the diaphragm (along the lower lung boundary) also appear bright on the CXR and sometimes have a very similar appearance to pneumonia (see Figure 1). A key goal of the algorithms is to be able to discern such subtle confounding effects.

Figure 7 shows examples of abnormality maps generated using our best performing method `imreg + identity` and CAM. Figures 7(a) and 7(b) show cases where pneumonia appears localized to one/two regions in the lungs. We see that our method produces a high-intensity response corresponding to those regions, successfully identifying all the pixels suspected for pneumonia. In other CXRs, pneumonia appears diffused throughout the lungs. Figure 7(c) shows two such examples with varying degrees of abnormality (the top example has a brighter pattern compared to the bottom). The corresponding abnormality maps are bright across both the lungs, highlighting regions of abnormality successfully. The next three scenarios present some challenging cases for localization. Although pneumonia appears mild (low-intensity values on the CXR) in Figure 7(d), it has a nodular/patchy appearance in Figure 7(e). Figure 7(f) shows cases where the CXR itself is partially occluded due to incomplete imaging (top) and tilted (bottom). We see that our method successfully highlights the regions of abnormality in every case. Interestingly, the method also picks up the metal device in Figure 7(e) (bottom), which we will explore more in the discussion section. We would also like to note that our method is able to distinguish normal from abnormal regions, even in cases where they have very similar appearances with an almost invisible boundary. Figures 7(a) and 7(c) show two such examples, where our abnormality maps do not extend into the diaphragm in the lower lungs, in spite of overlapping abnormal regions with similar intensities. Similarly, in Figure 7(b), the abnormality response does not extend into the heart, beyond the diagonal edge in the lower left lung. The abnormality maps do not respond to the ribs in any case. Finally, we observe that CAM also produces higher intensity response corresponding to the regions of abnormality. But the key differences from our method are that the responses are (1) partial, not covering the abnormality fully and sometimes missing it (e.g., Figure 7d); (2) have a fractured appearance, with many holes and missing regions. Additional results for each of the above scenarios along with the generated *pseudo* normal CXRs can be found in Figures D.1, D.2, and D.3 in Online Appendix D.

Figure 8 presents a detailed comparison of all our methods. Figure 8(a) shows single-site pneumonia cases where we see that all our methods respond to the regions of abnormality successfully, with `imreg + identity` producing the most accurate response. Figure 8(b) shows challenging multisite pneumonia cases, where the top example has abnormalities of different sizes (1) diffused across the entire right lung and (2) a relatively small localized abnormality in the left lung) and the bottom example has abnormalities that have a nodular/patchy appearance. We see that the methods with the additional losses perform better than `plain` in all the cases and also pick up the infusion port in the bottom example. Figure 8(c) shows cases where pneumonia appears diffused across both the lungs. The top example is another challenging case where both the lungs appear bright and homogeneous. As we can see, there is no visible boundary along the diaphragm at the bottom of the lungs and a barely visible boundary along the sides of the lungs. Interestingly, all the methods are still able to infer anatomical locations and highlight the regions of abnormality. The bottom example is another interesting case where the ground truth BB masks do not cover all the regions of lung opacity (essentially present across the entirety of both the lungs). Here, all the methods successfully highlight abnormal regions beyond the ground truth BBs, with the best methods producing a response covering entire lungs. We also see that `bboxmask`'s response focuses around the BB mask in the right lung, missing regions of abnormality outside of it. This is likely due to the increased effect of the bounding loss in the absence of other constraints such as image registration and identity loss. Finally, Figure 8(d) shows some cases where pneumonia has a nodular and mild appearance. Again, we see the methods with the additional losses performing better than `plain`. The bottom example has mild lung opacity beyond the regions marked by the ground truth BBs, which the top performing models respond well to. From these results, we observe that `imreg`, `imreg + identity + bboxmask`, and `imreg + identity` generally produce better abnormality maps compared to the others, with the later two producing the most accurate ones. This concludes our qualitative results.

