

# Self-Guided Multiple Instance Learning for Weakly Supervised Disease Classification and Localization in Chest Radiographs

Constantin Seibold<sup>1,3</sup>, Jens Kleesiek<sup>2</sup>,  
Heinz-Peter Schlemmer<sup>2</sup> and Rainer Stiefelhagen<sup>1</sup>

<sup>1</sup> Karlsruhe Institute of Technology  
{constantin.seibold, rainer.stiefelhagen}@kit.edu

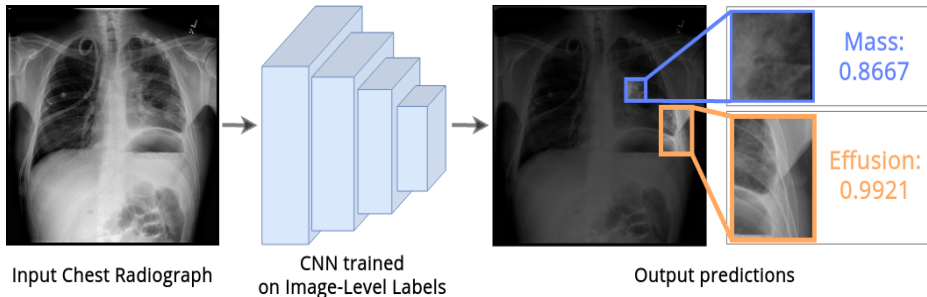
<sup>2</sup>German Cancer Research Center Heidelberg  
{j.kleesiek,h.schlemmer}@dkfz-heidelberg.de

<sup>3</sup>HIDSS4Health - Helmholtz Information and Data Science School for Health,  
Karlsruhe/Heidelberg, Germany

**Abstract.** The lack of fine-grained annotations hinders the deployment of automated diagnosis systems, which require human-interpretable justification for their decision process. In this paper, we address the problem of weakly supervised identification and localization of abnormalities in chest radiographs. To that end, we introduce a novel loss function for training convolutional neural networks increasing the *localization confidence* and assisting the overall *disease identification*. The loss leverages both image- and patch-level predictions to generate auxiliary supervision. Rather than forming strictly binary from the predictions as done in previous loss formulations, we create targets in a more customized manner, which allows the loss to account for possible misclassification. We show that the supervision provided within the proposed learning scheme leads to better performance and more precise predictions on prevalent datasets for multiple-instance learning as well as on the NIH ChestX-Ray14 benchmark for disease recognition than previously used losses.

## 1 Introduction

With millions of annually captured images, chest radiographs (*CXR*) are one of the most common tools assisting radiologists in the diagnosing process [1]. The emergence of sizeable CXR datasets such as Open-I or ChestX-ray14 [2–6], allowed deep Convolutional Neural Networks (*CNN*) to aid the analysis for the detection of pulmonary anomalies [6–24]. Despite the success of deep learning, inferring the correct abnormality location from the network’s decision remains challenging. While for supervised tasks, this is achieved through algorithms such as Faster R-CNN [25–27], the necessary amount of fine-grained annotation for CXR images to train these models is vastly missing and expensive to obtain. Instead, models are trained using image-level labels parsed from medical reports, which might be inaccurate [5]. As such, the problem of pulmonary pathology identification and localization is at best weakly supervised.



**Fig. 1.** In our framework, the network reads chest X-ray images and produces overall image-level pathology prediction scores and their corresponding locations.

Existing work for weakly-supervised pathology localization in CXR builds either upon network saliency or Multiple-Instance Learning (*MIL*). Saliency-based methods [6–13] focus primarily on the multi-class recognition task and predict locations implicitly through saliency visualization methods such as CAM, Grad-CAM, or excitation backpropagation [28–30]. These methods employ global average pooling to merge spatial features during the classification process. However, through this process the CNN makes less indicative decisions, as healthy regions are heavily outweighing the few regions of interest containing the abnormality. The other direction combines Fully Convolutional Networks (*FCN*) with MIL to implicitly learn patch-level predictions used for localization [20–23]. In MIL-based methods, the input data is regarded as a bag of instances where the label is only available on bag-level. The bag will be assigned a positive label if and only if there exists at least one positive instance. This problem formulation fits for diagnosis in medical images as small regions might define the existence of a pathology within the overall image.

In this paper, we focus on MIL-based approaches to diagnose and localize pulmonary abnormalities in CXRs. Much MIL-related work investigated the use of different pooling functions resembling a max-function to aggregate either predictions or embeddings [6, 31–37]. By balancing all given outputs, networks learn implicitly from the bag label. We argue that this approach overlooks the explicit use of instance-level predictions into training. We present a novel loss formulation split into two stages. While the first stage leads through conventional bag-level classification, the second stage leads to more definitive predictions by generating auxiliary supervision from instance-level predictions. By segregating the prediction maps into foreground, background, and ambiguous regions, the network can provide itself instance-wise targets with differing levels of certainty.

The main contributions of our study can be summarized as follows: We provide a novel loss function that applies prediction maps for self-guidance to achieve better classification and localization performance without the necessity to expand a given fully convolutional network architecture. We present the effect of this loss on MIL-specific datasets as well as the ChestX-Ray14 benchmark. The experiments demonstrate competitive results with the current state-of-the-art for weakly supervised pathology classification and localization in CXRs.

## 2 Related Work

**Automated Chest Radiograph Diagnosis.** With the release of large-scale CXR datasets [2–6] the development of deep learning-based automated diagnosis methods made noticeable progress in both abnormality identification [6–24, 38–40] and the subsequent step of report generation [17, 19, 41]. However, despite CNNs, at times, surpassing the accuracy of radiologists in detecting pulmonary diseases [11, 14], inferring the correct pathology location remains a challenge due to the lack of concretely annotated data. Initial work such as done by Wang *et al.* [6] or Rajpurkar *et al.* [11] uses CAM [28] to obtain pathology locations. Due to the effectiveness and ease of use, saliency-based methods like CAM became a go-to method for showcasing predicted disease regions [6, 11–14, 37]. As such, there exists work to improve CAM visualizations through the use of auxiliary modules or iterative training [42–44].

