IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 7, 2021, accepted April 20, 2021, date of publication April 26, 2021, date of current version May 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075555

Learning to Localize Objects Using Limited Annotation, With Applications to Thoracic Diseases

EYAL ROZENBERG[1], **DANIEL FREEDMAN**[2], **(Member, IEEE), AND ALEX A. BRONSTEIN**[1], **(Fellow, IEEE)**
[1]Department of Computer Science, Technion-Israel Institute of Technology, Haifa 32000, Israel
[2]Google Research, Haifa 3190500, Israel

Corresponding author: Eyal Rozenberg (eyalr@cs.technion.ac.il)

ABSTRACT **Motivation:** The localization of objects in images is a longstanding objective within the field of image processing. Most current techniques are based on machine learning approaches, which typically require careful annotation of training samples in the form of expensive bounding box labels. The need for such large-scale annotation has only been exacerbated by the widespread adoption of deep learning techniques within the image processing community: deep learning is notoriously data-hungry. **Method:** In this work, we attack this problem directly by providing a new method for learning to localize objects with limited annotation: most training images can simply be annotated with their whole image labels (and no bounding box), with only a small fraction marked with bounding boxes. The training is driven by a novel loss function, which is a continuous relaxation of a well-defined discrete formulation of weakly supervised learning. Care is taken to ensure that the loss is numerically well-posed. Additionally, we propose a neural network architecture which accounts for both patch dependence, through the use of Conditional Random Field layers, and shift-invariance, through the inclusion of anti-aliasing filters. **Results:** We demonstrate our method on the task of localizing thoracic diseases in chest X-ray images, achieving state-of-the-art performance on the ChestX-ray14 dataset. We further show that with a modicum of additional effort our technique can be extended from object localization to object detection, attaining high quality results on the Kaggle RSNA Pneumonia Detection Challenge. **Conclusion:** The technique presented in this paper has the potential to enable high accuracy localization in regimes in which annotated data is either scarce or expensive to acquire. Future work will focus on applying the ideas presented in this paper to the realm of semantic segmentation.

INDEX TERMS Weakly supervised learning, object localization, deep learning, X-ray, limited annotation.

I. INTRODUCTION

Large-scale labelled datasets are one of the key ingredients in many recent algorithms in image processing and computer vision. The combination of such datasets with deep learning techniques has resulted in state-of-the-art (SOTA) algorithms in many tasks, including classification, detection, and segmentation. However, a problematic aspect of the standard deep learning approach is the cost of labelling, particularly in localization tasks – either detection or segmentation. In the case of detection, one wishes to find an object in an image by placing a bounding box around it; in the more fine-grained task of segmentation, one wishes to localize an object with pixel-level granularity. Generally, in order to use deep learning to perform either of these tasks in standard fashion, one requires a fair amount of images with annotations that mirror the desired output: bounding boxes in the case of detection, and pixel-level masks in the case of segmentation. The major problem is that collecting such annotations can be very expensive. Indeed, these annotations are much more expensive than their counterparts in the corresponding classification task, in which the annotator must simply specify a label for the image.

Our goal in this paper is to learn to perform localization with considerably fewer annotated examples. In particular, we consider the following setting: only a very small number

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

of examples have bounding box or segmentation mask labels, while relatively cheap whole image (i.e., classification-style) labels are available for each image in the dataset. Algorithms that successfully solve this kind of problem have wide applicability in computer vision, but would show perhaps their strongest impact in medical imaging. This is due to the fact that while annotation is expensive generally, it is even more costly in the medical setting where the annotator must generally be a physician.

A particular application which might benefit from this approach is the detection of thoracic diseases within chest X-rays. In general, this kind of detection is known to be a complicated task for radiologists, as a scan can potentially contain varying patterns each of which matches several different pathologies. Indeed, there can be considerable variability in the interpretation of chest radiographs, even amongst experts [1], leading to lower reliability of these findings [2], [3]. Thus, systems based on computer-aided diagnosis (CAD) are desirable, as they can perform automatic detection of disease and pathologies in a consistent way, and perhaps with greater accuracy than human experts. To this end, Rajpurkar *et al.* proposed a deep-learning-based algorithm [4] which outperforms radiologists in disease classification on the ChestX-ray14 dataset [5]. However, as noted above, standard CAD systems are costly to train due to the need for large amounts of annotation.

In formulating a localization algorithm trained with limited annotation, our point of departure is the approach of Li *et al.* [6]. This approach is in the spirit of multiple instance learning and has achieved SOTA results on the ChestX-ray14 dataset. However, the method has a number of shortcomings, both in terms of the underlying probabilistic model – which, for example, assumes patch independence – as well in terms of the numerical problems that arise from the formulation. To address these shortcomings, we propose a new technique which resolves these problems through the introduction of a novel loss function as well as a new architecture.

In particular, the main contributions of this paper are as follows:

1) We propose a novel loss function for object localization with limited annotation. This loss is a continuous relaxation of a well-defined discrete formulation of weakly supervised learning, and is numerically well-posed.
2) We propose a new architecture for localization which accounts for both patch dependence and shift-invariance, through the inclusion of Conditional Random Field (CRF) layers and anti-aliasing filters, respectively.
3) We validate our technique on the problem of localizing thoracic diseases in chest X-rays, achieving SOTA performance on the ChestX-ray14 dataset.
4) We show to how to extend our localization technique to the problem of object detection.

The remainder of the paper is organized as follows. Section II reviews related work. Section III formulates the problem of localization with limited annotation and reviews the approach of Li *et al.* [6]. Section IV describes our novel loss function, focusing on its advantageous numerical properties. Section V proposes the new network architecture, detailing both the CRF layers and anti-aliasing filters. Section VI discusses various aspects of the algorithm's implementation. Section VII presents the experiments, including a discussion of the data, results and ablation studies. Section VIII concludes the paper.

We note that a preliminary version of this paper was presented at the Machine Learning for Health Workshop at NeurIPS 2019 [7].

II. RELATED WORK

Thoracic Disease Localization: As our results are demonstrated on the ChestX-ray14 dataset of Wang *et al.* [5], we begin with a brief discussion of this dataset and corresponding algorithmic research. The dataset is a collection of over 100K front-view X-ray images. Using automatic extraction methods from the associated radiological reports based on natural language processing, each image is labelled with up to 14 different thoracic pathology classes; in addition, a small number of images with pathologies are manually annotated with bounding boxes for a subset of 8 of the 14 diseases. Together with the release of the dataset, Wang *et al.* [5] presented the first benchmark for classification and localization by a weakly supervised convolutional neural network (CNN) architecture. This benchmark only used the whole image labels for training and ignored the bounding box annotations.

Following the release of the dataset and initial benchmark, several works proposed more sophisticated networks for more accurate classification or localization results. Yao *et al.* [8] leveraged the inter-dependencies among all 14 diseases using long short-term memory (LSTM) network for disease identification, outperforming Wang *et al.* [5] on 13 of 14 classes. Rajpurkar *et al.* [4] proposed classifying multiple thoracic pathologies by using a 121-layer Dense Convolutional Network (DenseNet) [9], yielding SOTA results for the classification task for all 14 diseases. Unlike both previous methods, which do not exploit any of the bounding box annotations, Li *et al.* [6] took advantage of these annotations to simultaneously perform disease identification and localization through the same underlying model. Although their method did not surpass other methods on the classification task, they did achieve a new SOTA for localization. Subsequent works [10], [11] have focused on improving the SOTA in classification; by contrast, our focus is localization rather than classification, so we shall not elaborate further on these results.

Object Detection: Another generally related area of interest is object detection. Current SOTA object detectors are of two types: two-stage or single-stage. The two-stage family is represented by the "regions with CNN features" (R-CNN) framework [12], comprised of a region proposal stage followed by the application of a classifier to each of these candidate proposals. This architecture, through a sequence of advances [13]–[15], consistently achieves SOTA results on the challenging COCO benchmark [16]. The initial single-stage detectors, such as YOLO [17] and SSD [18], exhibited greater run-time speed at the expense of some accuracy. More recently Lin *et al.* proposed RetinaNet [19], whose training is based on the ''focal loss''; this network was able to match the speed of previous single-stage detectors while surpassing the accuracy of all existing SOTA two-stage detectors. These detection approaches, however, are not aimed at tasks that contain a small number of annotated examples, as in our setting of interest, and are often prone to low accuracy on such datasets. Other works have directly addressed the challenging task of weakly supervised object detection: by leveraging prior knowledge to impose constraints or regularizers on the model architecture [20], [21]; through the use of reinforcement learning to gradually mine desirable object regions under a region searching paradigm [22]; and by integrating multiple instance learning and self-paced learning within the same framework [23].

