

# SODA: Detecting Covid-19 in Chest X-rays with Semi-supervised Open Set Domain Adaptation

Jielu Zhou\*  
 zhoujeli777@hotmail.com  
 Carnegie Mellon University  
 PA, USA

Baoyu Jing\*  
 baoyuj2@illinois.edu  
 University of Illinois at  
 Urbana-Champaign  
 IL, USA

Zeya Wang  
 zw17.rice@gmail.com  
 Rice University  
 TX, USA

## ABSTRACT

The global pandemic of COVID-19 has infected millions of people since its first outbreak in last December, and the daily new cases are still climbing by hundreds of thousands as of May 2020. A key challenge for preventing and controlling COVID-19 is how to quickly, widely, and effectively implement the test for the disease, because testing is the first step to break the chains of transmission. To assist and speed up the diagnosis of the disease, radiology imaging is used to complement the screening process and triage patients into different risk levels. Deep learning methods have been considered as very powerful tools and have taken a more active role in automatically detecting COVID-19 disease in chest x-ray images, as witnessed in many recent works in the past few weeks. Most of these works first train a Convolutional Neural Network (CNN) on an existing large-scale chest x-ray image dataset and then fine-tune it with a COVID-19 dataset at a much smaller scale. However, direct transfer across datasets from different domains may lead to poor performance due to domain shift, especially on the biomedical datasets which can be collected and preprocessed quite differently from different hospitals. Also, the small scale of the COVID-19 dataset on the target domain can make the training fall into the overfitting trap. To solve all these crucial problems and fully exploit the available large-scale chest x-ray image dataset[32], we formulate the problem of COVID-19 chest x-ray image classification in a semi-supervised open set domain adaptation setting, through which we are motivated to reduce the domain shift and avoid overfitting when training on a very small dataset of COVID-19. In addressing this formulated problem, we propose a novel Semi-supervised Open set Domain Adversarial network (SODA), which is able to align the data distributions across different domains in a general domain space and also in a common subspace of source and target data. In our experiments, SODA achieves a leading classification performance compared with recent state-of-the-art models, as well as effectively separating COVID-19 with common pneumonia.

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*Conference'17, July 2017, Washington, DC, USA*

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

## KEYWORDS

COVID-19, Medical Image Analysis, Domain Adaptation, Open Set Domain Adaptation, Semi-Supervised Learning

### ACM Reference Format:

Jielu Zhou, Baoyu Jing, and Zeya Wang. 2020. SODA: Detecting Covid-19 in Chest X-rays with Semi-supervised Open Set Domain Adaptation. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/1122445.1122456>

## 1 INTRODUCTION

Since the Coronavirus disease 2019 (COVID-19) was first declared as a Public Emergency of International Concern (PHEIC) on January 30, 2020<sup>1</sup>, it has quickly evolved from a local outbreak in Wuhan, China to a global pandemic, costing millions of lives and dire economic loss worldwide. In the US, the total COVID-19 cases grew from just one confirmed on Jan 21, 2020 to over 1 million on April 28, 2020 in a span of 3 months. Despite drastic actions like shelter-in-place and contact tracing, the total cases in US kept increasing at an alarming daily rate of 20,000 - 30,000 throughout April, 2020. A key challenge for preventing and controlling COVID-19 right now is the ability to quickly, widely and effectively test for the disease, since testing is usually the first in a series of actions to break the chains of transmission and curb the spread of the disease.

COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)<sup>2</sup> By far, it is the most reliably diagnosed through Reverse Transcription Polymerase Chain Reaction (RT-PCR)<sup>3</sup> in which a sample is taken from the back of throat or nose of the patients and tested for viral RNA. While taking samples from the patients, aerosol pathogens could be released and would put the healthcare workers at risk. Furthermore, once the sample is collected, the testing process usually takes several hours and recent study reports that the sensitivity of PT-PCR is around 60-70% [1], which suggests that many people tested negative for the virus may actually carry it thus could infect more people without knowing it. On the other hand, the sensitivity of chest radiology imaging for COVID-19 was much higher at 97% as reported by [1, 8].