Alternatively, Li *et al.* [21] propose a slightly modified FCN trained in MIL-fashion to address the problem. Here, each image patch is assigned a likelihood of belonging to a specific pathology. These likelihood-scores are aggregated using a noisy-OR pooling for the means of computing the loss. This approach is extended by Yao *et al.* [39] and Liu *et al.* [22] who while using different architecture or preprocessing methods stick with the same MIL-based training regime. Similarly, Rozenberg *et al.* [23] expand Li *et al.*’s approach through the usage of further postprocessing steps such as the integration of CRFs.

All of these methods approach this task through image-level supervision and try to gain improved localization through changes in architecture, iterative training or postprocessing. In contrast, rather than modifying a given architecture, we leverage network predictions within the same training step to achieve more confident localization.

**Multiple Instance Learning.** MIL has become a widely adopted category within weakly supervised learning. It was first proposed for drug activity prediction [45] and has since found a use for applications such as sound [34, 35, 46] and video event tagging [33] as well as weakly supervised object detection [47–49]. While max- and mean- pooling have been common choices for deep MIL networks, recent work investigates the use of the pooling function to combine instance embeddings or predictions to deliver a bag-level prediction [32–36, 39]. The choice of pooling function will often resemble the max-operator or an approximation of such to stay in line with common MIL-assumptions. Static functions such as Noisy-OR, Log-Sum-Exp or Softmax [6, 21, 34] along with learnable ones like adaptive generalized mean, auto-pooling or attention have been proposed [32–34]. While the choice of pooling function is a vital part of the overall inference and loss computation in training step in MIL, it in itself does not provide sufficient information as the optimization will still occur only based on the bag-level prediction. In order to accurately impact the training, instance-level predictions are necessary to influence the loss. There exist few methods that leverage the use of artificial supervision within a MIL setting to train the network additionally through instance-level losses [33, 50–52]. One direction is to

introduce artificial instance-labels for prediction scores above a specified threshold [33,51,52]. The loss function splits into a bag-level loss acting in standard MIL fashion by aggregating the predictions and an instance-level prediction where the network gains pixel-wise supervision based on a set prediction threshold [33]. While this approach provides supervision for each instance, it is heavily depending on the initialization potentially introducing a negative bias. On the other side, Morfi *et al.* [50] introduce the MMM loss for audio event detection. This loss provides direct supervision for the extreme values of the bag, whereas the overall bag accumulated using a mean pooling for a bag-level prediction. Despite all instances influencing the optimization, the supervision of this method is limited as it disregards the association for probable positive/negative instances.

In an ideal scenario, each positive instance should have a near maximal prediction whereas negative ones should be minimal. However, often, the case presents where the amount of positive bags will sway a classifier towards a biased prediction due to class imbalance. This might lead to all instances within a bag for a certain class to get either high or low prediction scores making strict thresholding difficult to apply. Furthermore, as long as the prediction value distribution within a bag is not separable but rather clumped or uniform existing methods cannot account for a fitting expansion of the decision boundary. In contrast, we adopt instance-level supervision in an adapting way, where the distinctness of the prediction directly defines the influence of the loss.

### 3 Methodology

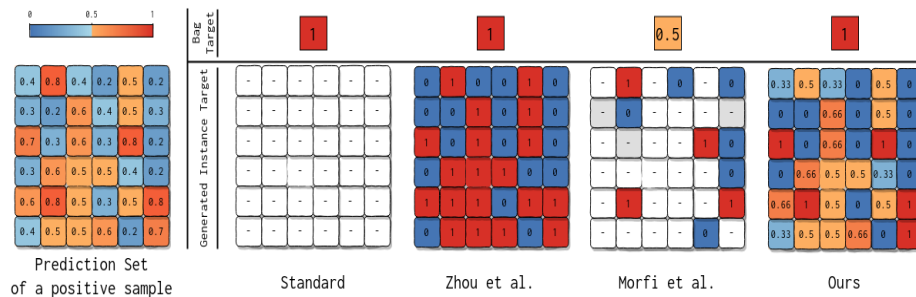
We start this section by defining multiple-instance learning. We, then, introduce our proposed Self-Guiding Loss (*SGL*) and how it differs from existing losses. Lastly, we address the use of *SGL* for classification and weakly supervised localization of CXR pathologies in a MIL setting.

#### 3.1 Preliminaries of Multiple-Instance Learning

Assume, we are given a set of bag-of-instances of size  $N$  with the associated labels  $\mathcal{B} = \{(B_1, y_1), \dots, (B_N, y_N)\}$ . Let  $B_i, i = \{1, \dots, N\}$  be the  $i$ -th bag-of-instances and  $B_{i,j} \in B_i, j \in \{1, \dots, N_i\}$  be the  $j$ -th instance with  $N_i$  being the number of instances of the  $i$ -th bag. The associated labels  $y_i \in \{0, 1\}^C$  describe the presence or absence of classes, which can occur independently of each other. Let  $c \in \{1, \dots, C\}$  describe a certain class out of  $C$  classes in total. The label of a bag and an instance for a specific class  $c$  is thus shown by  $y_i^c \in \{0, 1\}$  and  $y_{i,j}^c \in \{0, 1\}$ , respectively. We refer to a target of 1 as positive and 0 as negative. The MIL-assumption requires that  $y_i^c = 1$  if and only if there exists at least a single positive instance, hence we can define

$$y_i^c = \max_j y_{i,j}^c. \tag{1}$$

Note that while the bag-level annotation  $y_i^c$  is available within the training data, the instance-level annotation  $y_{i,j}^c$  is unknown.



**Fig. 2.** Illustration of supervision for different loss function concepts for MIL. Strict bag-level supervision (left) provided, Zhou *et al.*’s BIL [33] (center left), Morfi *et al.*’s MMM [50] (center right) and on the right our proposed SGL.