Multiple Instance Learning: We also note the multiple instance learning (MIL) literature [24] as a type of weakly supervised learning. MIL is a very promising approach to the setting in which the goal is to localize or detect objects within images, but only whole image labels are available for training purposes. The approach treats an image as a *bag of patches* (instances); the image is considered negative if all of the patches are negative and positive if at least one patch is labeled positive. A variety of applications are possible using this approach. Recent applications include a progressive learning framework for weakly supervised object detection [25], localization of action segments in video [26], and algorithms within the fields of geoscience and remote sensing [27], [28]. The approach has also been adopted in medical imaging. In particular, several studies have combined MIL with CNNs, utilizing local patch information in weakly supervised tasks. For example, Yan *et al.* [29] have used the MIL framework to reveal which local regions are discriminative, using patches to better localize body part identifiers in CT slices, without manual annotations, and with only image-level labelling. Zhu *et al.* [30] have eliminated the need for costly annotation of training data using Deep MIL for mass classification based on whole mammograms. Hou *et al.* [31] have trained a CNN classifier that aggregates patch-level predictions to automatically locate discriminative patches within a whole slide tissue image, by formulating a novel Expectation-Maximization (EM) MIL based algorithm. Schwab *et al.* [32] took advantage of MIL to improve the explainability of their detection algorithm, by jointly performing classification and localization of critical findings in X-rays. In our work, in order to incorporate examples that are labelled for a classification task (i.e. with whole image labels) and do not possess bounding box annotations, we utilize the MIL approach to enforce a variant of the constraint that a positive image must contain at least one patch that belongs to the corresponding disease.

Conditional Random Fields: Finally, we mention Conditional Random Fields (CRFs) that are often used for structured prediction. Unlike standard classifiers that predict labels for each pixel/patch without explicit regard for the labels of other pixels/patches, the CRF explicitly takes account of the neighboring samples by means of a graphical model. Dealing with dense graphs was initially considered problematic from a complexity point of view, but a novel and efficient solution to this problem was proposed by Krähenbühl and Koltun [33]. In this work, the pairwise edge potentials are defined by a linear combination of Gaussian kernels. In a more recent version, the edge potentials are learned parametrically [34]. In either case, the CRF is optimized via a series of mean-field iterations. In subsequent years, other densely connected CRFs that can be learned end-to-end as part of neural networks have been reported with corresponding improved accuracy and complexity. For example, Chen *et al.* [35], [36] proposed appending a fully connected CRF to the final network layer, yielding improvements in semantic segmentation accuracy. In another instance, Bhatkalkar *et al.* [37] presented a way to improve the performance of a CNN for optical disk segmentation in fundus images by the inclusion of CRFs in the model. Finally, we note that the main problem of incorporating CRFs into neural networks is the slow training and inference speeds that result; more recent works have addressed this important issue [38], [39].

In our particular case, we are interested in localization. Thus, we would like to introduce the same spatial dependency between patches in an explicit manner. To this end, we rely on the recent pixel-adaptive convolution (PAC) approach of Su *et al.* [39], due to its simplicity, excellent performance, and ability to learn end-to-end in an efficient manner.

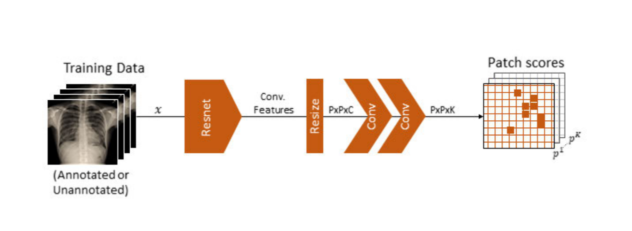


FIGURE 1. **Base model overview [6]. Input images are processed by a CNN, extracting their feature maps. The latter are then resized and processed by two subsequent convolutional layers to finally output a $P \times P \times K$ tensor of patch scores.**

III. PROBLEM FORMULATION

A. THE BASE MODEL

As our starting point, we take the approach of Li *et al.* [6], which proposes a technique for the classification and localization of abnormalities in radiological images. This approach is very appealing in that it allows for localization to be achieved with a very limited number of bounding box annotations. We now give a brief summary of the technique. The architecture used in [6] is shown in Figure 1.

A preact-ResNet network [40], with the final classification layer and global pooling layer removed, is used as the backbone; this part of the architecture encodes the images into a set of C feature maps. These feature maps are subsequently divided into a $P \times P$ grid of patches. Through an application of two convolutional layers (including batch normalization and ReLU activation), the number of channels is modified to K, where K is the number of possible disease types. A per-patch probability score for each disease class is then derived by the application of a sigmoid function; this is denoted p_j^k, where the probability is that the j^{th} patch of the image belongs to class k. Note that a sigmoid function is applied, rather than a softmax, as a particular patch may belong to more than one disease.

As mentioned above, it is assumed that some images have bounding box annotations, while most do not. Let us define some terms: the image is x; for a disease k, the label $y^k = 1$ if the disease is present, $y^k = 0$ otherwise; if the disease k is annotated with a bounding box b^k, then $a^k = 1$, otherwise $a^k = 0$. We note that in practice, the bounding box b^k is produced by mapping the manually annotated bounding box coordinates to their nearest locations on the $P \times P$ grid. As a result, b^k is a subset of the P^2 patches. Now, the loss function can then be broken into two cases, in terms of whether a bounding box annotation is supplied or not. In the case in which there is a bounding box for a disease of class k, i.e. $a^k = 1$, the annotated loss is taken to be

$$L_{ann}^k = -\log p(y^k = 1|x, b^k) \tag{1}$$

where $p(y^k = 1|x, b^k)$ denotes the probability that disease k is within bounding box b^k of image x, and is given by

$$p(y^k = 1|x, b^k) = \prod_{j \in b^k} p_j^k \prod_{j \in \bar{b}^k} (1 - p_j^k) \tag{2}$$

and $\bar{b}^k$ is the complement of bounding box b^k. The above formula is simply the standard formula for combining independent patch probabilities. In the case in which no bounding box is supplied, i.e. $a^k = 0$, the unannotated loss is

$$L_{un}^k = -y^k \log p(y^k = 1|x) - (1 - y^k) \log(1 - p(y^k = 1|x)) \tag{3}$$

where

$$p(y^k = 1|x) = 1 - \prod_j (1 - p_j^k) \tag{4}$$

The latter probability is simply the probability that there is at least one patch with disease k, again assuming independence of patches. Finally, the overall loss per image is

$$L = \sum_k \left(\lambda_{ann} a^k L_{ann}^k + (1 - a^k) L_{un}^k \right) \tag{5}$$

We refer to this model – the architecture and the loss – as the *base model*. It was shown to attain SOTA performance in terms of localization on the NIH Chest X-ray dataset [5].

B. ISSUES WITH THE BASE MODEL

The probabilistic formulation of Li *et al.* is a nice approach to localization tasks with a very limited number of bounding box annotations. In spite of this, there are several issues with the technique that limit its performance:

(i) **Single Patch ⇒ Positive Declaration:** In Equation (4) of the above derivation for the unannotated loss, only a single patch needs to be positive for a positive disease detection within the image. In general, this assumption is prone to false positives. One would like for multiple patches to be present for a declaration; in particular, a single positive detection could easily be caused by noise. We would, therefore, like to do away with this assumption, by using a novel loss function.

(ii) **Numerical Issues:** The paper refers to a particular numerical problem that results from the multiplication of many small numbers, as in Equations (2), (4). This numerical underflow is fatal to the approach, and the problem is circumvented in [6] through a series of unjustified heuristics as they normalized the patch probabilities from [0,1] to [0.98,1]; resulting in patch-level "probability scores" that do not necessarily reflect the meaning of probability anymore. We propose a formulation of the loss in which these issues never arise.

(iii) **Patch Independence:** In both Equations (2) and (4), the probabilities of each patch containing an object of a particular class (p_j^k) are treated as independent between patches. This is not correct in practice as we would like to integrate more elaborate terms that model contextual relationships between object/patch classes. We solve this through the use of a Conditional Random Field model.

(iv) **Lack of Shift-Invariance:** As has been pointed out by Zhang [41], modern CNNs are not technically shift-invariant. In order to improve the performance of the localization, one can, therefore, address this issue through the addition of anti-aliasing filters prior to downsampling, as suggested in [41].

To summarize, our technique is related to the base model but is differentiated in two key ways:

- The use of a novel loss function, which addresses many of the aforementioned issues.
- Modifications to the architecture, specifically (a) the incorporation of Conditional Random Field layers and (b) the inclusion of anti-aliasing filters.

We now elaborate on each of these, in turn.