Due to the shortage of viral testing kits, the long period of waiting for results, and low sensitivity rate of RT-PCR, radiology imaging has been used as a complementary screening process to assist the diagnosis of COVID-19 and triage patients into different risk levels.

<sup>1</sup><https://www.statnews.com/2020/01/30/who-declares-coronavirus-outbreak-a-global-health-emergency/>

<sup>2</sup>[https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)

<sup>3</sup><https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/how-do-coronavirus-tests-work>

Unlike PT-PCR, imaging is readily available in most healthcare facilities around the world, and the whole process can be done rapidly. In recent years, with the rapid advancement in deep learning and computer vision, many breakthroughs have been developed in using Artificial Intelligence (AI) for medical imaging analysis, especially disease detection [14, 32, 33] and report generation [4, 16, 17, 21], and some AI models achieve expert radiologist-level performance [19]. Right now, with most healthcare workers busy at front lines saving lives, the scalability advantage of AI-based medical imaging systems stand out more than ever. Some AI-based chest imaging systems have already been deployed in hospitals to quickly inform healthcare workers to take corresponding actions<sup>4</sup>.

Annotated datasets are required for training AI-based methods, and a small chest x-ray dataset with COVID-19 is collected recently: COVID-ChestXray [6]. In the last few weeks, several works [2, 20, 22, 23] apply Convolutional Neural Networks (CNN) and transfer learning to detect COVID-19 cases from chest x-ray images. They first train a CNN on a large dataset like CheXpert [14] and ChestXray14 [32], and then fine-tune the model on the small COVID-19 dataset. By far, due to the lack of large-scale open COVID-19 chest x-ray imaging datasets, most works only used a very small amount of positive COVID-19 imaging samples [6]. While the reported metrics like accuracy and AUC are high, it is likely that these models overfit on this small dataset and may not achieve the reported performance on a larger COVID-19 x-ray dataset. Besides, these methods suffer a lot from label domain shift: these newly trained models lose the ability to detect common thoracic diseases like “Effusion” and “Nodule” since these labels do not appear in the new dataset. Moreover, they also ignored the visual domain shift between the two datasets. On the one hand, the large-scale datasets like ChestXray14 [32] and CheXpert [14] are collected from top U.S. health institutes like National Institutes of Health (NIH) clinical center and Stanford University, which are well-annotated and carefully processed. On the other hand, COVID-ChestXray [6] is collected from a very diverse set of hospitals around the world and they are of very different qualities and follow different standards, such as the viewpoints, aspect ratios and lighting, etc. In addition, COVID-ChestXray contains not only chest x-ray images but also CT scan images.

In order to fully exploit the limited but valuable annotated COVID-19 chest x-ray images and the large-scale chest x-ray image dataset at hand, as well as prevent the above-mentioned drawbacks of those fine-tuning based methods, we define the problem of learning a classifier for COVID-19 from the perspective of open set domain adaptation (Definition 1) [25]. Different from traditional unsupervised domain adaptation which requires the label set of both source and target domain to be the same, the open set domain adaptation allows different domains to have different label sets. This is more suitable for our problem because COVID-19 is a new disease which is not included in the ChestXray14 or CheXpert dataset. However, since our task is to train a new classifier for COVID-19 dataset, we have to use some annotated samples. Therefore, we further propose to view the problem as a Semi-Supervised Open Set Domain Adaptation problem (Definition 2).

<sup>4</sup><https://spectrum.ieee.org/the-human-os/biomedical/imaging/hospitals-deploy-ai-tools-detect-covid19-chest-scans>

Under the given problem setting, we propose a novel Semi-supervised Open set Domain Adversarial network (SODA) comprised of four major components: a feature extractor  $G_f$ , a multi-label classifier  $G_y$ , domain discriminators  $D_g$  and  $D_c$ , as well as common label recognizer  $R$ . SODA learns the domain-invariant features by a two-level alignment, namely, domain level and common label level. The general domain discriminator  $D_g$  is responsible for guiding the feature extractor  $G_f$  to extract domain-invariant features. However, it has been argued that the general domain discriminator  $D_g$  might lead to false alignment and even negative transfer [26, 34]. For example, it is possible that the feature extractor  $G_f$  maps images with “Pneumonia” in the target domain and images with “Cardiomegaly” in the source domain into similar positions, which might result in the miss-classification of  $G_y$ . In order to solve this problem, we propose a novel common label discriminator  $D_c$  to guide the model to align images with common labels across domains. For labeled images,  $D_c$  only activates when the input image is associated with a common label. For unlabeled images, we propose a common label recognizer  $R$  to predict their probabilities of having a common label.