We aim at learning a classifier to predict the likelihood of each instance in regard to each class within a bag-of-instances. In several works for deep MIL, this classifier might consist of a convolutional backbone  $\Psi$  linked with a pooling layer  $\Phi$  to combine predictions or features. The class-wise likelihood of a single instance is denoted by  $p_{i,j}^c(B_{i,j}) \in [0, 1]$  with

$$\mathbf{p}_i^c(B_i) = \{p_{i,1}^c(B_{i,1}), p_{i,2}^c(B_{i,2}), \dots, p_{i,N_i}^c(B_{i,N_i})\} = \Psi_c(B_i) \quad (2)$$

being the set of all instance-level predictions for class  $c$  of the  $i$ -th bag. These instance-level predictions are aggregated using a pooling layer to obtain bag-level predictions

$$p_i^c(B_i) = \Phi_c(\mathbf{p}_i^c(B_i)) \quad (3)$$

with  $p_i^c(B_i) \in [0, 1]$ . For brevity, we omit the arguments of the presented functions from this point on.

### 3.2 Self-Guiding Loss

The SGL is designed to address the MIL-setting. Here, one faces an inherent lack of knowledge of the correct instance labels joined with an imbalance between positive and negative instances. Commonly used MIL approaches merge instance predictions and train entirely by optimizing any loss function using the bag’s label  $y$  and the bag prediction  $\mathbf{p}$ . This level of supervision is illustrated on the left in Fig 2. The bag label is presented in the top row, while the types of instance supervision are displayed in the bottom. Numbers designate the target label whereas “-” denotes no existing supervision for that particular instance. While this level of supervision will lead the network to accurate bag-level predictions, inferring the determining instance is not ensured.

Rather than just utilizing the bag, our loss formulation is split in two parts. The first part defines the bag-level loss, while the second part describes how the network’s predictions induce artificial supervision to train the network.

**Bag-Level Loss.** The bag-level loss behaves as in classic MIL approaches. A bag-level prediction is generated by aggregating the network’s instance-level predictions. We calculate the loss of this stage using common loss functions  $\mathcal{L}$  such as the binary cross-entropy by passing the prediction and target for all classes and bags as follows:

$$\mathcal{L}_{Bag}(\mathcal{B}, y) = \frac{1}{C \cdot N} \sum_c \sum_i \mathcal{L}(p_i^c, y_i) \quad (4)$$

with  $i \in \{1, \dots, N\}$  and  $c \in \{1, \dots, C\}$ . This loss is, hereby, depending on the choice of the pooling function  $\Phi$  and provides leeway for the instance-level loss to step in.

**Instance-Level Loss.** To outline the instance-level loss, we start with the assumption that a network trained just from bag labels will inevitably assign some positive instances a noticeably higher prediction score than most negative instances. From this, we derive three types of instance predictions. Instances with a high score are likely to be considered positive, whereas instances with a low score as negative. Instances with scores close to the decision boundary are rather ambiguous as they may easily be swayed in the course of training and as such do not pose an as concrete implication about the actual class of the instance. Pursuing this line of thought we establish three types of supervision based on the certainty level of each prediction within a bag.

Our first step is to normalize the prediction set using the common min-max feature scaling. We apply this to avoid cases of biases stemming from either algorithmic decisions such the choice of the pooling function or general data imbalance. We denote the resulting rescaled bag of predictions  $\theta$  by

$$\theta_{ij}^c = \frac{p_{ij}^c - \min(\mathbf{p}_i^c)}{\max(\mathbf{p}_i^c) - \min(\mathbf{p}_i^c)} \quad (5)$$

with min and max being functions returning the minimal and maximal values within a set respectively. The normalized predictions are then used within a ternary mask  $M$  depicting targets stemming from the previously named cases similar to Hou *et al.* [53] and Zhang *et al.* [42]. For this, we define a higher and lower threshold to partition the prediction set,  $\delta_h$  and  $\delta_l$  respectively with  $\delta_h + \delta_l = 1$  and  $\delta_h \geq \delta_l \geq 0$ . Everything larger than the upper threshold  $\delta_h$  will be regarded as a positive instance and all instances with scores lower than  $\delta_l$  as negative. The target mask  $M$  is then defined for each instance  $j$  in the bag  $i$  for class  $c$  by

$$M_{i,j}^c = \begin{cases} 0 & , \text{ if } \theta_{i,j}^c < \delta_l \text{ or } y_i^c = 0 \\ \theta_{i,j}^c & , \text{ if } \delta_l \leq \theta_{i,j}^c \leq \delta_h \\ 1 & , \text{ if } \delta_h < \theta_{i,j}^c \end{cases} \quad (6)$$

For distinctly positive and negative predictions, we obtain instance-wise supervision with a target value of 1 and 0 respectively. We can also presume based on

Eq. 1. that each instance within negative bags is also negative. Thus, we can set all values of their masks to 0. The remaining uncertain regions, however, do not allow for an explicit label assignment. While we want to enforce the networks decision process, we also have to account for possible missassignment. Thus, rather than setting a fixed target value, we set the target to be  $\theta$ . This process shows some similarity to the popular label smoothing procedure [54]. Rather than using maximal valued targets, the maps adjusted value is inserted into the loss function as target value. This slightly pushes the loss into the direction of the most extreme predictions within the uncertain instance set. By doing so we steadily increase the amount distinctly positive and negative predictions over the course of training.

We can construct the loss using a fundamental loss function  $\mathcal{L}$  like binary cross entropy by utilizing  $M$  as target. The instance-level loss is then defined as

$$\mathcal{L}_{Inst}(\mathcal{B}, M) = \sum_i \sum_c \sum_j 2^{\alpha_i^c - 1} \cdot \mathcal{L}(p_{i,j}^c, M_{i,j}^c), \quad (7)$$

where each part is being normalized by the number of pixels with the respective supervision types. This way, we strengthen the networks decision process for its more certain instances. We, further, consider a weighing factor  $\alpha$  to influence the bag's impact based on the overall certainty of its prediction. We define  $\alpha$  by

$$\alpha_i^c = \max(\max(\mathbf{p}_i^c) - \text{median}(\mathbf{p}_i^c), 1 - y_i) \quad (8)$$