IV. THE NEW LOSS FUNCTION

A. NOTATION

As described above, the output of the base model is a tensor of shape the of $P \times P \times K$; we will continue to denote the output as p_j^k, the probability is that the j-th patch of the image belongs to class k. This output will then be fed into a series of layers that implement a CRF model, with p_j^k representing the unary terms in the CRF. The output of the CRF is denoted as z_j^k. We will discuss the details of the CRF model in Section V; for now, we may think of z_j^k as a sharper estimate of whether a particular disease k is present in patch j. z^k indicates the length P^2 vector of all patch values for a given disease k.

B. THE LOSS FUNCTION: FIRST PASS

We first consider an annotated example with $a^k = 1$, i.e. one with a bounding box. As described previously , b^k, which consists of a subset of the patches, contains an object of class k. In this case, the following loss function is natural:

$$\begin{aligned} L_{ann}^k = -\mathbb{I}[z^k &\text{ contains a blob of size} \geq \tau^k|b^k| \text{ within } b^k \\ &\text{AND contains blobs of total size} \leq \rho^k|\bar{b}^k| \text{ within } \bar{b}^k] \end{aligned} \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function; $\tau^k, \rho^k \in [0, 1]$ are thresholds; and $\bar{b}^k$ is the complement of the bounding box b^k. A blob may be made precise as a connected component; however, this will not lead to a nice differentiable loss. So we make the following continuous relaxation of the above discrete formulation:

$$\begin{aligned} L_{ann}^k = &-\sigma(\mathbf{1}^T(\mathbb{I}[b^k] \odot z^k) - \tau^k|b^k|) \\ &\cdot \sigma(\rho^k|\bar{b}^k| - \mathbf{1}^T(\mathbb{I}[\bar{b}^k] \odot z^k)) \end{aligned} \tag{7}$$

where $\mathbb{I}[b^k]$ is now an indicator function on the bounding box; $\odot$ is the Hadamard product; $\mathbf{1}$ is the vector of all 1's; and σ is a sigmoid function, i.e. a smooth approximation to the indicator function. What this says is that there must be a total of $\tau^k|b^k|$ patches within b^k which detect class k; the relaxation is that the total no longer has to be in a single connected component. This is a reasonable relaxation, especially since the CRF already encourages smoothness. In addition, we require that there be fewer than $\rho^k|\bar{b}^k|$ patches outside of b^k which detect class k.

Note that this logic extends in a straightforward manner to the case of an unannotated example with $a^k = 0$, when there is no bounding box specified. In the case of a positive example (i.e. one in which disease k is present), the loss is simply

$$L_{un|pos}^k = -\sigma(\mathbf{1}^T z^k - \hat{\tau}^k) \tag{8}$$

so that the threshold $\hat{\tau}^k$ now has a meaning in absolute terms, i.e. the absolute number of patches vs. the number of patches relative to the size of a bounding box. In the case of a negative example – where disease k is absent – we have an equation analogous to (8):

$$L_{un|neg}^k = -\sigma(\hat{\rho}^k - \mathbf{1}^T z^k) \tag{9}$$

where $\hat{\rho}^k$ is another threshold, whose meaning is in terms of the absolute number of patches, similar to $\hat{\tau}^k$. Finally, we can combine Equations (8) and (9) to get

$$L_{un}^k = -y^k\sigma(\mathbf{1}^T z^k - \hat{\tau}^k) - (1 - y^k)\sigma(\hat{\rho}^k - \mathbf{1}^T z^k) \tag{10}$$

C. ADDRESSING ISSUES (I) AND (II)

The above formulation addresses Issues (i) and (ii) raised in Section III. Regarding Issue (i), Equation (8) requires more than a single patch in order to make a positive declaration; the number of patches must be equal to $\hat{\tau}^k$, which is a per-class parameter that can be chosen. Regarding Issue (ii), neither Equations (7) or (10) involve the multiplication of many small values; indeed, both are well-posed from a numerical point of view.

D. DEALING WITH VANISHING GRADIENTS

Due to the presence of sigmoid functions in Equations (7) and (10), in practice, we experience issues of vanishing gradients during training. We propose the following remedy, based on a different relaxation. We replace Equation (7) with

$$\begin{aligned} L_{ann}^k = &|b^k|^{-1} \text{ ReLU}(\tau^k|b^k| - \mathbf{1}^T(\mathbb{I}[b^k] \odot z^k)) \\ &+ |\bar{b}^k|^{-1} \text{ ReLU}(\mathbf{1}^T(\mathbb{I}[\bar{b}^k] \odot z^k) - \rho^k|\bar{b}^k|) \end{aligned} \tag{11}$$

Note that there are three fundamental differences between Equations (7) and (11). First, we have replaced the sigmoid functions, σ, with ReLU functions; this has the effect of still leading to minimal loss (in this case, zero) once the constraints are satisfied, but leads to more nicely behaved gradients. Second, we have replaced the multiplication with addition. Once sigmoids have been replaced by ReLU's, the notion of a "fuzzy AND" relaxation is no longer relevant; in this case, an addition makes more sense, and again leads to better-behaved gradients. Finally, sigmoids are scaled between 0 and 1, whereas ReLU's can grow without bound; this necessitates the insertion of a scaling factor of $|b^k|^{-1}$ and $|\bar{b}^k|^{-1}$, to ensure that the two terms in the sum are properly balanced.

Similarly, for unannotated examples we replace Equation (10) with:

$$L_{un}^k = y^k \text{ ReLU}(\hat{\tau}^k - \mathbf{1}^T z^k) + (1 - y^k) \text{ ReLU}(\mathbf{1}^T z^k - \hat{\rho}^k) \tag{12}$$

The thresholds $\tau^k, \hat{\tau}^k, \rho^k, \hat{\rho}^k$ can be treated as parameters of the network, that can be optimized during training, or can be considered hyperparameters. Note that the losses described in Equations (11) and (12) do not suffer from any numerical issues. This is due to the fact that they are not the product of many individual probabilities; rather, they aggregate information across patches in such a way that the resulting loss is numerically stable.

E. BALANCING FACTORS AND THE FINAL LOSS FUNCTION

There are two sources of data imbalance to account for. The first is the large imbalance of negative (non-diseased) vs. positive (diseased) examples in the data. To deal with this, we modify Equation (12) slightly, to read

$$L_{un}^k = y^k \text{ ReLU}(\hat{\tau}^k - \mathbf{1}^T z^k) + \gamma(1 - y^k) \text{ ReLU}(\mathbf{1}^T z^k - \hat{\rho}^k) \tag{13}$$

where γ is the ratio of positive to negative examples in the data. The latter de-emphasizes negative examples relative to the positive examples.

The second form of imbalance we must account for is that between the annotated and unannotated examples, in case both are used in training. In practice, there are usually many more unannotated examples available. However, this is already accounted for by the factor λ_{ann} in Equation (5), which is set to a value greater than one. Combining our own annotated and unannotated losses in Equations (11) and (13), respectively, using Equation (5), we arrive at the final form of the per-example loss:

$$L(x, y, a, b) = \sum_k L^k(x, y, a, b) \tag{14}$$

where

$$\begin{aligned} L^k =\ & \lambda_{ann}\, a^k \Big[|b^k|^{-1}\ \mathrm{ReLU}(\tau^k |b^k| - \mathbf{1}^T(\mathbb{I}[b^k] \odot z^k(x))) \\ & + |\bar{b}^k|^{-1}\ \mathrm{ReLU}(\mathbf{1}^T(\mathbb{I}[\bar{b}^k] \odot z^k(x)) - \rho^k |\bar{b}^k|) \Big] \\ & + (1 - a^k) \Big[y^k\, \mathrm{ReLU}(\hat{\tau}^k - \mathbf{1}^T z^k(x)) \\ & + \gamma(1 - y^k)\, \mathrm{ReLU}(\mathbf{1}^T z^k(x) - \hat{\rho}^k) \Big] \end{aligned} \tag{15}$$

and the dependence of the z variables on the image x has been made explicit.

V. ARCHITECTURAL MODIFICATIONS

A. CRF MODEL

As mentioned in Issue (iii) in Section III, we would like to do away with the assumption of patch independence. Indeed, in our derivation of Equation (15), we did not make use of such assumptions. However, to further bolster the dependence between neighboring patches, we introduce a CRF model into our network. The CRF introduces, in an explicit manner, a spatial dependency between patches. The effect of the CRF is to increase the confidence for a given patch's predicted label, and thereby to improve localization. There are several choices amongst neural network-compatible CRFs; we choose the recent pixel-adaptive convolution (PAC) approach of Su *et al.* [39], due to its simplicity and excellent performance. We thus integrate the PAC-CRF modification to our base network and train the model end-to-end.