The main contributions of the paper are summarized as follows:

- To the best of our knowledge, we are the first to tackle the problem of COVID-19 chest x-ray image classification from the perspective of domain adaptation.
- We formulate the problem in a novel semi-supervised open set domain adaptation setting.
- We propose a novel two-level alignment model: Semi-supervised Open set Domain Adversarial network (SODA).
- We present a comprehensive evaluation to demonstrate the effectiveness of the proposed SODA.

## 2 PRELIMINARY

### 2.1 Problem Definition

#### Definition 1. *Unsupervised Open Set Domain Adaptation*

Let  $\mathcal{D}^s = \{(\mathbf{x}_n^s, y_n^s)\}_{n=1}^{N^s}$  be a source domain with  $N^s$  labeled samples, and  $\mathcal{D}^t = \{(\mathbf{x}_n^t)\}_{n=1}^{N^t}$  be a target domain with  $N^t$  unlabeled samples, where the underlying label set  $\mathcal{L}^t$  of the target domain might be different from the label set  $\mathcal{L}^s$  of the source domain. Define  $\mathcal{L}^c = \mathcal{L}^s \cap \mathcal{L}^t$  as the *set of common labels* shared across different domains,  $\bar{\mathcal{L}}^s = \mathcal{L}^s \setminus \mathcal{L}^c$  and  $\bar{\mathcal{L}}^t = \mathcal{L}^t \setminus \mathcal{L}^c$  be *sets of domain-specific labels* which only appear in the source and the target domain respectively.

The task of *Unsupervised Open Set Domain Adaptation* is to build a model which could accurately assign common labels in  $\mathcal{L}^c$  to samples  $\mathbf{x}_n^t$  in the target domain, as well as distinguish those  $\mathbf{x}_n^t$  belonging to  $\bar{\mathcal{L}}^t$ .

#### Definition 2. *Semi-supervised Universal Domain Adaptation*

Given a source domain  $\mathcal{D}^s = \{(\mathbf{x}_n^s, y_n^s)\}_{n=1}^{N^s}$  with  $N^s$  labeled samples, and a target domain  $\mathcal{D}^t \cup \mathcal{D}'$  consisting of  $\mathcal{D}^t = \{(\mathbf{x}_n^t)\}_{n=1}^{N^t}$  with  $N^t$  unlabeled samples and  $\mathcal{D}' = \{(\mathbf{x}_n', y_n')\}_{n=1}^{N'}$  with  $N'$  labeled samples.

The task of *Semi-supervised Open Set Domain Adaptation* is to build a model to assign labels from  $\mathcal{L}^t$  to unlabeled samples in  $\mathcal{D}^t$ .

## 2.2 Notations

We summarize the symbols used in the paper and their descriptions in Table 1.