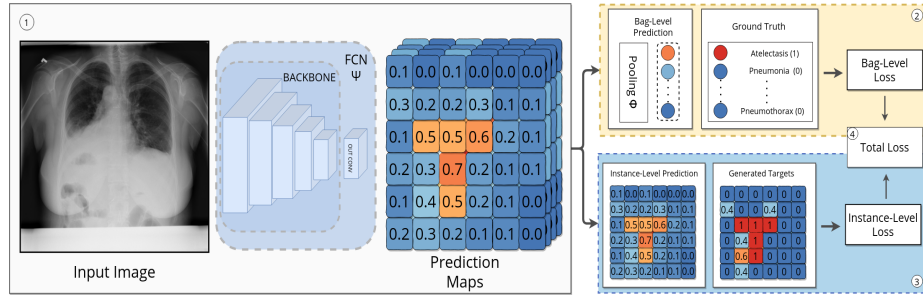
Since a positive bag in a common MIL setting should have a low valued median due to a limited amount of positive instances, it is weighted highly if the network is able to clearly separate positive from negative predictions. Thus, for positive bags,  $\alpha = 0$  if all predictions result in the same value and  $\alpha = 1$  if the network is able to clearly separate positive from negative instances under the assumption that the number of positive instances is vastly smaller than the number of negative ones. For negative bags,  $\alpha = 1$  holds due to the given supervision.

The complete loss is then defined by

$$\mathcal{L}_{SGL}(p_i, y_i) = \mathcal{L}_{Bag} + \lambda \cdot \mathcal{L}_{Inst}, \quad (9)$$

with  $\lambda$  denoting the weighing hyperparameter of the instance-level loss.

An example of the final supervision for our loss is displayed in Fig 2. The standard approach on the left uses no instance-level supervision. In the center left, Zhou *et al.*'s BIL provides a positive label for each instance above the 0.5 threshold and a negative else, while maintaining the bag supervision. The MMM loss by Morfi *et al.*, in the center right, considers positive labels for the maximum instances and negative ones for minimal instances. It further uses the target of 0.5 for a mean pooled prediction. Opposed to this, our loss adapts its assumed supervision to the produced predictions. Rather than just using the maximum or applying set thresholding, we threshold on a rescaled set of predictions, thus avoiding a common problem occurring with imbalanced data. Our formulation incorporates all instance predictions while providing a margin of error based on the networks certainty over the smoothed targets  $\theta$  and the weighing factor  $\alpha$ .



**Fig. 3.** Overview of considered framework for thoracic disease identification and localization. A chest X-ray is passed through an FCN and produces a prediction map. The prediction map is used to compute the instance- and bag-level losses.

### 3.3 MIL for Chest Radiograph Diagnosis

We consider a MIL scenario for CXR diagnosis. We build upon the assumption that singular patches (instances) of an image (bag-of-instances) can infer the occurrence of such a pathology (class). An example of this is the class “nodules”, which can take up minimal space within the image. We are given just image-level labels for pathologies, while more detailed information such as bounding box or pixel-level supervision remains hidden. The bag is associated with a class if and only if there exists at least one instance causing such implication. The goal is to learn a model that when given a bag of instances can predict the bag’s label based on the instance information. By classifying the bag’s instances the model provides insight regarding which regions are affected by a pathology.

**Overview.** In Figure 3, we illustrate an overview of the considered scheme for CXR diagnosis. Firstly, an FCN processes CXR-images, which results in patch-wise classification scores for each abnormality. The number of patches stems from their perceptual field, which is a result of backbone architecture. Each patch is independently processed via a  $1 \times 1$  convolutional classification layer. In this work, we do not add specific modules to our backbone network. These patch predictions are aggregated in the second part using a pooling layer, which produces a bag-level prediction for our bag-level loss. The third part applies an instance-level loss function based on the patch-wise predictions. In the fourth part, both the instance and bag-level losses join for optimization. Here, we further penalize the occurrence of non-zero elements in  $M$  using an  $L_2$ -Norm.

**Choice of Pooling Function.** The choice of the correct pooling function is vital for any MIL-setting to produce accurate bag-level predictions. Methods like max and mean pooling will lead to imprecise decisions. In the context of MIL in CXR diagnosis, Noisy-OR found use, but this function suffers from the numerical instability stemming from the product of a multitude of instances. Rather than letting singular instances influence the decision process, we choose to employ the Softmax-pooling, which has found success in audio event detection [34,35]. It provides a meaningful balance between instance-level predictions to let each instance influence the bag level loss based on its intensity.



## 4 Experiments

### 4.1 Datasets

**MNIST-Bags.** In a similar fashion to Ilse *et al.* [32], we use the MNIST-bags [32,55] dataset to evaluate our method for a MIL-setting. A bag is created grayscale MNIST-images of size  $28 \times 28$ , which are resized to  $32 \times 32$ . A bag is considered positive if it contains the label “9”. The number of images in a bag is Gaussian-distributed based on a fixed bag size. We investigate different average bag sizes and amounts of training bags. During evaluation 1000 bags created from the MNIST test set of the same bag size as used in training. We average the results of ten training procedures.<sup>1</sup>

**CIFAR10-Bags.** We build CIFAR10-bags from CIFAR10 [56] in a similar fashion to MNIST-bags. We choose to create 2500 and 5000 training and test bags respectively with fixed bag sizes. A bag here is considered positive if it contains the label “dog”. We investigate in these experiments the influence of a varying number of positive instances per bag. We average five training runs.

**NIH ChestX-ray14.** To present the effect of our loss for medical diagnosis, we conduct experiments on the NIH ChestX-ray14 dataset [6]. It contains 112,120 frontal-view chest X-rays taken from 30,805 patients with 14 disease labels. Unless further specified, we resize the original image size of  $1024 \times 1024$  to  $512 \times 512$ . We use the official split between train/val and test, as such we get a 70%/10%/20% split. Also, 880 images with a total of 984 images with bounding boxes for 8 of the 14 pathologies from the test set.

### 4.2 Implementation Details

For all MNIST-Bags-experiments, we use a LeNet5 model [55] as Ilse *et al.* [32]. We apply max-pooling,  $\delta_l = 0.3$  and  $\lambda = 1$  for our method unless further specified. We train BIL [33] using mean-pooling as we found it unable to train with max-pooling.