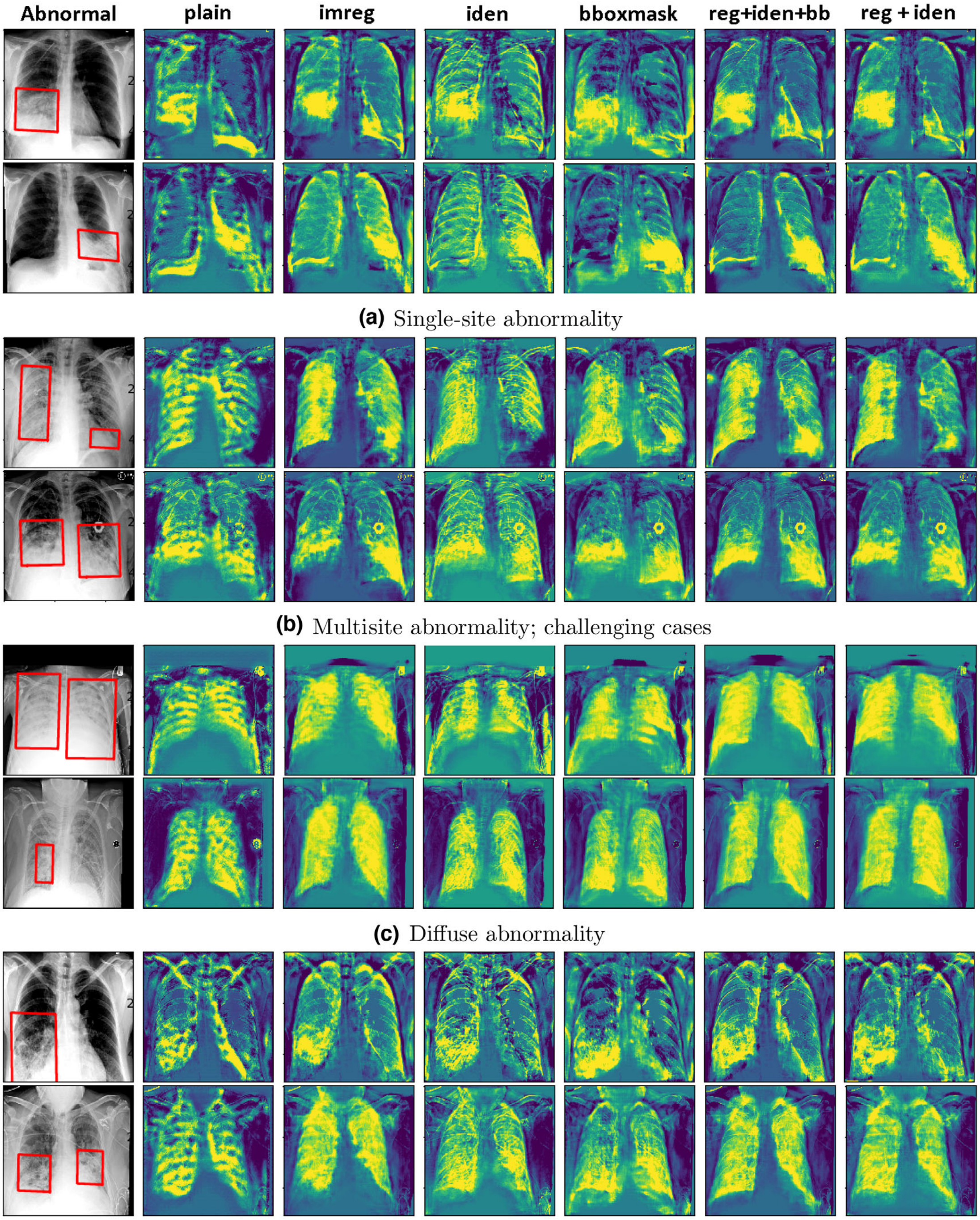
4 | DISCUSSION

4.1 | Summary of results

To summarize the results: (1) our GAN-based weakly supervised approaches are able to identify pixels suspicious for pneumonia, with the highest ROC AUC of 0.77, sensitivity of 0.74, specificity of 0.77, accuracy of 0.76, and a DR of 85%, when measured against radiologist annotated BBs; (2) all our approaches outperform CAM, a popular state-of-the-art weakly supervised approach for localizing abnormalities in images; (3) the performance of the GAN improves as we add additional prior information/knowledge constraints. For example, `imreg` and `identity` perform better than `plain`. In particular, registering images to a common reference space is an effective prior; (4) the combination of all the weakly supervised losses, `imreg + identity`, achieves the best performance compared to including each of them individually, potentially capturing complementary information.