Given the patch probability outputs p_j^k of the base model, the unary potentials of the CRF are simply taken as the $\psi_j^{(u)}(k) = p_j^k$. Thus, in the absence of neighbor dependence, the CRF will simply choose $z_j^k = p_j^k$. Neighbour dependence may be introduced through pairwise potentials. As in [39], we take this potential to be $\psi_{jl}^{(p)}(k_j, k_l) = G(f_j, f_l) W_{k_j k_l} (\xi_j - \xi_l)$, where f are a set of learnable features on the $P \times P$ grid; G is a fixed Gaussian kernel; ξ_j is the pixel coordinates of patch j; and W is the inter-class compatibility function, which varies across different spatial offsets, and is also learned. The pairwise connections are defined over a fixed window Ω around each patch. As our unary model outputs a $P \times P \times K$ tensor, we insert two 2D convolution layers immediately prior to the PAC-CRF model ; each such layer is followed by a rectified linear unit and batch-normalization. The output of this part of the network is a tensor with the same size as the input image. Please refer to Figure 2.

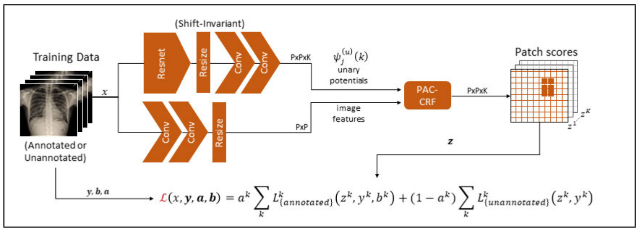

(Shift-Invariant)
Training Data
(Annotated or Unannotated)
x
Resnet
Resize
Conv
Conv
PxPxK
$\psi_j^{(u)}(k)$
unary potentials
Conv
Conv
Resize
PxP
image features
PAC-CRF
PxPxK
Patch scores
z^1
z^K
z
y, b, a
$\mathcal{L}(x, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = a^k \sum_k L^k_{\{annotated\}}(z^k, y^k, b^k) + (1 - a^k) \sum_k L^k_{\{unannotated\}}(z^k, y^k)$

FIGURE 2. The network architecture. In the diagram, the image is x; for a disease k, the label $y^k = 1$ if the disease is present, $y^k = 0$ otherwise; if the disease k is annotated with a bounding box b^k, then $a^k = 1$, otherwise $a^k = 0$. (y and a are one-hot vectors.) Our network consists of two branches: the upper one has the same architecture as in Figure 1, though modified with shift-invariant anti-aliasing filters, and computes the unary terms of the CRF model. The lower branch extracts features from the input images to form a feature-tensor of the same size ($P \times P$) as the unary terms, which are used in the pairwise terms. Both enter into the PAC-CRF [39], outputting a $P \times P \times K$ tensor of patch scores $\{z^k(x)\}_{k=1}^K$ for the input image x. The latter are used to calculate the loss function in equations (14) and (15).

B. ANTI-ALIASING

An important property of any model whose goal is to perform localization or segmentation is that the output of the model should be shift-invariant with regard to its input. However, as Zhang [41] has noted, standard CNNs use downsampling layers while ignoring sampling theorem, and are therefore not shift-invariant; this is in spite of the fact that CNNs are commonly used as the backbone of many localization/segmentation tasks. To circumvent this problem, an anti-aliasing filter is required prior to every downsampling part of the network. In particular, Zhang [41] proposed the insertion of a blur kernel as a low-pass filter prior to each downsampling step in the network, and thereby demonstrated an increased accuracy across several commonly used architectures and tasks. Following [41], we thus modify the backbone of our base model and integrate such low-pass filters as part of the preact-ResNet network [40]. This effectively addresses Issue (iv) in Section III.

VI. IMPLEMENTATION DETAILS

The flow of our overall training regime is illustrated in Algorithm 1 and Figure 3; we now describe the training procedure in more detail. In Stage I, we train the unary terms using the model without the PAC-CRF until convergence. Once Stage I is complete, we freeze this model; in Stage II, we train only the PAC-CRF part. It then remains to learn the per-class thresholds τ^k, $\hat{\tau}^k$, ρ^k, $\hat{\rho}^k$. These thresholds are related to the empirical distribution of the areas occupied by the corresponding pathology within the image. It is natural

Algorithm 1 Training procedure

```
Initialize and freeze the per-class thresholds: $\tau^k$, $\hat{\tau}^k$, $\rho^k$, $\hat{\rho}^k$

Stage I:
    • Train the model without PAC-CRF to obtain unary terms $\rightarrow \psi_j^{(u)}(k)$
Stage II:
    • Freeze $\psi_j^{(u)}(k)$
    • Train PAC-CRF module to obtain patch probabilities $\rightarrow z_j^k$
while per-class thresholds have not converged do
    Freeze the model and train $\tau^k$, $\hat{\tau}^k$, $\rho^k$, $\hat{\rho}^k$
    Freeze $\tau^k$, $\hat{\tau}^k$, $\rho^k$, $\hat{\rho}^k$ and repeat Stage II
end
```

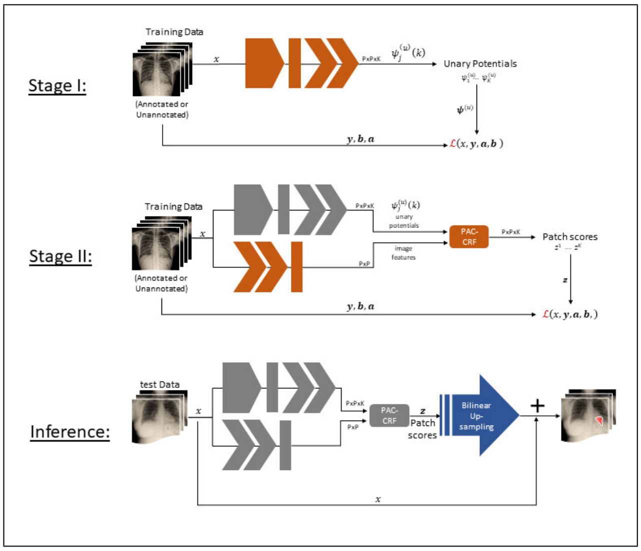


FIGURE 3. The stages of the training regime described in Algorithm 1. For reference, the overall architecture of the network is shown in detail in Figure 2. For each stage, the blocks marked in orange are those whose parameters are optimized within that stage. By contrast, the blocks marked in gray have their parameters frozen (i.e. not optimized) for that stage. The block marked in blue represents the post-processing stage of upsampling at inference time.

to treat those thresholds as parameters of the network, which can be optimized during training, thereby allowing them to adapt to the area distributions. Unfortunately, this quickly leads to the degenerate solution, i.e: $\tau, \hat{\tau} \rightarrow 0$, $\rho \rightarrow 1$ and $\hat{\rho} \rightarrow P^2$. In order to avoid reaching the degenerate solution, we employ the following procedure: we begin by freezing the thresholds, and training only the network weights; we then freeze the network weights, and find the optimal thresholds; we continue to alternate this procedure until convergence. At inference time we simply use the trained model to obtain patch probability scores of the objects, and then use bilinear upsampling to scale back to the image dimensions.

We now discuss the complexity of various parts of our model, as expressed in terms of the number of parameters. The entire model contains 35M parameters. The dominant component is the ResNet50 model, which is responsible for 23M parameters. The next largest component derives from the convolutional layers which are appended to the end of the ResNet model and are responsible for the unary terms, see the upper branch in Figure 2; these account for 10M parameters. The remainder of the parameters (2M) are due to the pairwise terms (lower branch in Figure 2) and the PAC CRF model. The latter number is quite small, which is especially true when the patch grid is of low resolution, which is the case in all of our experiments (we use 20 × 20). Finally, note that the inclusion of anti-aliasing filters does not add any parameters. In terms of the computational complexity (i.e. algorithm run-time), the same conclusions apply: nearly all of the run-time is due to the use of the ResNet backbone, while the CRF adds a negligible computation overhead.

VII. EXPERIMENTS

A. DATASET

We have evaluated our model on the NIH Chest X-ray dataset [5]. The NIH Chest X-ray dataset consists of 112,120 frontal-view X-ray images with a resolution of 1024 × 1024, and annotated with 14 disease labels (each image can have multiple labels). 60,361 of the images (about 54%) are labelled as having "no finding"; out of the remaining 51,759 images which have at least one disease label, only 880 images have bounding box annotations. To summarize: only 0.8% of images in the dataset possess bounding boxes.

Images can have more than one label or one bounding box, so there are 81,176 class labels and a total of 984 labelled bounding boxes. The 984 bounding boxes annotations are only given for 8 of the 14 disease types. Refer to Table 1 for further details.