For all CIFAR10-bags-experiments, we train a ResNet-18 [57] with the same optimizer hyperparameters and batchsize of 64 for 50 epochs. We apply max-pooling,  $\delta_l = 0.3$  and  $\lambda = 1$  for our method.

For the experiments on NIH ChestX-ray14, each network is initialized using an Image-Net pretraining. We use the same base model as Wang *et al.* [6] by employing a ResNet-50 [57]. We replace the final fully connected and pooling layers with a convolutional layer of kernel size  $1 \times 1$ , resulting in the same number of parameters as Wang *et al.* We follow standard image normalization [58]. For training, we randomly crop the images to size 7/8-th of the input image size, whereas we use the full image size during test time. We train the network for 20 epochs using the maximum batch-size for our GPU using Adam [59] with a learning rate, weight decay,  $\beta_1$  and  $\beta_2$  of  $10^{-4}$ ,  $10^{-4}$ , 0.9 and 0.999 respectively. We decay the learning rate by 0.1 every 10 epochs. We set  $\delta = 0.3$  and  $\lambda =$

<sup>1</sup> [www.github.com/ConstantinSeibold/SGL](http://www.github.com/ConstantinSeibold/SGL)

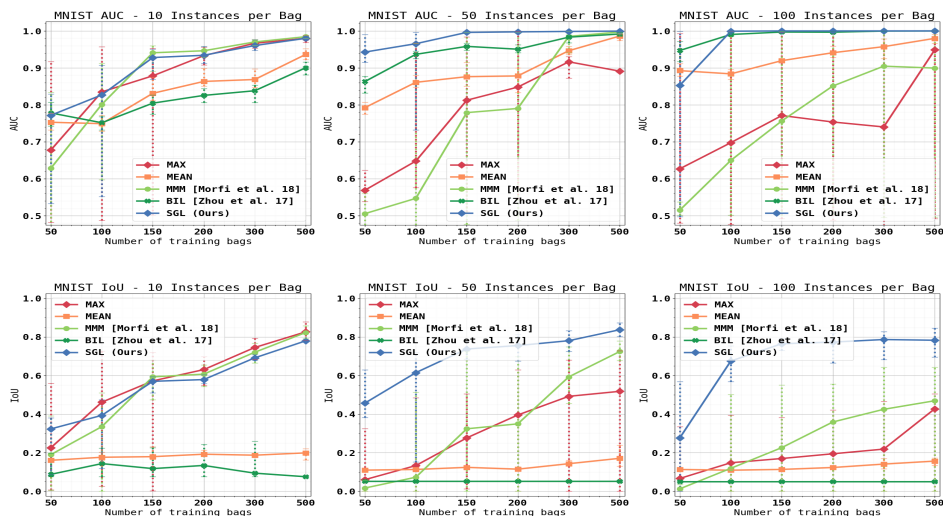


Fig. 4. Test AUC and IoU for MNIST-Bags for differing avg. instances per bag.

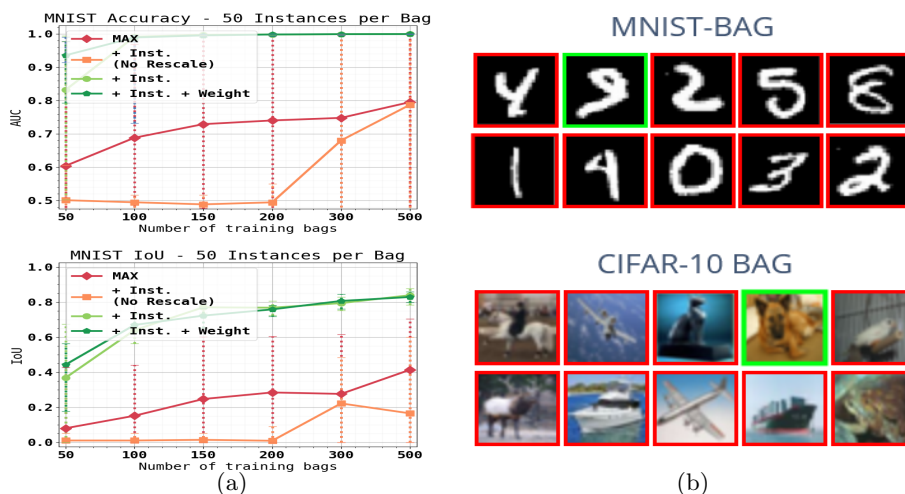
20. We increase  $\lambda$  to keep the two losses on similar magnitudes. The model is implemented using Pytorch [60].

### 4.3 Evaluation Metrics

We evaluate the classification ability of our network via the area under the ROC-curve ( $AUC$ ). To evaluate the localization ability we apply average intersection-over-union ( $IoU$ ) to calculate the class-wise localization accuracy similar to Rusakovsky *et al.* [58]. To compute the localization scores, we threshold the probability map at the scalar value  $T_p$  to get the predicted area and compute the intersection between predicted and ground truth area to compute the IoU. In the case of MNIST- and CIFAR10-bags IoU is computed as the intersection between predicted positive instances and ground truth positive instances at  $T_p = 0.5$ . The localization accuracy is calculated by  $\frac{\#hit}{\#hit + \#miss}$ , where an image has the correct predicted localization ( $hit$ ) iff it has the correct class prediction and a higher overlap than a predefined threshold  $T_{IoU}$ .