**Table 1: Notations**

Symbols	Description
$\mathcal{D}^s$	set of labeled samples in the source domain
$\mathcal{D}^t$	set of unlabeled samples in the target domain
$\mathcal{D}'^t$	set of labeled samples in the target domain
$\mathcal{L}^s$	set of labels for the source domain
$\mathcal{L}^t$	set of labels for the target domain
$\mathcal{L}^c$	set of common labels across domains
$\hat{\mathcal{L}}^s$	set of domain-specific labels in the source domain
$\hat{\mathcal{L}}^t$	set of domain-specific labels in the target domain
$\mathcal{L}$	set of all labels from all domains
$N^s$	number of labeled samples in the source domain
$N^t$	number of unlabeled samples in the target domain
$N'^t$	number of labeled samples in the target domain
$G_f$	feature extractor
$G_y$	multi-label classifier for $\mathcal{L}$
$G_{y_l}$	binary classifier for label $l$ (part of $G_y$ )
$R$	common label recognizer
$D_c$	domain discriminator for common labels $\mathcal{L}^c$
$D_g$	general domain discriminator
$L_{G_y}$	loss of multi-label classification over the entire dataset
$L_R$	loss of $R$ over the entire dataset
$L_{D_g}$	loss of $D_g$ over the entire dataset
$L_{D_c}$	loss of $D_c$ over the entire dataset
$\lambda$	the coefficient of losses
$\mathbf{x}$	input image
$\mathbf{h}$	hidden features
$y$	ground-truth label
$\hat{y}$	predicted probability
$\hat{d}$	predicted probability that $\mathbf{x}$ belongs to source domain
$\hat{r}$	predicted probability that $\mathbf{x}$ has common labels

## 3 METHODOLOGY

### 3.1 Overview

An overview of the proposed Semi-supervised Open Set Domain Adversarial network (SODA) is shown in Fig. 1. Given an input image  $\mathbf{x}$ , it will be first fed into a feature extractor  $G_f$ , which is a Convolutional Neural Network (CNN), to obtain its hidden feature  $\mathbf{h}$  (green part). The binary classifier  $G_{y_l}$  (part of the multi-label classifier  $G_y$ ) takes  $\mathbf{h}$  as input, and will predict the probability  $\hat{y}_l$  for the label  $l \in \mathcal{L}$  (blue part).

We propose a novel two-level alignment strategy for extracting the domain invariant features across the source and target domain. On the one hand, we perform *domain alignment* (Section 3.2), which leverages a general domain discriminator  $D_g$  to minimize the domain-level feature discrepancy. On the other hand, we emphasize the *alignment of common labels*  $\mathcal{L}^c$  (Section 3.3) by introducing another domain discriminator  $D_c$  for images associated

with common labels. For labeled images in  $\mathcal{D}^s$  and  $\mathcal{D}'^t$ , we compute loss for  $D_c$  and conduct back-propagation only if the input image  $\mathbf{x}$  is associated with a common label  $l \in \mathcal{L}^c$ . As for unlabeled data in  $\mathcal{D}^t$ , we propose a common label recognizer  $R$  to predict the probability  $\hat{r}$  that an image  $\mathbf{x}$  has a common label, and use  $\hat{r}$  as a weight in the losses of  $D_c$  and  $D_g$ .

### 3.2 Domain Alignment

Domain adversarial training [10] is the most popular method for helping feature extractor  $G_f$  learn domain-invariant features such that the model trained on the source domain can be easily applied to the target domain. The objective function of the domain discriminator  $D_g$  can be written as:

$$\begin{aligned} L_{D_g} = & -\mathbb{E}_{(\mathbf{x}^s \in \mathcal{D}^s)}[\log \hat{d}_g] \\ & -\mathbb{E}_{(\mathbf{x}^t \in \mathcal{D}^t \cup \mathcal{D}'^t)}[\log(1 - \hat{d}_g)] \end{aligned} \quad (1)$$

where  $\hat{d}_g$  denotes the predicted probability that the input image belongs to the source domain.

In SODA, we use a Multi-Layer Perceptron (MLP) as the general domain discriminator  $D_g$ .

### 3.3 Common Label Alignment

In the field of adversarial domain adaptation, most of the existing methods only leverage a general domain discriminator  $D_g$  to minimize the discrepancy between the source and target domain. Such a practice ignores the label structure across domains, which will result in false alignment and even negative transfer [26, 34]. If we only use a general domain discriminator  $D_g$  in the open set domain adaptation setting (Definition 1 and Definition 2), it is possible that the feature extractor  $G_f$  will map the target domain images with a common label  $l \in \mathcal{L}^c$ , say “Pneumonia”, and the source domain images with a specific label  $l \in \hat{\mathcal{L}}^s$ , “Cardiomegaly”, to similar positions in the hidden space, which might lead to the classifier miss-classifying a “Pneumonia” image in the target domain as “Cardiomegaly”.