TABLE 1. The distribution of class-labels and bounding-box annotations for each disease in the NIH Chest X-ray dataset. No-finding is the case in which no disease was detected.

Class	# Labels	# Bounding Boxes	% Bounding Boxes
Atelectasis	11,559	180	18.3 %
Cardiomegaly	2,776	146	14.8 %
Effusion	13,317	153	15.6 %
Infiltration	19,894	123	12.5 %
Mass	5,782	85	8.6 %
Nodule	6,331	79	8.0 %
Pneumonia	1,431	120	12.2 %
Pneumothorax	5,302	98	10.0 %
Consolidation	4,667	**Total:** 984	100 %
Emphysema	2,516		
Pleural Thickening	3,385		
Edema	2,303		
Fibrosis	1,686		
Hernia	227		
No-Finding	60,361		

B. MODEL SETTINGS

We use ResNet-50 as the backbone of our model, and take $P = 20$ for a 20 × 20 patch grid. Both of these selections are made to ensure a fair comparison vis-a-vis Li's model [6], which has produced SOTA results for localization on the NIH Chest X-ray dataset. More specifically, our backbone is a preact-ResNet-50 network [40] (as in [6]). We initialize the network weights based on ImageNet [42] pre-training and then allow them to evolve as training proceeds. The images are 1024 × 1024, but we have resized them to 512 × 512 for faster processing; we have also normalized the image range to $[-1, 1]$, as we found this led to faster convergence. We take the batch size to be 48. Hyperparameter tuning yields the following settings: $\lambda_{ann} = 70$, an exponentially decaying learning rate initialized to 0.001, and use of the ADAM optimizer [43] accompanied by a weight decay regularization coefficient equal to 0.01. Finally, for the PAC-CRF module we use a 19 × 19 PAC filter, to accommodate the 20 × 20 patch grid resolution.

C. EVALUATION METRICS

We are interested in localization accuracy, so we evaluate solely over annotated examples. We use Intersection-over-Union (IoU) and Intersection-over-Region (IoR) to measure localization accuracy. A patch is taken to be positive, i.e. the disease k is present in patch j, if its value is greater than 0.5: $z_j^k \geq 0.5$. The union of all positive patches is the detected region. The IoU and IoR can then be computed between the ground truth bounding box and the detected region.

A localization is taken to be correct if IoU $\geq T$ or IoR $\geq T$ for given threshold T; following the practice of [6], we use $T = 0.1$. Performance statistics are then computed over 5-fold cross-validation of the annotated examples.

D. EVALUATION METHODOLOGY

We compare our results with those of Li *et al.* [6], which represent the SOTA for the localization task on the NIH Chest X-ray dataset; and those of Wang *et al.* [5], that presented the first benchmark for this task. We examine three separate settings: (a) the model is trained using only 80% of the annotated examples, with the 80% representing the training part of the fold; (b) and (c) the model is trained using 80% of the annotated examples as described in (a), as well as 10% or 20%, respectively, of the unannotated examples, representing about 10k or 20K examples (selected randomly). More specifically, in settings (b) and (c) the model is pre-trained for the unary terms of setting (a), and then all parameters are allowed to evolve as training proceeds. In all three settings, the results are evaluated on the remaining 20% of the annotated examples of each fold. This specification was chosen to agree with that of [6] and to make a relatively fair comparison with [5] that evaluated over the entire annotated dataset – since we use 5-fold cross-validation, the complete set of annotated images has been evaluated. Note that the division of images is made at the patient level, keeping unique patients in each fold.

We also compare our localization results with a state of the art detection network RetinaNet [19] on the same NIH Chest X-ray dataset. RetinaNet exploits only the annotated examples and is not generally intended to exploit the positive unannotated examples to improve detection. Thus RetinaNet results are compared with ours only for setting (a) – for annotated samples only (i.e 0% unannotated samples). The implementation [44] we used for RetinaNet leads to very similar results as those presented in [19]. We allowed 100 detections per image, while selecting the bounding box with the highest score for each class to measure the IoU and IoR accuracies. Training was stopped at the epoch which possessed the best average validation results over all five folds, i.e. before overfitting the training data. We chose the remainder of the parameters according to the optimal settings in [19]: for example, the focal parameter γ was set to 2, and the weighting factor α was set to 0.25.

E. RESULTS

Overall Results: In Table 2 we present two versions of the localization accuracy, with the first based on IoU and the second based on IoR. The standard deviation of the reported results is shown whenever they are given in their work; best results are shown in bold. Our method outperforms that of Li *et al.*, Wang *et al.* and presents superior results vs. RetinaNet. In particular, we outperform Li *et al.* for every disease class, for both IoU and IoR, as well as for both settings in common - with no extra unannotated data added, and with 20% unannotated data.

Examining our IoU accuracy more closely as compared to that of Li *et al.*, Table 2 reveals several patterns. First, we perform considerably better than Li *et al.* when no unannotated data is added; for example, the accuracy on Atelectasis and Mass is nearly double that of Li *et al.*, whereas the performance on Nodule is five times better. Second, with the addition of unannotated data, the gaps narrow – for example, Nodule is now slightly less than double, and many other disease classes have a smaller gap – but the gap is still present for each disease class. We hypothesize that our improvement is less in most cases simply because the algorithm trained with no unannotated data already has a fairly high performance; thus, the marginal benefit of adding the unannotated data is smaller.

We also present superior results over RetinaNet, as well as more stable results with regard to standard deviation. Wang *et al.* [5] did not train with annotated samples but only used the annotated examples for validation; this may explain why they achieve poorer results. It may also give a good indication as to how crucial a small dataset of annotated examples can be to significantly improving localization results.

The Role of Unannotated Data: In examining our own results, we may see that the addition of unannotated data often helps, but does not always do so. In particular, comparing 0% to 20% shows that there is an increase in localization accuracy for four of the eight diseases – Infiltration, Mass, Nodule, Pneumothorax – while two diseases, Cardiomegaly and Pneumonia, undergo little or no change. The remaining two, Atelectasis and Effusion, actually suffer a decrease in accuracy due to the addition of the extra unannotated data.

TABLE 2. IoU and IoR disease localization accuracy, with 5-fold cross-validation (cv). 80% of the annotated (*ann*[%]) examples were used for training, in addition to either 0%, 10% or 20% of the unannotated (*un*[%]) examples (selected randomly).

IoU Accuracy											
	Train		Test								
Model	*ann*[%]	*un*[%]	*ann*[%]	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
ours	80%	0%	20% (cv)	**0.818±0.05**	**1±0.00**	**0.882±0.05**	0.927±0.03	0.695±0.10	0.404±0.10	0.918±0.07	0.726±0.1
		10%		0.779±0.07	**1±0.00**	0.824±0.06	0.933±0.03	0.684±0.11	0.419±0.12	**0.924±0.06**	0.695±0.08
		20%		0.779±0.08	**1±0.00**	0.843±0.07	**0.945±0.03**	0.709±0.04	**0.444±0.12**	0.915±0.06	**0.732±0.08**
Li	80%	0%	20% (cv)	0.488	0.989	0.693	0.842	0.342	0.081	0.715	0.437
		20%		0.687	0.978	0.831	0.9	0.634	0.241	0.568	0.576
		50%		0.71±0.05	0.98±0.02	0.87±0.03	0.92±0.05	**0.71±0.10**	0.40±0.10	0.60±0.11	0.63±0.09
Wang	0%	70%	100%	0.688	0.938	0.660	0.707	0.400	0.139	0.633	0.377
RetinaNet	80%	0%	20% (cv)	0.546±0.05	0.906±0.18	0.621±0.08	732±0.13	0.454±0.09	0.263±0.33	0.517±0.05	0.536±0.12

IoR Accuracy											
	Train		Test								
Model	*ann*[%]	*un*[%]	*ann*[%]	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
ours	80%	0%	20% (cv)	**0.889±0.04**	**1±0.00**	**0.92±0.08**	0.95±0.03	0.773±0.07	**0.580±0.10**	0.933±0.06	0.767±0.10
		10%		0.884±0.04	**1±0.00**	0.909±0.08	0.951±0.04	0.794±0.13	0.550±0.13	0.930±0.06	0.779±0.08
		20%		0.844±0.07	**1±0.00**	0.896±0.07	**0.967±0.03**	**0.808±0.09**	0.520±0.12	**0.935±0.07**	**0.806±0.10**
Li	80%	0%	20% (cv)	0.528	**1.000**	0.753	0.875	0.452	0.111	0.786	0.473
		20%		0.724	0.991	0.874	0.921	0.674	0.271	0.644	0.624
		50%		0.77±0.06	0.99±0.01	0.91±0.04	0.95±0.05	0.75±0.08	0.40±0.11	0.69±0.09	0.68±0.10
Wang	0%	70%	100%	0.622	**1.000**	0.797	0.911	0.588	0.152	0.858	0.520
RetinaNet	80%	0%	20% (cv)	0.587±0.05	0.906±0.18	0.643±0.09	0.765±0.15	0.486±0.05	0.273±0.32	0.558±0.07	0.570±0.11