### 4.4 Results

**MNIST-Bags.** The AUC and IoU results for the mean bag sizes of 10, 50, and 100 with a varying number of given training bags are displayed from left to right in the top and bottom row of Fig. 4. We present the average of the runs as well as the best and worst runs for each method. For small bags, our method performs similarly to the simple max-pooling in both AUC and IoU. We attribute this average performance to the small number of instances in a bag, which does not allow to make proper use of our ternary training approach. As we increase the bag size to 50 and 100 our proposed loss performs better than

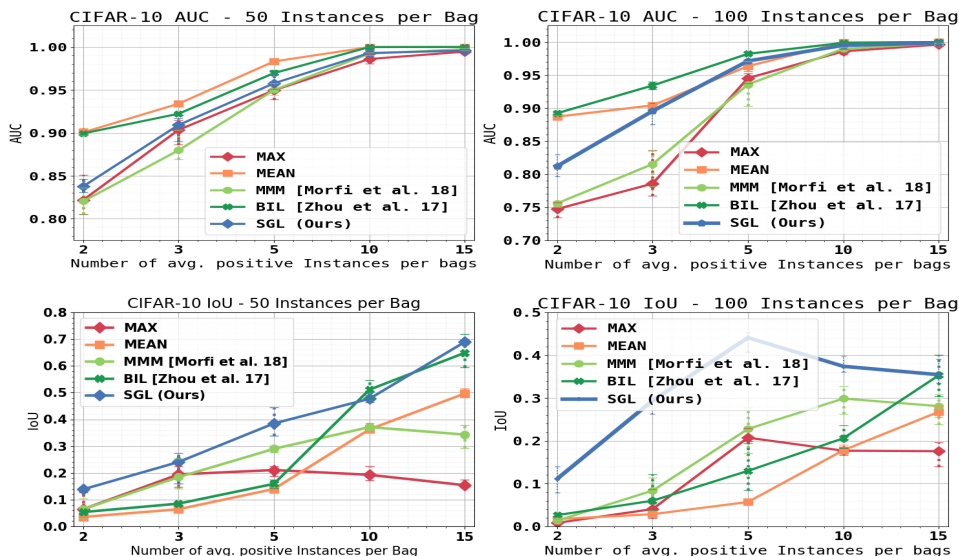


**Fig. 5.** Ablation study of SGL in Figure (a). Figure (b) displays exemplary positive bags of size 10 for MNIST-bags (top) and CIFAR10-bags (bottom). Positive instances denoted in green, negative ones in red.

the max-pooling baseline but also than the other methods for both metrics. We can see the difference notably in the IoU, where our loss achieves nearly double the performance of the next best method for almost all amounts of training bags. We notice that our approach does not pose a trade-off between confident predictions and overall AUC but manages to facilitate a training environment which improves both metrics. It is also worth mentioning that while increasing the amount of training bags improves the method for any bag size our loss achieves exceptional performance for both AUC and IoU with a relatively small number training examples for larger bag sizes. We can reason that the further use of self-guidance can potentially improve a method regardless of dataset size.

In Fig. 5 (a), we present ablation studies involving different constellations of the loss. When considering the loss components, we start with just max-pooling baseline and successively add parts of SGL. Max-pooling alone struggles with the identification of positive/negative bags, however, improves slowly in terms of IoU and AUC with increasing numbers of training bags. When adding the proposed loss without the rescaling mentioned in Eq. 3.2 and weighting component (shown by *Inst. (No Rescale)*) the method becomes incapable to learn as even random initializations might skew the network towards incorrect conclusions. When adding the rescaling component (shown by *Inst.*) the model drastically outperforms prior parts in both metrics. Doing so achieves higher maximums than with the applied weighting factor  $\alpha$  displayed by *Inst. + Weight*. However, the addition of the weighting factor provides a more stable training, specifically for smaller amounts of training data.

**CIFAR10-Bags.** The AUC and IoU results for the mean bag sizes of 50 and 100 with a varying number of positive instances per bag are in the top and bottom row of Fig. 6. For smaller bag sizes, we observe that straight forward mean-pooling achieves the best AUC scores for CIFAR10-Bags. Overall SGL im-



**Fig. 6.** Test AUC and IoU for CIFAR10-Bags for differing number of average positive instances per bag with a bag size of 50 and 100.

proves over straight forward max-pooling for any number of instances. In regards to IoU, our method manages to outperform other methods for nearly any number of positive instances per bag. For larger bag sizes SGL achieves roughly the same performance as BIL which trained using mean-pooling in terms of AUC while outperforming it in IoU for all amounts of positive instances per bag. The addition of self-guidance manages to bridge shortcomings of max-pooling, boosting its classification accuracy for any bag size or number of positive instances.

**NIH ChestX-Ray 14: Multi-Label Pathology Classification.** Table 1 shows the AUC scores for all the disease classes. We compare the results of our loss function with a common classification approach by Wang *et al.* [6], the MIL-based methods proposed by Li *et al.* [21] and Liu *et al.* [22]. The latter two employ noteworthy architectural adaptations and train using bounding box

Pathologies Method	At.	Card.	Cons.	Ed.	Eff.	Emph.	Fib.	Hernia	Inf.	Mass	Nod.	Pl. Th.	Pn.	Pt.	Mean
Wang <i>et al.</i>	0.70	0.81	0.70	0.81	0.76	0.83	0.79	0.87	0.66	0.69	0.67	0.68	0.66	0.80	0.75
Li <i>et al.</i> *	<b>0.80</b>	0.87	<b>0.80</b>	0.88	0.87	0.91	0.78	0.70	0.70	0.83	0.75	0.79	0.67	0.87	0.81
Liu <i>et al.</i> *	0.79	0.87	0.79	<b>0.91</b>	<b>0.88</b>	0.93	0.80	0.92	0.69	0.81	0.73	0.80	<b>0.75</b>	0.89	<b>0.83</b>
ResNet-50+SGL	0.78	<b>0.88</b>	0.75	0.86	0.84	<b>0.95</b>	<b>0.85</b>	<b>0.94</b>	<b>0.71</b>	<b>0.84</b>	<b>0.81</b>	<b>0.81</b>	0.74	<b>0.90</b>	<b>0.83</b>

**Table 1.** Comparison of classification performance for CXR pathologies on the NIH ChestX-Ray14 dataset. Here, 70% of all images were used for training with no bounding box annotations available. Evaluations were performed on the official test split containing 20% of all images. “\*” denotes usage of additional bounding box supervision.