The better performance of our methods over CAM can likely be attributed to the ability of GANs to impose higher order data consistency via the critic function. This enables modeling the entire image and learning coarse-to-fine image details accurately, without having to explicitly specify the features of interest. The Wasserstein critic provides a meaningful metric for GAN training that correlates well with the quality of the generated samples. Incorporating prior knowledge into the model additionally helps the GAN focus on the regions of interest relevant to the problem. In contrast, the CAM output is incomplete and noisy. This is likely due to the classifier focusing only on the most important local features of the image that are useful for the classification task, while ignoring the others.

We want to emphasize the weakly supervised nature of our methods, not requiring any location annotations such as BBs/segmentations for training.



(a) Single-site abnormality

(b) Multisite abnormality; challenging cases

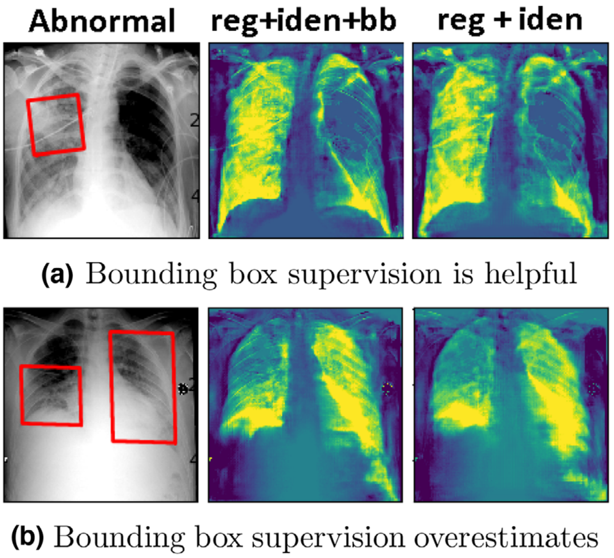
(c) Diffuse abnormality

(d) Nodular abnormality; challenging cases

FIGURE 8 Comparison of all our methods for different abnormality scenarios. Note that `imreg`, `imreg + identity + bboxmask`, and `imreg + identity` generally produce superior responses compared to others, with the later two (last two columns) producing the best

4.2 | Comparison to BB supervision

It is interesting to note that the BB supervised approaches, `bboxmask` and `imreg + identity + bboxmask`, seemingly do not perform better than the best weakly supervised model, `imreg + identity`. `bboxmask` does better than `plain`, but worse than `imreg + identity`. `imreg + identity + bboxmask` and `imreg + identity` have equivalent performances, with the former doing slightly better in some cases (e.g., it produces a marginally better (fuller) response to the abnormal regions in Figures 8a [bottom example], 8(b), and 8(d) [top example]). This suggests that the incorporated knowledge constraints, that is, registering images



(a) Bounding box supervision is helpful

(b) Bounding box supervision overestimates

FIGURE 9 Effect of bounding box supervision. Notice that `imreg + identity + bboxmask` produces a slightly better abnormality map than `imreg + identity` in (a), whereas the opposite is true in (b)

to a common reference space, identity mapping for normal cases, and other losses, may already be constraining the GAN enough to respond to only regions of abnormality, which otherwise require training with BB supervision. Although this observation requires further validation, there are some possible explanations: (1) although the BBs help the GAN focus on the regions of abnormality (e.g., in Figure 9a, `imreg + identity + bboxmask` responds better to the abnormality compared to `imreg + identity`), the BBs do not always cover the abnormality exactly; some of the BBs overestimate abnormal regions, causing the algorithms to overestimate as well. For example, in Figure 9(b), the BB extends across the entire left lung and beyond, with `imreg + identity + bboxmask` responding to even normal regions in the top part of the left lung. `imreg + identity` correctly identifies only the amorphous region at the bottom of the left lung. Another example is the bottom case of Figure 8(c), where we see that the entire right lung has an amorphous appearance, whereas the BB annotation covers only a small region. `bboxmask` responds only to the BB region, missing the upper lung, while `imreg + identity` responds to the entire lung; (2) we implemented one way of incorporating BB supervision and have optimized its relative importance/weight. There may be room for improvement here, enabling better utilization of BB supervision.

We also note that comparison to the RetinaNet model in Figure 6 alludes to the current gap that exists between our weakly supervised algorithm and a state-of-the-art BB-supervised algorithm. We hope that this gap can be improved in the future.