In examining the data in Table 1, the solution to this puzzle becomes apparent: Atelectasis and Effusion have the largest number of annotated examples out of the eight disease classes, with 18.3%, 15.6% out of the 984 bounding boxes annotations. This explains why they have quite high localization accuracies to begin with, when no unannotated data has been added (0.818 and 0.882, respectively); and why the addition of extra unannotated examples does not increase accuracy. On the flip side, Nodule and Mass have the smallest number of annotated examples, with 8% and 8.6% out of the 984 bounding boxes annotations, which explains why adding unannotated data helps the most in these cases. It is interesting to note that Pneumonia has high accuracy and a decent number of annotated examples, 12.2%; the main reason it differs from Atelectasis and Effusion is that it also has a relatively small number of unannotated examples, 1,431 out of all unannotated examples, compared to 11,559 and 13,317 respectively for Atelectasis and Effusion. Thus, the addition of a relatively small number of unannotated examples does not have a strong influence on the accuracy of Pneumonia detection. We note that the IoR data is fairly similar to the IoU data, and most of the observations above hold in this case as well. The exception is Nodule, for which we see a decrease in IoR as we increase the unannotated examples, due to the very small size of the pathology.

Interestingly, despite the fact that the 984 bounding box annotations are only given for 8 of the 14 disease types, we have noticed that for the case in which we incorporate unannotated examples during training, we obtain superior results by learning on the full complement of 14 disease types. This seems to imply some interesting interdependence amongst disease classes.

Qualitative Examples and Failure Modes: Qualitative examples of the algorithm's localizations are shown in Figure 4, where each column represents different pathology, and the last row demonstrates failure modes. These images show good localization results for a highly varied distribution of examples; there is generally a good fit to the bounding box positions, while not overfitting to the bounding boxes' rectangular structure. Some examples are also shown for the case of multiple-class annotations, which the algorithm handles without difficulty.

Regarding the failure modes, i.e. IoU < 0.1, we see a variety of different issues: examples where the position of the disease is incorrectly detected and examples where the area of the disease is too large. We hypothesize that the main reason stems from the small number of annotated examples used for training. In such a case, when the algorithm is unable to correctly identify the location of the disease, the most statistically probable location is chosen. This may lead to a wider localization area covering one or two of the lungs, or cause a bias with localizing the disease. It is possible that using augmentation may do away with this kind of overfitting and reduce these occurrences.

The Per-Class Thresholds: We now discuss the optimal thresholds. For all classes k, τ^k and ρ^k converged to 1 and 0.05 respectively; the former encourages high patch values around the disease and the latter encourages low values outside of the bounding box. Similarly, for all classes k, $\hat{\rho}^k$ converged to 0, encouraging low patch values within X-ray

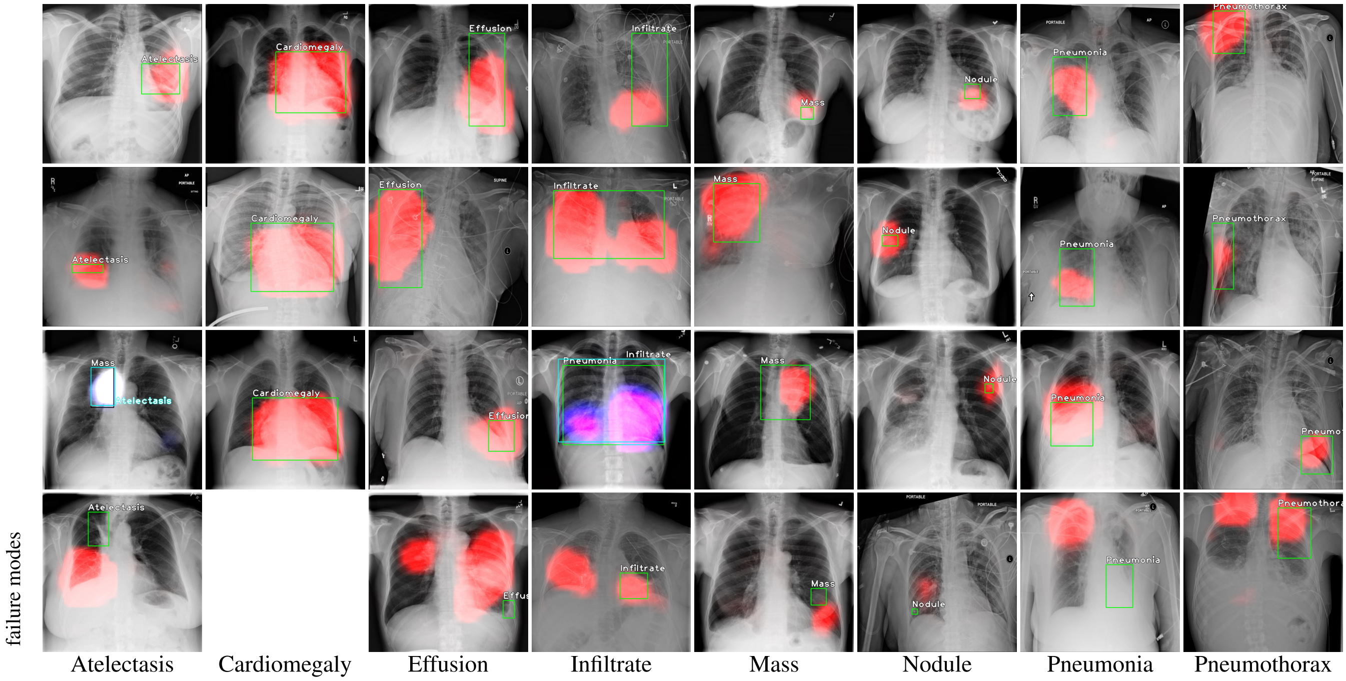


FIGURE 4. Examples of localized pathologies on test images. The colored blob is the result of our localization algorithms (prior to thresholding), while the ground truth is marked by a bounding box. Each column represents a different pathology, indicated in the caption. In images where two diseases are present, there is a double annotation. The last row demonstrates failure modes.

images that do not contain any disease. This is quite intuitive: in the former case, the network is willing to tolerate some appearance of the disease outside of the bounding box; in contrast, in the latter case no disease is present, and the network would prefer to see absolutely no positive patches anywhere. One can similarly explain the value of τ^k converging to 1: our accuracy is measured via the IoU and IoR metrics, both of which encourage positive labelling within the entire bounding box.

The $\hat{\tau}^k$ values converged to more interesting values that are specific to each disease; at convergence, the $\hat{\tau}$ values of Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, and Pneumothorax are 37, 80, 41, 63, 41, 10, 41 and 30 respectively.[1] These values are related to the empirical distribution of the areas occupied by the corresponding pathology within the image. Indeed, the $\hat{\tau}^k$ values are actually very close to the mean area of their bounding boxes (measured on a 20 × 20 patch grid) and therefore possess information regarding the average size of the disease. For example, note the $\hat{\tau}$ values for two diseases, Cardiomegaly (80) and Nodule (10); these values accord with empirical reality, as Cardiomegaly is known to be relatively larger than other diseases while Nodule is relatively smaller.

F. ABLATION STUDY: THE ROLE OF THE CRF MODEL

In Table 3 we measure the influence of adding the CRF submodel into our model and compare localization accuracy with and without it. The results confirm our hypothesis: having spatial dependence between patches and dependence between channels (diseases) leads to an increase in the patch-score confidence for the predicted label. Indeed, we see that the accuracy improves with the CRF model in all cases.

Figure 5 compares the localization outcome with and without adding the CRF model. It can be seen that the CRF model is crucial for better localization performance. In particular: in the case of Effusion, Mass and Pneumonia it eliminates irrelevant segments; in the case of Nodule and Pneumothorax, it refines the localization of the disease; and in the case of Atelectasis and Infiltrate, it more accurately detects the disease.

G. EXTENSION TO DISEASE DETECTION - A PROOF OF PRINCIPLE

With slight modification, our proposed method can be extended to performing traditional detection with bounding boxes (rather than localization) of diseases in chest X-ray images. The goal of this extension is to demonstrate the ability of our model to find rectangular chunks that are spatially well separated. This extension is essentially a proof of principle, as the focus of our work is and remains localization. Nevertheless, it is instructive to see the system in action on a parallel task.