$T_{IoU}$	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.1	Wang <i>et al.</i> [6]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.57
	Li <i>et al.</i> [21]*	<b>0.71</b>	<b>0.98</b>	<b>0.87</b>	<b>0.92</b>	<b>0.71</b>	0.40	0.60	<b>0.63</b>	<b>0.73</b>
	Liu <i>et al.</i> [22]	0.39	0.90	0.65	0.85	0.69	0.38	0.30	0.39	0.60
	SGL (Ours)	0.67	0.94	0.67	0.81	<b>0.71</b>	<b>0.41</b>	<b>0.66</b>	0.43	0.66
0.3	Wang <i>et al.</i> [6]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	Li <i>et al.</i> [21]*	<b>0.36</b>	<b>0.94</b>	<b>0.56</b>	<b>0.66</b>	0.45	<b>0.17</b>	<b>0.39</b>	<b>0.44</b>	<b>0.50</b>
	Liu <i>et al.</i> [22]	0.34	0.71	0.39	0.65	<b>0.48</b>	0.09	0.16	0.20	0.38
	SGL (Ours)	0.31	0.76	0.30	0.43	0.34	0.13	<b>0.39</b>	0.18	0.36
0.5	Wang <i>et al.</i> [6]	0.05	0.18	0.11	0.07	0.01	0.01	0.01	0.03	0.06
	Li <i>et al.</i> [21]*	0.14	<b>0.84</b>	<b>0.22</b>	0.30	0.22	0.07	<b>0.17</b>	<b>0.19</b>	<b>0.27</b>
	Liu <i>et al.</i> [22]	<b>0.19</b>	0.53	0.19	<b>0.47</b>	<b>0.33</b>	0.03	0.08	0.11	0.24
	SGL (Ours)	0.07	0.32	0.08	0.19	0.18	<b>0.10</b>	0.12	0.04	0.13
0.7	Wang <i>et al.</i> [6]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	Li <i>et al.</i> [21]*	0.04	<b>0.52</b>	0.07	0.09	0.11	<b>0.01</b>	<b>0.05</b>	0.05	0.12
	Liu <i>et al.</i> [22]	<b>0.08</b>	0.30	<b>0.09</b>	<b>0.25</b>	<b>0.19</b>	<b>0.01</b>	0.04	<b>0.07</b>	<b>0.13</b>
	SGL (Ours)	0.02	0.01	0.1	0.00	0.04	0.00	0.03	0.01	0.01

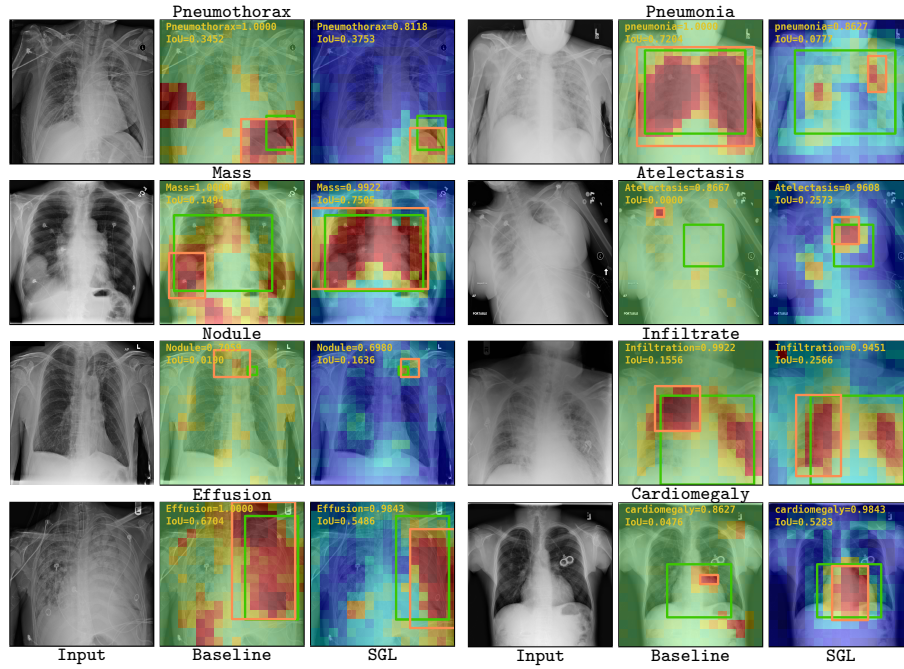
**Table 2.** Disease localization accuracy are evaluated with a classification threshold of 0.5. “\*” denotes additional bounding box supervision.

supervision. All of the named methods utilize a ResNet-50 as backbone network. We outperform the baseline ResNet-50 of Wang *et al.* in all categories. We observe that our loss formulation achieves better classification performance than all other methods in **9** of **14** classes in total. We also reach a better mean performance than other methods, which use further bounding box annotations and architectural modifications such employing additional networks [22] or further convolutional layer [21, 22].

**NIH ChestX-Ray 14: Pathology Localization.** We evaluate the localization ability of the prior named methods through the accuracy over an IoU threshold. For our method, we upsample each prediction map using Nearest-Neighbor-Interpolation. We construct bounding boxes around the connected component of the maximum prediction after applying common morphological operations. The results are displayed in Table 2 for the IoU thresholds  $T_{IoU} \in \{0.1, 0.3, 0.5, 0.7\}$ . We, further, display qualitative examples for each pathology in Figure 7. For the visualization, we use no morphological postprocessing. We compare our method against a baseline version of our model trained only using the bag-level loss with a mean-pooling function. The expert annotation is displayed by a green bounding box, while the predicted one is orange.

Our method achieves favourable performance across all pathologies on a threshold of  $T_{IoU} = 0.1$ . It generally outperforms the baseline of Wang *et al.* [6]. For higher thresholds, our model falls behind the more specified approaches of Li *et al.* [21] and Liu *et al.* [22]. We ascribe the suboptimal quantitative performance to the factors of low spatial output resolution, which can hinder passing the IoU threshold especially for naturally small classes such as *Nodules*, and the overall coarse annotation as can be seen in Figure 7 e.g. the pathology *Infiltrate*. Here, an infiltrate affects the lung area, which the model correctly marks, yet the bounding box naturally includes the cardiac area, thus diminishing the IoU.

In Figure 7, we see that our proposed method can generally make more precise predictions compared to the baseline model. Furthermore, the model can more



**Fig. 7.** We compare the patch-wise predictions between a mean-pooling trained baseline to our proposed method for different diseases. The value ranges from 0 (blue) to 1 (red). We show prediction boxes (orange) around the connected component of the maximum prediction and Ground-Truth bounding boxes (green).

distinctly separate between healthy and abnormal tissue. These results indicate the ability of our loss to lead itself towards more refined predictions.