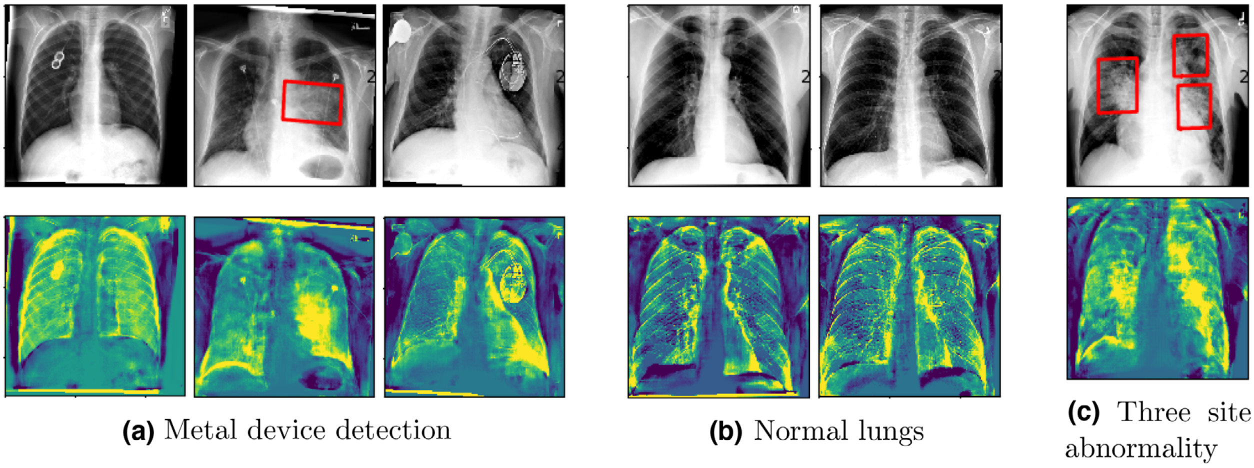
4.3 | Metal device detection

Another interesting feature of our methods is their ability to detect metal devices in CXRs. Figure 10(a) shows three examples, where the CXR on the left has an infusion port, the middle has ECG leads (tiny wires in either lung; these are external to the patient, placed on the skin), and the right has a pacemaker. We see that all of them are highlighted well in the abnormality maps, including the wires of the pace maker and the ECG leads that are only a few pixels wide. We attribute this to the relatively stronger intensity of metal devices compared to the background, which are picked up by the GAN. Figure 10(b) shows two examples of normal CXRs with no abnormality (see Figure D.4 in Online Appendix D for more examples). We see that the method produces a minimal response with no particular region highlighted. Finally, Figure 10(c) shows a case where the method successfully highlights all three distinct localized regions of abnormality.

We performed additional analysis to assess the effect of image resolution, dependence on data set size, loss functions to encourage spatial localization in the predicted abnormality maps, and different methods for incorporating BB supervision. Please refer to Online Appendix E for details.

4.4 | Failure cases and limitations

Although our methods achieve compelling results in many cases, they are far from uniformly successful. Figure 11 presents examples of common scenarios where the predicted abnormality maps differ from the ground truth BBs. Figure 11(a) shows false positives, where our method highlights regions outside the BBs as well. Some of these regions have mild opacity (e.g., upper lungs in the top example, by the heart in the middle example, etc.), whereas others are just errors made by the algorithm. A relatively common error is to highlight the lower lung and diaphragm boundary, even when there is no abnormality (e.g., lower right lung in the bottom example of Figure 11a). We believe the reason for this to lie in the varying shapes of normal lungs due to different breathing phases and variability across patients. Figure 11(b) shows common false negative scenarios. The top example is a challenging case, where the entire left lung is bright with no discernible lung boundary. Although the method detects the upper lung, it completely misses the lower part due to lack of any structure/recognizable pattern. The middle example is another case where the method appears to identify the abnormal pixels correctly whereas the BBs overestimate the abnormality, extending into normal lung regions. In the bottom example, the method appears to identify the amorphous regions not included in the