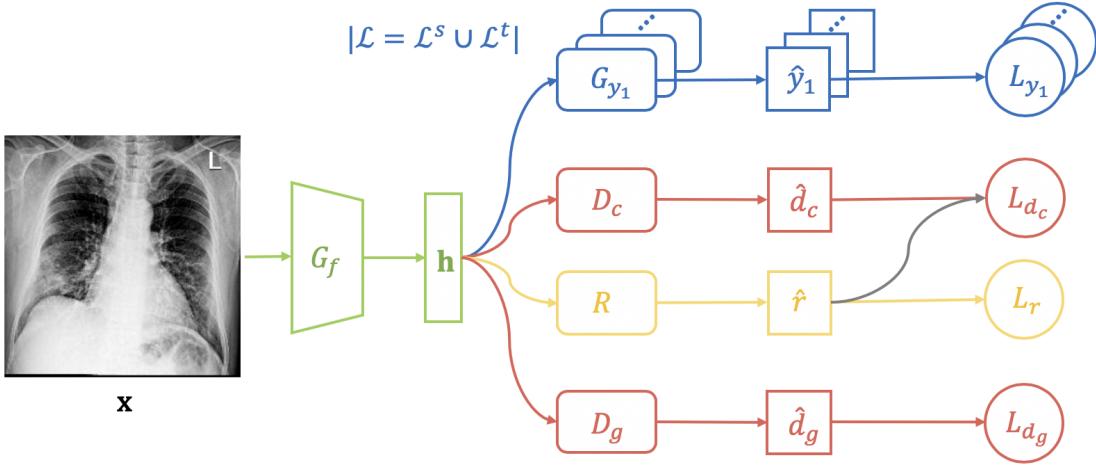
To address the problem of the miss-matching between the common and specific label sets, we propose a domain discriminator  $D_c$  to distinguish the domains for the images with a common label. For the labeled data from the source domain  $\mathcal{D}^s$  and the target domain  $\mathcal{D}^t$ , we know whether an image  $\mathbf{x}$  has a common label or not, and we only calculate the loss  $L_{D_c}$  for  $D_c$  on the samples with common labels:

$$\begin{aligned} L_{D_c}^{label} = & -\mathbb{E}_{(\mathbf{x}^s \in \mathcal{D}^s, y^s \in \mathcal{L}^c)}[\log \hat{d}_c] \\ & -\mathbb{E}_{(\mathbf{x}^t \in \mathcal{D}^t, y^t \in \mathcal{L}^c)}[\log(1 - \hat{d}_c)] \end{aligned} \quad (2)$$

where  $\hat{d}_c$  denotes the predicted probability that the input images is associated with a common label.

However, a large number of images in the target domain are unlabeled, and thus extra effort is required for determining whether an unlabeled image is associated with a common label. To address this problem, we propose a novel common label recognizer  $R$  to predict the probability  $\hat{r}$  whether an unlabeled image has at least one common label. The probability  $\hat{r}$  will be used as a weight in the loss function of  $D_c$ <sup>5</sup>:

<sup>5</sup>Note that gradients stop at  $\hat{r}$  in the training period.



**Figure 1: Architecture overview of the proposed SODA model.** Given an input image  $x$ , the feature extractor  $G_f$  will extract its hidden features  $h$  (green part), which will be fed into a multi-label classifier  $G_y$  (blue part), a common label recognizer  $R$  (yellow part) and a domain discriminator  $D$  (red part) to predict the probability  $\hat{y}$  of disease labels, the probability  $\hat{r}$  that  $x$  is associated with a common label and the probability  $\hat{d}$  that  $x$  belongs to the source domain.  $L_y$ ,  $L_r$  and  $L_d$  denote the losses of image classification, common label classification and domain classification.  $D_g$  is the general domain discriminator, and  $D_c$  is the domain discriminator for images associated with a common label.  $G_{y_1}$  denotes the image classifier for the first label in the label set of the entire dataset  $\mathcal{L} = \mathcal{L}^s \cup \mathcal{L}^t$ . Note that the gradients from  $L_{d_c}$  and  $L_{d_g}$  are not allowed to pass through  $\hat{r}$  (grey arrows).