We evaluate our algorithm on a dataset in which the goal is to detect a single disease, namely Pneumonia, whose details we describe shortly. Given the fact that we are now interested in detection, and specifically detection of a single disease,

[1] For the remaining six pathologies, each of which has zero images annotated with bounding boxes, the $\hat{\tau}$ values each converged to 40.

TABLE 3. The influence of the CRF model. IoU and IoR disease localization accuracy, with 5-fold cross-validation (cv). 80% of the annotated (*ann*[%]) examples were used for training, with 0% unannotated examples. +/− denote including/not including the CRF model.

IoU accuracy										
train	test									
ann[%]	*ann*[%]	CRF	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
80%	20%	+	**0.818±0.05**	**1±0.0**	**0.882±0.05**	**0.927±0.03**	**0.695±0.1**	**0.404±0.1**	**0.918±0.07**	**0.726±0.1**
	(cv)	-	0.814±0.07	1±0.0	0.868±0.05	0.919±0.04	0.698±0.09	0.362±0.12	0.884±0.06	0.707±0.1

IoR accuracy										
train	test									
ann[%]	*ann*[%]	CRF	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
80%	20%	+	**0.889±0.04**	**1±0.0**	**0.920±0.08**	**0.95±0.03**	**0.773±0.07**	**0.580±0.1**	**0.933±0.06**	**0.767±0.1**
	(cv)	-	0.889±0.05	1±0.0	0.908±0.08	0.95±0.04	0.735±0.08	0.526±0.14	0.908±0.06	0.735±0.13

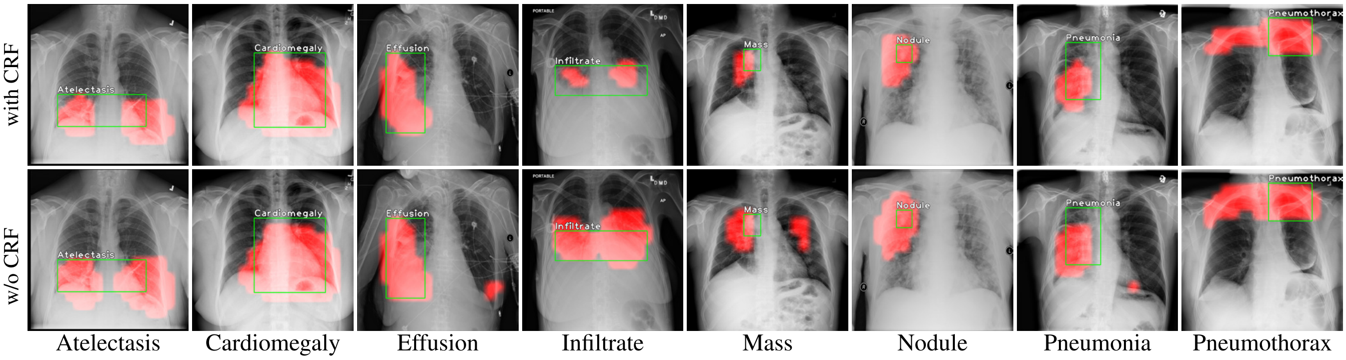


FIGURE 5. Comparison of localized pathologies on test images with (upper row) or without (lower row) CRF included. The colored blob is the result of our localization algorithm, while the ground truth is marked by a green bounding box. Each column represents a different pathology, indicated in the caption.

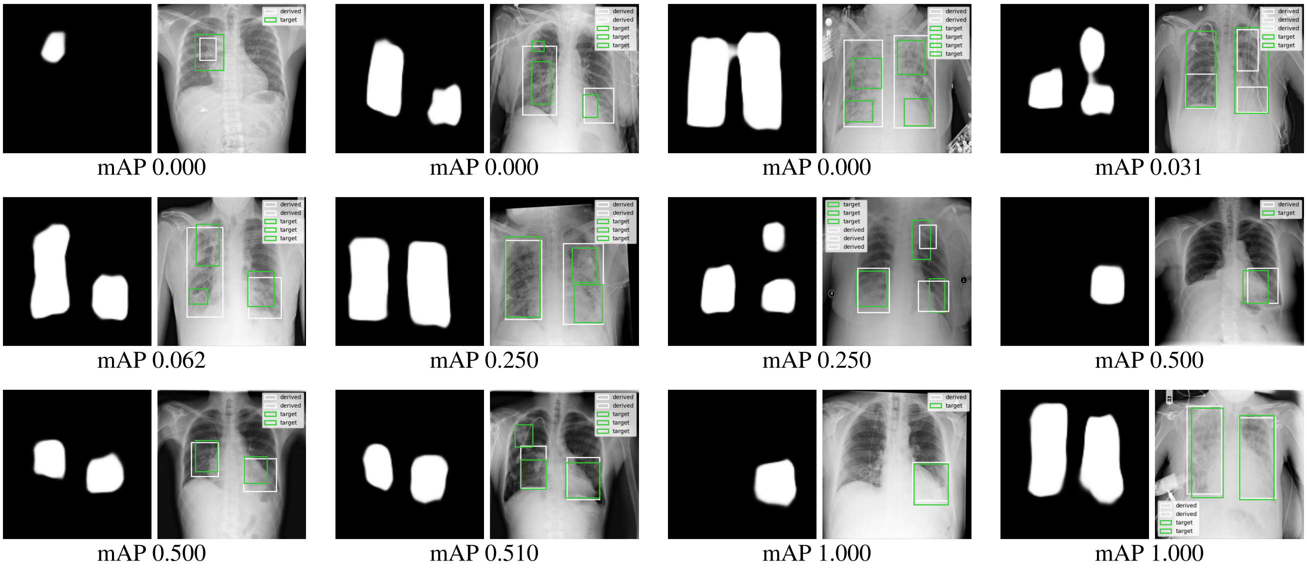
we make three modifications to the structure of our model. (1) We change the parameters of the output tensor – the patch grid is set to 128 × 128 to improve detection separability, and we take the number of channels to be $K = 4$. This $128 \times 128 \times 4$ output tensor is then summed over the channel domain, resulting in a localization heatmap. (2) The threshold indicating a patch in the localization heatmap to be positive, i.e. indicating the presence of pneumonia, is increased to 0.995. The latter number was chosen by a validation procedure. (3) We use ResNet-101 as the backbone, which showed slightly better performance on the detection task.

The remaining parts of the architecture, including all hyperparameters, remain unchanged. The network is trained for 50 epochs with batch size of 27. To convert the binary mask into bounding boxes, a simple non-tunable algorithm [45] is applied. Finally, the confidence level for each bounding box is derived by averaging the patch scores within the box by its area.

The dataset we use for evaluation is the Kaggle RSNA Pneumonia Detection Challenge dataset [46], which comprises 26,684 chest X-ray images. Each image may contain a number of annotations to indicate the presence of pneumonia; an image is considered positive if one or more annotations are present. The dataset also includes an independent private test set with 3,000 images and a published private leaderboard. The challenge evaluates the performance using the mean average precision (mAP) score with different intersection over union (IoU) thresholds; the threshold values range from 0.4 to 0.75 with a step size of 0.05. A score of 1 means a good overlap between the derived bounding boxes and the ground truth. A score of 0 can be due to either a false positive or a detection with insufficient overlap, i.e. all detections have bounding boxes with IoU less than 0.4 with a ground truth object. The bounding boxes are evaluated in order of their confidence levels. We use 80% of the public data for training while the remaining 20% is used for validation.

In general, our algorithm was not designed to deal with a dataset like the RSNA Kaggle Challenge for two reasons. First, as we have already emphasized, our algorithm is designed to be a localization algorithm rather than a detection algorithm. That is, it generates blobs with arbitrary shape rather than rectangular bounding boxes. Second, our algorithm focuses on the scenario of weak supervision; the RSNA Kaggle Challenge, by contrast, is not concerned with weak supervision, as all positive images have a bounding box. As a result, our algorithm is not perfectly suited to the RSNA Kaggle Challenge dataset. Nevertheless, we were able to demonstrate good performance on this dataset, and confirmed that the algorithm is indeed able to generate boxes that are spatially well separated, thereby identifying multiple locations of the disease.