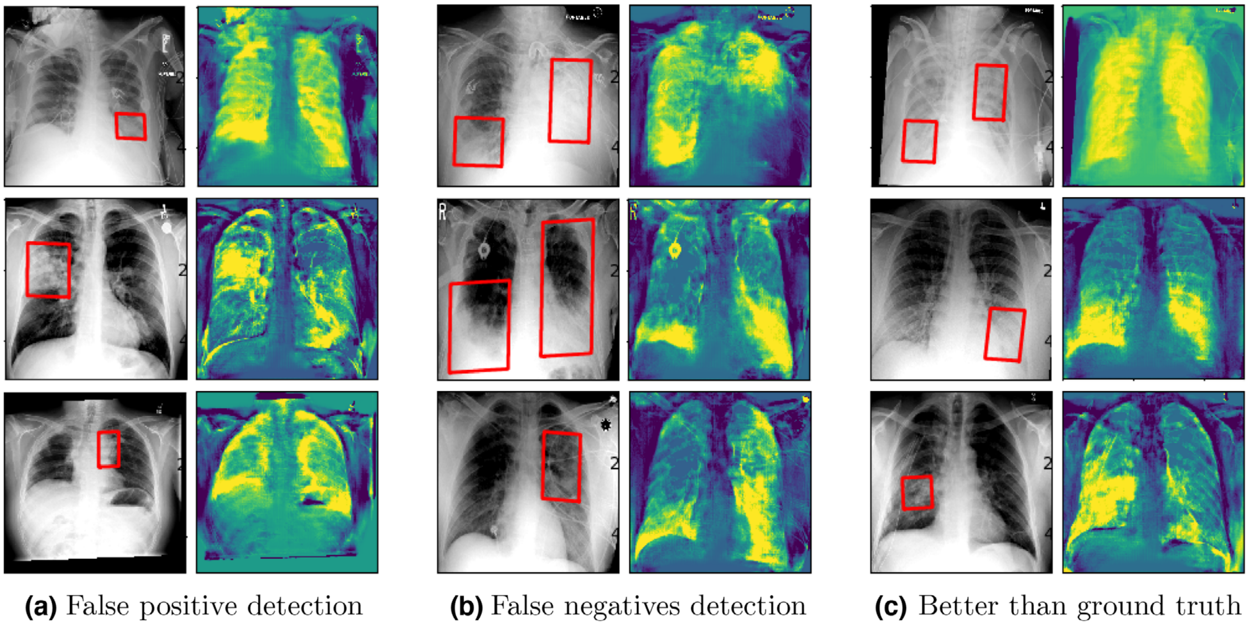


(a) Metal device detection

(b) Normal lungs

(c) Three site abnormality

FIGURE 10 Examples of metal device detection, response to normal lungs, and a three site abnormality. The CXRs are shown in the top row and the corresponding abnormality shown in the bottom. Notice that in (a), the method successfully highlights the infusion port (left), ECG leads (center), and pacemaker (right), including the tiny wires. There is no region highlighted in (b) and all three regions highlighted in (c), as expected



(a) False positive detection

(b) False negatives detection

(c) Better than ground truth

FIGURE 11 Examples of cases where the results differ from ground truth

BBs correctly, but misses the BB region itself. These examples illustrate the challenges in automatically interpreting CXRs due to complex anatomical variability, particularly in important rare cases. Finally, Figure 11(c) presents some examples where the method does seemingly better than ground truth BBs by successfully identifying regions of opacity that were not included by the BBs.

Although the quantitative results are very encouraging, there is room for improvement, particularly in the F_1 score. These again allude to the challenges in interpreting CXRs mentioned in the introduction. But we also would like to note that: (1) the ground truth BBs in our data set were not always accurate and some of the false positives/negatives were actually found to be valid detections as seen above; we believe one of the reasons for this to be the web-based annotation tool used by the radiologists, which likely allows marking regions only using rectangular boxes as opposed to irregular shapes that can cover the abnormality exactly; (2) our validation compares pixel-level predictions with BB ground truth, which is not ideal. This points to the lack of detailed pixel-level annotations in the field; (3) many studies have reported significant variability among radiologists in interpreting CXRs.[71,72] This may be causing variability in the ground truth (BB annotations and normal/abnormal image class), possibly confounding algorithmic detection/evaluation; and lastly, (4) interpreting pneumonia from CXRs alone is a hard problem; when available, radiologists compare CXRs of the patients from different time points, along with clinical symptoms, vital signs, history, and laboratory exam to make a successful diagnosis.