## 5 Conclusion

In this paper, we propose a novel loss formulation in which one gathers auxiliary supervision from the network’s predictions to provide instance-level supervision. In comparison to existing MIL-based loss functions, we do not rely on initialization and still provide pixel-wise supervision driving the network. Due to the design of this loss, it can support any MIL-setting such as patch-based pathology diagnosis. We demonstrate our method on two MIL-based datasets as well as the challenging NIH ChestX-Ray14 dataset. We display promising classification and localization performance qualitatively and quantitatively.

**Acknowledgements.** The present contribution is supported by the Helmholtz Association under the joint research school “HIDSS4Health Helmholtz Information and Data Science School for Health”.

## References

1. : Nhs england: Diagnostic imaging dataset statistical release. <https://www.england.nhs.uk/> (2020)
2. Kohli MD, R.M.: Open-i: Indiana university chest x-ray collection. <https://openi.nlm.nih.gov> (2013)
3. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. arXiv preprint arXiv:1901.07441 (2019)
4. Johnson, A.E., Pollard, T.J., Berkowitz, S., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR: a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 **1** (2019)
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 590–597
6. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2097–2106
7. Cai, J., Lu, L., Harrison, A.P., Shi, X., Chen, P., Yang, L.: Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2018) 589–598
8. Shen, Y., Gao, M.: Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In: International Workshop on Machine Learning in Medical Imaging, Springer (2018) 389–397
9. Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M.: Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In: International Workshop on Machine Learning in Medical Imaging, Springer (2018) 249–258
10. Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports **9** (2019) 1–10
11. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: CheXnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
12. Wang, H., Xia, Y.: Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. arXiv preprint arXiv:1807.03058 (2018)
13. Park, B., Cho, Y., Lee, G., Lee, S.M., Cho, Y.H., Lee, E.S., Lee, K.H., Seo, J.B., Kim, N.: A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-pa x-ray screening for pulmonary abnormalities. Scientific reports **9** (2019) 1–9
14. Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., et al.: Deep learning for chest radiograph diagnosis: A retrospective comparison of the cheXnet algorithm to practicing radiologists. PLoS medicine **15** (2018) e1002686
15. Wang, Q., Cheng, J.Z., Zhou, Y., Zhuang, H., Li, C., Chen, B., Liu, Z., Huang, J., Wang, C., Zhou, X.: Low-shot multi-label incremental learning for thoracic

- diseases diagnosis. In: International Conference on Neural Information Processing, Springer (2018) 420–432
16. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 6666–6673
  17. Li, X., Cao, R., Zhu, D.: Vispi: Automatic visual perception and interpretation of chest x-rays. arXiv preprint arXiv:1906.05190 (2019)
  18. Li, Y., Pang, Y., Shen, J., Cao, J., Shao, L.: Netnet: Neighbor erasing and transferring network for better single shot object detection. arXiv preprint arXiv:2001.06690 (2020)
  19. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 9049–9058
  20. Yan, C., Yao, J., Li, R., Xu, Z., Huang, J.: Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. (2018) 103–110
  21. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L.: Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8290–8299
  22. Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., Yu, Y.: Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10632–10641
  23. Rozenberg, E., Freedman, D., Bronstein, A.: Localization with limited annotation for chest x-rays. (2019)
  24. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters (2018)
  25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
  26. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, Springer (2016) 21–37
  27. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
  28. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2921–2929
  29. Selvaraju, R.R., 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. (2017) 618–626
  30. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126** (2018) 1084–1102
  31. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)



32. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712 (2018)
33. Zhou, Y., Sun, X., Liu, D., Zha, Z., Zeng, W.: Adaptive pooling in multi-instance learning for web video annotation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2017) 318–327
34. McFee, B., Salamon, J., Bello, J.P.: Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **26** (2018) 2180–2193
35. Wang, Y., Li, J., Metze, F.: A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2019) 31–35
36. Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky-or network. *IEEE transactions on neural networks and learning systems* **30** (2019) 3484–3495
37. Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., Huang, J.: Deep multi-instance learning with dynamic pooling. In: Asian Conference on Machine Learning. (2018) 662–677
38. Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P.: Deep hierarchical multi-label classification of chest x-ray images. *Proceedings of Machine Learning Research* **1** (2019) 13
39. Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K.: Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv preprint arXiv:1803.07703 (2018)
40. Guendel, S., Ghesu, F.C., Grbic, S., Gibson, E., Georgescu, B., Maier, A., Comaniciu, D.: Multi-task learning for chest x-ray abnormality classification on noisy labels. arXiv preprint arXiv:1905.06362 (2019)
41. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. arXiv preprint arXiv:1904.02633 (2019)
42. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 597–613
43. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1325–1334
44. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1568–1576
45. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* **89** (1997) 31–71
46. Kong, Q., Xu, Y., Sobieraj, I., Wang, W., Plumbley, M.D.: Sound event detection and time-frequency segmentation from weakly labelled data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27** (2019) 777–787
47. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2843–2851
48. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence* **39** (2016) 189–203

49. Felipe Zeni, L., Jung, C.R.: Distilling knowledge from refinement in multiple instance detection networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2020) 768–769
50. Morfi, V., Stowell, D.: Data-efficient weakly supervised learning for low-resource audio event detection using deep learning. arXiv preprint arXiv:1807.06972 (2018)
51. Wang, X., Zhu, Z., Yao, C., Bai, X.: Relaxed multiple-instance svm with application to object discovery. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1224–1232
52. Shamsolmoali, P., Zareapoor, M., Zhou, H., Yang, J.: Amil: Adversarial multi-instance learning for human pose estimation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16** (2020) 1–23
53. Hou, Q., Jiang, P., Wei, Y., Cheng, M.M.: Self-erasing network for integral object attention. In: Advances in Neural Information Processing Systems. (2018) 549–559
54. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2818–2826
55. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86** (1998) 2278–2324
56. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. (2009)
57. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
58. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115** (2015) 211–252
59. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
60. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W. (2017)