$$L_{D_c}^{un} = -\mathbb{E}_{(x' \in \mathcal{D}', y' \in \mathcal{L}^c)}[\hat{r} \log(1 - \hat{d}_c)] \quad (3)$$

In addition, we also use  $\hat{r}$  to re-weigh unlabeled samples in  $D_g$  (Equation 1) to further emphasize the alignment of common labels:

$$\begin{aligned} L_{D_g} &= -\mathbb{E}_{(x^s \in \mathcal{D}^s)}[\log \hat{d}_g] \\ &\quad - \mathbb{E}_{(x' \in \mathcal{D}')}[ \log(1 - \hat{d}_g)] \\ &\quad - \mathbb{E}_{(x' \in \mathcal{D}')}[\hat{r} \log(1 - \hat{d}_g)] \end{aligned} \quad (4)$$

Finally, the recognizer  $R$  is trained on the labeled set  $\mathcal{D}^s \cup \mathcal{L}^t$  via cross-entropy loss:

$$\begin{aligned} L_R &= -\mathbb{E}_{(x \in \mathcal{D}^s \cup \mathcal{D}', y \in \mathcal{L}^c)}[\log \hat{r}] \\ &\quad - \mathbb{E}_{(x \in \mathcal{D}^s \cup \mathcal{D}', y \notin \mathcal{L}^c)}[\log(1 - \hat{r})] \end{aligned} \quad (5)$$

### 3.4 Overall Objective Function

The overall objective function of SODA can be written as a min-max game between classifiers  $G_y$ ,  $R$  and discriminators  $D_g$ ,  $D_c$ :

$$\min_{G_y, R} \max_{D_g, D_c} L_{G_y} + \lambda_R L_R - \lambda_{D_g} L_{D_g} - \lambda_{D_c}^{label} L_{D_c}^{label} - \lambda_{D_c}^{un} L_{D_c}^{un} \quad (6)$$

where  $L_R$ ,  $L_{D_g}$ ,  $L_{D_c}^{label}$  and  $L_{D_c}^{un}$  are respectively defined in Equation 5, 4, 2 and 3;  $L_{G_y}$  denotes the cross-entropy loss for multi-label classification;  $\lambda$  denotes the coefficient for each loss function.

## 4 EXPERIMENTS

### 4.1 Experiment Setup

#### 4.1.1 Datasets.

*Source Domain.* We use ChestXray-14 [32] as the source domain dataset. This dataset is comprised of 112,120 anonymized chest x-ray images from the National Institutes of Health (NIH) clinical center. The dataset contains 14 common thoracic disease labels: “Atelectasis”, “Consolidation”, “Infiltration”, “Pneumothorax”, “Edema”, “Emphysema”, “Fibrosis”, “Effusion”, “Pneumonia”, “Pleural thickening”, “Cardiomegaly”, “Nodule”, “Mass” and “Hernia”.

*Target Domain.* The newly collected COVID-ChestXray [6] is adopted as the target domain dataset. This dataset contains images collected from various public sources and different hospitals around the world. This dataset (by the time of this writing) contains 328 chest x-ray images in which 253 are labeled positive as the new disease “COVID-19”, whereas 61 are labeled as other well-studied “Pneumonia”.

*4.1.2 Evaluation Metrics.* We evaluate our model from four different perspectives. First, to test the classification performance, following the semi-supervised protocol, we randomly split the 328 x-ray images in COVID-ChestXray into 40% labeled set, and 60% unlabeled set. We run each model 3 times and report the average AUC-ROC score. Second, we compute the Proxy- $\mathcal{A}$  Distance (PAD) [3] to evaluate models’ ability for minimizing the feature discrepancy across domains. Thirdly, we use t-SNE to visualize the feature distributions of the target domain. Finally, we also qualitatively evaluate the models by visualizing their saliency maps.