4.5 | Clinical impact

Our methodology for abnormality localization could have broader reaching clinical implications for medical imaging diagnostics. Radiologists spend years training their visual memory to understand anatomy and physiology, and distinguish normal from abnormal. The methodology of this study, using GANs, attempts a similar approach and is able to bring the abnormal regions to attention. Specifically, chest radiograph interpretation is one of the harder tasks in radiology. Such an algorithm can have several benefits. First, there is potential for it to serve as a screening tool, aiding interpretation of scans. In a busy practice, radiologists may be asked to interpret hundreds of CXRs every day. This tool may help radiologists to focus more attention on the abnormal CXRs, potentially helping to improve the timeliness and accuracy of CXR reports. Second, it can serve to augment the training of radiology residents who are training their visual memory of chest imaging. We would like to note here that the algorithm presented in this paper was trained to detect variations from a "normal" class of images. Indeed, such variations may include not only pneumonia, but any other finding outside of what is seen in a CXR performed on a healthy, normal individual, for example, metal implantable devices, and even postsurgical findings. At this time, a reasonable first step toward clinical implementation could be for the radiological community to use this tool as a guide to supplement their own independent reads of a CXR. The detected abnormalities could undergo review by a radiologist, with the final decision remaining in their control. Lastly, using appropriate data, the model can be trained to detect other abnormalities in CXRs. More generally, the technique can be extended to create algorithms capable of detecting and characterizing other diseases in different anatomical regions.

4.6 | Directions for future work

Based on the above results and discussion, we list some possible avenues for further exploration: (1) combining deformable image registration techniques with the GAN model for accurately learning the anatomical variability and localizing abnormality; (2) incorporating additional prior information such as organ models (e.g., lung) and constraints based on the textural/spatial characteristics of the disease; (3) creating pixel-level ground truth for accurate evaluation of localization techniques; and (4) extending our approach to other lung abnormalities such as pneumothorax, emphysema, and COVID-19.

5 | CONCLUSION

In this paper, we presented GAN-based weakly supervised approaches for localizing pneumonia on standard CXRs. Our approaches do not require expensive location annotations such as BBs/segmentations for training and produce pixel-wise abnormality maps at the same resolution as the input abnormal CXR. We evaluated our approaches on a large set of CXRs from the open-source RSNA pneumonia detection challenge,[28] consisting of abnormalities differing in size, shape, location, orientation, number, intensity, and textural patterns. Quantitative and qualitative results show the ability of our approaches to localize abnormality, even in challenging scenarios of abnormality occurrence. Additionally incorporating prior knowledge/constraints into the model was observed to help improve localization performance. We also showed that our approaches produce abnormality maps that are superior to CAMs. Finally, we discussed discrepancy from ground truth BBs, detection in special cases with metal devices and avenues for improvement.

ACKNOWLEDGMENTS

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748

CONFLICT OF INTEREST

None.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available as part of the RSNA pneumonia detection challenge at [https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview], reference number.[28]

ORCID

Krishna Nand Keshavamurthy
https://orcid.org/0000-0003-2408-7414
Carsten Eickhoff
https://orcid.org/0000-0001-9895-4061
Krishna Juluru
https://orcid.org/0000-0001-8203-8894