Figure 6 show qualitative examples of our detection results. It shows the localization heatmap (prior to thresholding) and the resulting bounding boxes for several example images from the validation set. These images display a wide range of mAP scores, and a varying number of annotations. Observe that although our algorithm is not perfectly suited to this kind of task, it demonstrates good localization performance:


mAP 0.000 mAP 0.000 mAP 0.000 mAP 0.031

mAP 0.062 mAP 0.250 mAP 0.250 mAP 0.500

mAP 0.500 mAP 0.510 mAP 1.000 mAP 1.000

FIGURE 6. Qualitative performance on the Kaggle RSNA Pneumonia Detection Challenge validation set. In every image pair there are: left image - the localization heatmap (prior to thresholding), right image - derived bounding boxes (white) and ground-truth (green) on the example image. Below each image pair is its mean average precision (mAP) score. We show a broad range of mAP scores, as well as a varying number of annotations per image.

for example, the topmost left image localizes the pneumonia quite well despite the fact that the mAP score is 0. Additionally, we see examples where the algorithm obtains relatively high mAP scores for two or three target bounding boxes; this confirms that we can indeed identify multiple discrete locations. The overall mAP for the private dataset is 0.134. While this result is not at the same level as the dedicated ensemble methods winning first place in the challenge (mAP scores in the range of 0.22-0.255), our algorithm nevertheless demonstrates good performance given that it is designed for localization rather than detection; and weakly supervised data rather than fully supervised data.

VIII. CONCLUSION

We have presented a new technique for localization with limited annotation. The training of our network requires very few bounding box annotations, instead relying in large part on much cheaper whole image annotations. As a result, the method is widely applicable to situations in which box-based annotations are expensive, such as medical imaging. The method is based on a novel loss function, which is mathematically and numerically well-posed; and an architecture that explicitly accounts for patch non-independence and shift-invariance. We demonstrate our algorithm's efficacy on the task of thoracic disease localization in chest X-rays. The algorithm is able to localize multiple diseases in a given image, and we demonstrate SOTA results for the localization of 8 different classes in the ChestX-ray14 dataset. Additionally, we show how our algorithm can be extended from the task of localization to that of object detection, with the resulting detector achieving high quality results on the Kaggle RSNE Pneumonia Detection dataset. Future work will focus on applying the ideas presented in this paper to the realm of semantic segmentation.

REFERENCES

[1] R. M. Hopstaken, T. Witbraad, J. M. A. van Engelshoven, and G. J. Dinant, "Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections," *Clin. Radiol.*, vol. 59, no. 8, pp. 743–752, Aug. 2004.

[2] M. I. Neuman, E. Y. Lee, S. Bixby, S. Diperna, J. Hellinger, R. Markowitz, S. Servaes, M. C. Monuteaux, and S. S. Shah, "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children," *J. Hospital Med.*, vol. 7, no. 4, pp. 294–298, Apr. 2012, doi: 10.1002/jhm.955.

[3] H. D. Davies, E. E.-L. Wang, D. Manson, P. Babyn, and B. Shuckett, "Reliability of the chest radiograph in the diagnosis of lower respiratory infections in young children," *Pediatric Infectious Disease J.*, vol. 15, no. 7, pp. 600–604, Jul. 1996.

[4] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[5] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[6] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8290–8299.

[7] E. Rozenberg, D. Freedman, and A. Bronstein, "Localization with limited annotation for chest X-rays," in *Proc. Mach. Learn. Health Workshop (ML4H), Conjunct With NIPS*, 2019, pp. 52–65.

[8] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*. [Online]. Available: http://arxiv.org/abs/1710.10501

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[10] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*. [Online]. Available: http://arxiv.org/abs/1801.09927

[11] S. Gündel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest X-rays with location-aware dense networks," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds. Cham, Switzerland: Springer, 2019, pp. 757–765.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 740–755.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[20] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1081–1089.

[21] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, Apr. 2019.

[22] D. Zhang, J. Han, L. Zhao, and T. Zhao, "From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5549–5560, Dec. 2020.

[23] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[24] B. Babenko, "Multiple instance learning: Algorithms and applications," PubMed, NCBI, Bethesda, MD, USA, 2008, pp. 1–19.

[25] M. Zhang and B. Zeng, "A progressive learning framework based on single-instance annotation for weakly supervised object detection," *Comput. Vis. Image Understand.*, vol. 193, Apr. 2020, Art. no. 102903.

[26] Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 729–745.

[27] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.

[28] M. Zhang, W. Shi, S. Chen, Z. Zhan, and Z. Shi, "Deep multiple instance learning for landslide mapping," *IEEE Geosci. Remote Sens. Lett.*, early access, Jul. 16, 2020, doi: 10.1109/LGRS.2020.3007183.

[29] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1332–1343, May 2016.

[30] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 603–611.

[31] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2424–2433.

[32] E. Schwab, A. Gooßen, H. Deshpande, and A. Saalbach, "Localization of critical findings in chest X-ray without local annotations using multi-instance learning," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1879–1882.

[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2011, pp. 109–117.

[34] P. Kraehenbuehl and V. Koltun, "Parameter learning and convergent inference for dense random fields," in *Proc. 30th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 28, S. Dasgupta and D. McAllester, Eds., Atlanta, GA, USA, Jun. 2013, pp. 513–521. [Online]. Available: http://proceedings.mlr.press/v28/kraehenbuehl13.html

[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062

[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[37] B. J. Bhatkalkar, D. R. Reddy, S. Prabhu, and S. V. Bhandary, "Improving the performance of convolutional neural network for the segmentation of optic disc in fundus images using attention gates and conditional random fields," *IEEE Access*, vol. 8, pp. 29299–29310, 2020.

[38] M. T. T. Teichmann and R. Cipolla, "Convolutional CRFs for semantic segmentation," 2018, *arXiv:1805.04777*. [Online]. Available: http://arxiv.org/abs/1805.04777

[39] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, "Pixel-adaptive convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11166–11175.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 630–645.

[41] R. Zhang, "Making convolutional networks shift-invariant again," 2019, *arXiv:1904.11486*. [Online]. Available: http://arxiv.org/abs/1904.11486

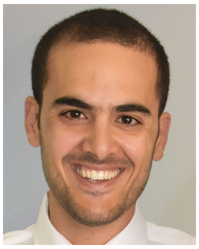
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[44] Y. Henon. (2018). *Pytorch Implementation of RetinaNet Object Detection*. [Online]. Available: https://github.com/yhenon/pytorch-retinanet

[45] C. Voglis. (2018). *From Masks to Bounding Boxes*. [Online]. Available: https://www.kaggle.com/voglinio/from-masks-to-bounding-boxes

[46] (2018). *RSNA Pneumonia Detection Challenge*. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview



EYAL ROZENBERG received the B.Sc. and M.Sc. degrees from the Faculty of Electrical Engineering, Tel Aviv University, in 2016 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Visual Sensing Theory and Applications (VISTA) Laboratory, Department of Computer Science, Technion–Israel Institute of Technology, under the supervision of Prof. Alex Bronstein. His master's thesis which dealt with nonlinear optics was performed under the supervision of Prof. Ady Arie. From 2016 to 2018, he worked as a Signal Processing Algorithm Developer at a start-up company in the autonomous vehicles sector, focusing on the task of lane localization. He has worked with the Chip Design Industry, Mellanox, which was later acquired by Nvidia, from 2014 to 2016, and with a small start-up company acquired by Cadence, from 2013 to 2014. His research interests include the intersection of deep learning and the field of physics, in particular quantum mechanics, nonlinear partial differential equations, and inverse problems.

DANIEL FREEDMAN (Member, IEEE) received the A.B. degree in physics from Princeton University, in 1993, and the Ph.D. degree in engineering sciences from Harvard University, in 2000. From 2000 to 2009, he served as a Professor of computer science with the Rensselaer Polytechnic Institute (RPI), Troy, NY, USA. In 2007, he became a Fulbright Fellow and a Visiting Professor with the Weizmann Institute of Science. Since 2009, he has been holding a number of corporate research positions at HP Labs, IBM Research, Microsoft Research, and finally Google Research. He is currently a Research Scientist at Google Research. He also works in the fields of computer vision, novel imaging modalities, and AI for medical applications. In addition to the Fulbright Fellowship, he received the National Science Foundation CAREER Award.

ALEX A. BRONSTEIN (Fellow, IEEE) is currently a Professor of computer science with the Technion–Israel Institute of Technology and a Principal Engineer with Intel Corporation. His research interests include numerical geometry, computer vision, and machine learning. He has authored over 100 publications in leading journals and conferences, over 30 patents and patent applications, the research monograph *Numerical Geometry of Non-Rigid Shapes*, and edited several books. Highlights of his research were featured in CNN, SIAM News, and Wired. In addition to his academic activity, he co-founded and served as the Vice President of technology in the Silicon Valley start-up company Novafora, from 2005 to 2009. He was a Co-Founder and one of the main inventors and developers of the 3D sensing technology with the Israeli startup Invision, subsequently acquired by Intel, in 2012. His technology is currently the core of the Intel RealSense 3D camera integrated into a variety of consumer electronic products. He is also a Co-Founder of the Israeli video search startup Videocites and the London-based startup Sibylla, where he serves as the Chief Scientist. He is a Fellow of the IEEE for his contribution to 3D imaging and geometry processing.