**4.1.3 Baseline Methods.** We compare the proposed SODA with two types of baselines methods: fine-tuning based transfer learning models and domain adaptation models. For fine-tuning based models, we select the two most popular CNN models DenseNet121 [13] and ResNet50 [12] as our baselines. These models are first trained on the ChestXray-14 dataset and then fine-tuned on the COVID-ChestXray dataset. As for the domain adaptation models, we compare our model with two classic models, Domain Adversarial Neural Networks (DANN) [10] and Partial Adversarial Domain Adaptation (PADA) [5]. Note that DANN and PADA were designed for unsupervised domain adaptation, and we implement a semi-supervised version of them.

**4.1.4 Implementation Details.** We use DenseNet121 [13], which is pretrained on the ChestXray-14 dataset [32], as the feature extractor  $G_f$  for SODA. The multi-label classifier  $G_y$  is a one layer neural network and its activation is the sigmoid function. We use the same architecture for  $D_g$ ,  $D_c$  and  $R$ : a MLP containing two hidden layers with ReLU [24] activation and an output layer. The hidden dimension for all of the modules:  $G_y$ ,  $D_g$ ,  $D_c$  and  $R$  is 1024. For fair comparison, we use the same setting of  $G_f$ ,  $G_y$  and  $D_g$  for DANN [10] and PADA [5]. All of the models are trained by Adam optimizer [18], and the learning rate is  $10^{-4}$ .

## 4.2 Classification Results

To investigate the effects of domain adaptation and demonstrate the performance improvement of the proposed SODA, we present the average AUC-ROC scores for all models in Table 2. Comparing the results for ResNet50 and DenseNet121, we observe that deeper and more complex models achieve better classification performance. For the effects of domain adaptation, it is obvious that the domain adaptation methods (DANN, PADA, and SODA) outperform those fine-tuning based transfer learning methods (ResNet50 and DenseNet121). Furthermore, the proposed SODA achieves higher AUC scores on both COVID-19 and Pneumonia than DANN and PADA, demonstrating the effectiveness of the proposed two-level alignment.

**Table 2: Target Domain Average AUC-ROC Score**

Model	COVID-19	Pneumonia
ResNet50 [12]	0.8143	0.8342
DenseNet121 [13]	0.8202	0.8414
DANN [10]	0.8785	0.8961
PADA [5]	0.8822	0.9038
SODA	<b>0.9006</b>	<b>0.9082</b>

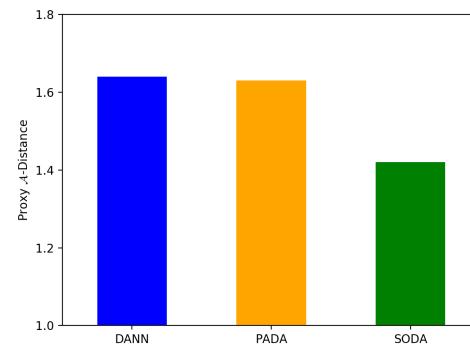
## 4.3 Proxy $\mathcal{A}$ -Distance

Proxy  $\mathcal{A}$ -Distance [3] has been widely used in domain adaptation for measuring the feature distribution discrepancy between the source and target domains. PAD is defined by

$$d_{\mathcal{A}} = 2(1 - 2 \min(\epsilon)) \quad (7)$$

where  $\epsilon$  is the domain classification error (e.g. mean absolute error) of a classifier (e.g. linear SVM [7]).

Following [10], we train SVM models with different  $C$  and use the minimum error to calculate PAD. In general, a lower  $d_{\mathcal{A}}$  means a better ability for extracting domain invariant features. As shown in Fig. 2, SODA has a lower PAD compared with the baseline methods, which indicates the effectiveness of the proposed two-level alignment strategy.



**Figure 2: Proxy  $\mathcal{A}$ -Distance**

## 4.4 Feature Visualization

We use t-SNE to project the high dimensional hidden features  $\mathbf{h}$  extracted by DANN, PADA, and SODA to low dimensional space. The 2-dimensional visualization of the features in the target domain is presented in Fig. 3, where the red data points are image features of “Pneumonia” and the blue data points are image features of “COVID-19”. It can be observed from Fig. 3 that SODA performs the best for separating “COVID-19” from “Pneumonia”, which demonstrates the effectiveness of the proposed common label recognizer  $R$  as well as the domain discriminator for common labels  $D_c$ .