REFERENCES

1. Centers for disease control and preventionhttps://www.cdc.gov/pneumonia/prevention.html. Accessed on 9th Sept. 2021.

2. WHO. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. Accessed on 9th Sept. 2021. 2001.
3. Delrue L, Gosselin R, Ilsen B, Van Landeghem A, Mey dJ, Duyck P. Difficulties in the interpretation of chest radiography. In: Coche EE, Ghaye B, Mey J, Duyck P, eds. *Comparative Interpretation of CT and Standard Radiography of the Chest*. Springer; 2011:27-49.
4. Ropp A, Waite S, Reede D, Patel J. Did i miss that: subtle and commonly missed findings on chest radiographs. *Curr Prob Diagn Radiol*. 2015;44:277-289.
5. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*. 2015;35:1668-1676.
6. Berlin L. Radiologic errors, past, present and future. *Diagnosis*. 2014;1:79-84.
7. Rao B, Zohrabian V, Cedeno P, Saha A, Pahade J, Davis MA. Utility of artificial intelligence tool as a prospective radiology peer reviewer–detection of unreported intracranial hemorrhage. *Acad Radiol*. 2021;28:85-93.
8. Gao Y, Geras KJ, Lewin AA, Moy L. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. *Am J Roentgenol*. 2019;212:300-307.
9. Bley TA, Baumann T, Saueressig U, et al. Comparison of radiologist and cad performance in the detection of ct-confirmed subtle pulmonary nodules on digital chest radiographs. *Invest Radiol*. 2008;43:343-348.
10. Franquet T. Imaging of community-acquired pneumonia. *J Thoracic Imaging*. 2018;33:282-294.
11. Barry K. The chest radiograph. *Ulster Med J*. 2012;81:143-148.
12. Novack V, Avnon LS, Smolyakov A, Barnea R, Jotkowitz A, Schlaeffer F. Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia. *Eur J Internal Med*. 2006;17:43-47.
13. Neuman MI, Lee EY, Bixby S, et al. Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. *J Hosp Med*. 2012;7:294-298.
14. Albaum MN, Hill LC, Murphy M, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. *Chest*. 1996;110:343-350.
15. Young M, Marrie TJ. Interobserver variability in the interpretation of chest roentgenograms of patients with possible pneumonia. *Arch Internal Med*. 1994;154:2729-2732.
16. Xia Y, Ying Y, Wang S, Li W, Shen H. Effectiveness of lung ultrasonography for diagnosis of pneumonia in adults: a systematic review and meta-analysis. *J Thoracic Dis*. 2016;8:2822-2831.
17. Claessens YE, Debray MP, Tubach F, et al. Early chest computed tomography scan to assist diagnosis and guide treatment decision for suspected community-acquired pneumonia. *Am J Respir Crit Care Med*. 2015;192:974-982. PMID: 26168322.
18. Maughan BC, Asselin N, Carey JL, Sucov A, Valente JH. False-negative chest radiographs in emergency department diagnosis of pneumonia. *Rhode Isl Med J*. 2014;97:20-23.
19. Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW. Efficient pneumonia detection in chest x-ray images using deep transfer learning. *Diagnostics*. 2020;10:417.
20. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016:2921-2929.
21. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.
22. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chest x-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017:3462-3471.
23. Baumgartner CF, Koch LM, Can Tezcan K, Xi Ang J, Konukoglu E. Visual feature attribution using wasserstein gans. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
24. Li Z, Wang C, Han M, et al. *Thoracic Disease Identification and Localization with Limited Supervision*. Springer International Publishing; 2019:139-161.
25. Taghanaki SA, Havaei M, Berthier T, et al. Infomask: masked variational latent representation to localize chest disease. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing; 2019:739-747.
26. Sirazitdinov I, Kholiavchenko M, Mustafaev T, Yixuan Y, Kuleev R, Ibragimov B. Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database. *Comput Electr Eng*. 2019;78:388-399.
27. Babenko B. *Multiple Instance Learning: Algorithms and Applications*. Springer; 2008.
28. Rsna pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview. Accessed on 9th Sept. 2021.
29. Pan I, Cadrin-Chênevert A, Cheng PM. Tackling the radiological society of North America pneumonia detection challenge. *Am J Roentgenol*. 2019;213:568-574.
30. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep cnn for image denoising. *IEEE Trans Image Process*. 2017;26:3142-3155.
31. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc.; 2014:2672-2680.
32. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
33. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
34. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose ct. *IEEE Trans Med Imaging*. 2017;36:2536-2545.
35. Mahapatra D, Bozorgtabar B, Hewavitharanage S, Garnavi R. Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2017:382-390.
36. Kim KH, Do WJ, Park SH. Improving resolution of mr images with an adversarial network incorporating images with different contrast. *Med Phys*. 2018;45:3120-3131.
37. Nie D, Trullo R, Lian J, et al. Medical image synthesis with context-aware generative adversarial networks. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, eds. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing; 2017:417-425.
38. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Igum I. Deep MR to CT synthesis using unpaired data. In: *SASHIMI@MICCAI*, 2017.
39. Xue Y, Xu T, Zhang H, Long LR, Huang X. Segan: adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics*. 2018;16:383-392.
40. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Precup D, Teh YW, eds. *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. International Convention Centre, Sydney, Australia; 2017: 214-223.

41. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:640-651.
42. Dosovitskiy A, Fischer P, Ilg E, et al. Flownet: learning optical flow with convolutional networks. In: *2015 IEEE International Conference on Computer Vision (ICCV)*; 2015:2758-2766.
43. Theano deep learning documentation. http://deeplearning.net/tutorial/fcn_2D_segm.html. Accessed on 1st Oct. 2020.
44. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing; 2015:234-241.
45. Rubner Y, Tomasi C, Guibas LJ. A metric for distributions with applications to image databases. In: *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*. IEEE Computer Society; 1998:59.
46. Hitchcock FL. The distribution of a product from several sources to numerous localities. *J Math Phys*. 1941;20:224-230.
47. Earth mover's distance. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/RUBNER/emd.htm. Accessed on 9th Sept. 2021.
48. Rachev ST, Ruschendorf L. *Mass transportation problems. Volume I: Theory, Probability and its Applications*, Vol. 1. Springer-Verlag; 1998.
49. Sohrab HH. *Basic Real Analysis*, Vol. 231. 2nd ed. Springer; 2003.
50. Villani C. *Optimal Transport: Old and New*, Vol. 338. Springer Science & Business Media; 2008.
51. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:5767-5777.
52. Zitová B, Flusser J. Image registration methods: a survey. *Image Vision Comput*. 2003;21:977-1000.
53. Kingma DP, Ba J. Adam: a method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
54. Shih G, Wu CC, Halabi SS, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiol Artif Intell*. 2019;1:e180041.
55. Pneumonia dataset annotation methods. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/discussion/64723. Accessed on 9th Sept, 2021.
56. Gabruseva T, Poplavskiy D, Kalinin A. Deep learning for automatic pneumonia detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.
57. Girosi F, Jones M, Poggio T. Regularization theory and neural networks architectures. *Neural Comput*. 1995;7:219-269.
58. Johnson HJ, McCormick MM, Ibanez L, et al. *The ITK Software Guide: Design and Functionality*. Kitware; 2015.
59. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:318-327.
60. Isensee F, Petersen J, Klein A, et al. Abstract: nnu-net: self-adapting framework for u-net-based medical image segmentation. In Handels H, Deserno TM, Maier A, Maier-Hein KH, Palm C, Tolxdorff T, eds. *Bildverarbeitung für die Medizin 2019*. Springer Fachmedien Wiesbaden; 2019:22.
61. Wu Y, Kirillov A, Massa F, Lo WY & Girshick R Detectron2. https://github.com/facebookresearch/detectron2, 2019. Accessed on 9th Sept. 2021
62. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32-35.
63. Kegelmeyer Jr WP, Pruneda JM, Bourland PD, MLNA. Computer-aided mammographic screening for spiculated lesions. *Radiology*. 1994;191:331-7.
64. Oliver A. *Automatic Mass Segmentation in Mammographic Images*. PhD thesis, Universitat De Girona; 2008.
65. Kallergi M, Carney GM, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. *Med Phys*. 1999;26:267-275.
66. Penedo M, Souto M, Tahoces PG, et al. Free-response receiver operating characteristic evaluation of lossy jpeg2000 and object-based set partitioning in hierarchical trees compression of digitized mammograms. *Radiology*. 2005;237:450-457. PMID: 16244253.
67. Moor dT, Rodriguez-Ruiz A, Mérida AG, Mann R, Teuwen J. Automated lesion detection and segmentation in digital mammography using a u-net deep learning network. In: Krupinski EA, ed. *14th International Workshop on Breast Imaging (IWBI 2018)*, Vol. 10718. International Society for Optics and Photonics, SPIE; 2018:23-29.
68. Jin L, Yang J, Kuang K, et al. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: development and validation of fracnet. *EBioMedicine*. 2020;62:103106.
69. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
70. The viridis color palettes. https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html.
71. Novack V, Avnon LS, Smolyakov A, Barnea R, Jotkowitz A, Schlaeffer F. Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia. *Eur J Internal Med*. 2006;17:43-47.
72. Albaum MN, Hill LC, Murphy M, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. *Chest*. 1996;110:343-350.
73. Chakraborty DP. *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with r-Based Examples*, CRC Press; 2017.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Keshavamurthy KN, Eickhoff C, Juluru K. Weakly supervised pneumonia localization in chest X-rays using generative adversarial networks. *Med. Phys.* 2021;48:7154–7171. https://doi.org/10.1002/mp.15185