## 4.5 Grad-CAM

Grad-CAM [28] is used to visualize the features extracted from all compared models. Fig. 4 shows the Grad-CAM results on seven different COVID-19 positive chest x-rays. These seven images have annotations (small arrows and box) indicating the pathology locations. We observe that ResNet50 and DenseNet121 can focus wrongly on irrelevant locations like the dark corners and edges. In contrast, domain adaptation models have better localization in general, and our SODA model gives more focused and accurate pathological locations than other models compared. In addition, we consult a professional radiologist with over 15 years of clinical experience from Wuxi People’s Hospital and received positive feedback on the pathological locations as indicated by the Grad-CAM of SODA. In the future, we plan to do a more rigorous evaluation study with more inputs from radiologists. We believe the features extracted from SODA can assist radiologists to pinpoint the suspect COVID-19 pathological locations faster and more accurately.

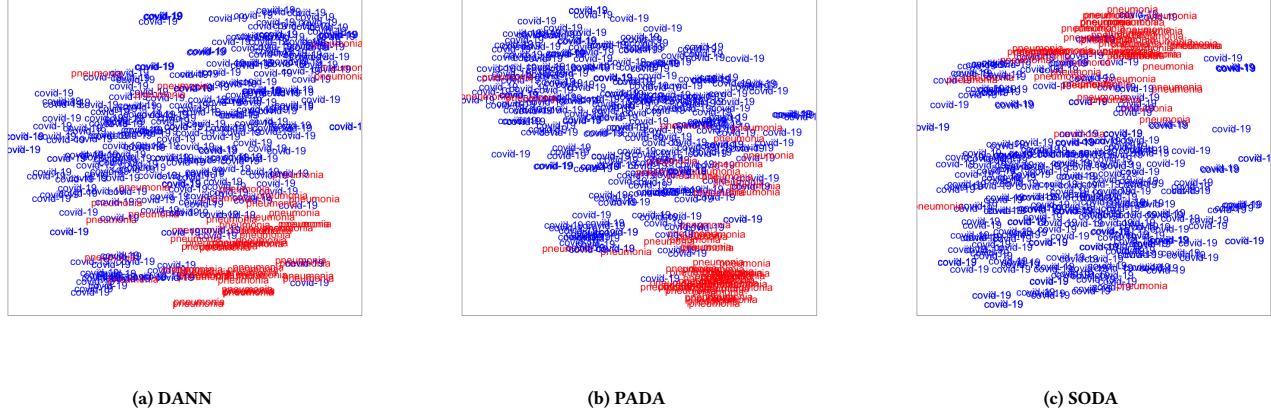


Figure 3: t-SNE visualization for DANN, PADA and SODA on the target domain.

## 5 RELATED WORK

### 5.1 Domain Adaptation

Domain adaptation is an important application of transfer learning that attempts to generalize the models from source domains to unseen target domains [9, 10, 15, 29–31, 36]. Deep domain adaptation approaches are usually implemented through discrepancy minimization [30] or adversarial training [9, 10, 29]. Adversarial training, inspired by the success of generative adversarial modeling [11], has been widely applied for promoting the learning of transfer features in image classification. It takes advantage of a domain discriminator to classify whether an image is from the source or target domains. On top of these methods, a couple of works have been presented for exploring the high-level structure in the label space, which aim at further improving the domain adaptation performance for multi-class image classification [34] or fundamentally solving the application problem when the label sets from source domains and target domains are different [36]. In order to meet the latter target, more and more researchers have started to study the open set domain adaptation problem, in which case the target domain has images that do not come from the classes in the source domain [25, 36]. Universal domain adaptation is the latest method that is proposed through using an adversarial domain discriminator and a non-adversarial domain discriminator to successfully solve this problem. [36]. Although domain adaptation has been well explored, its application in medical imaging analysis, such as domain adaptation for chest x-ray images, is still under-explored.

### 5.2 Semi-supervised Learning

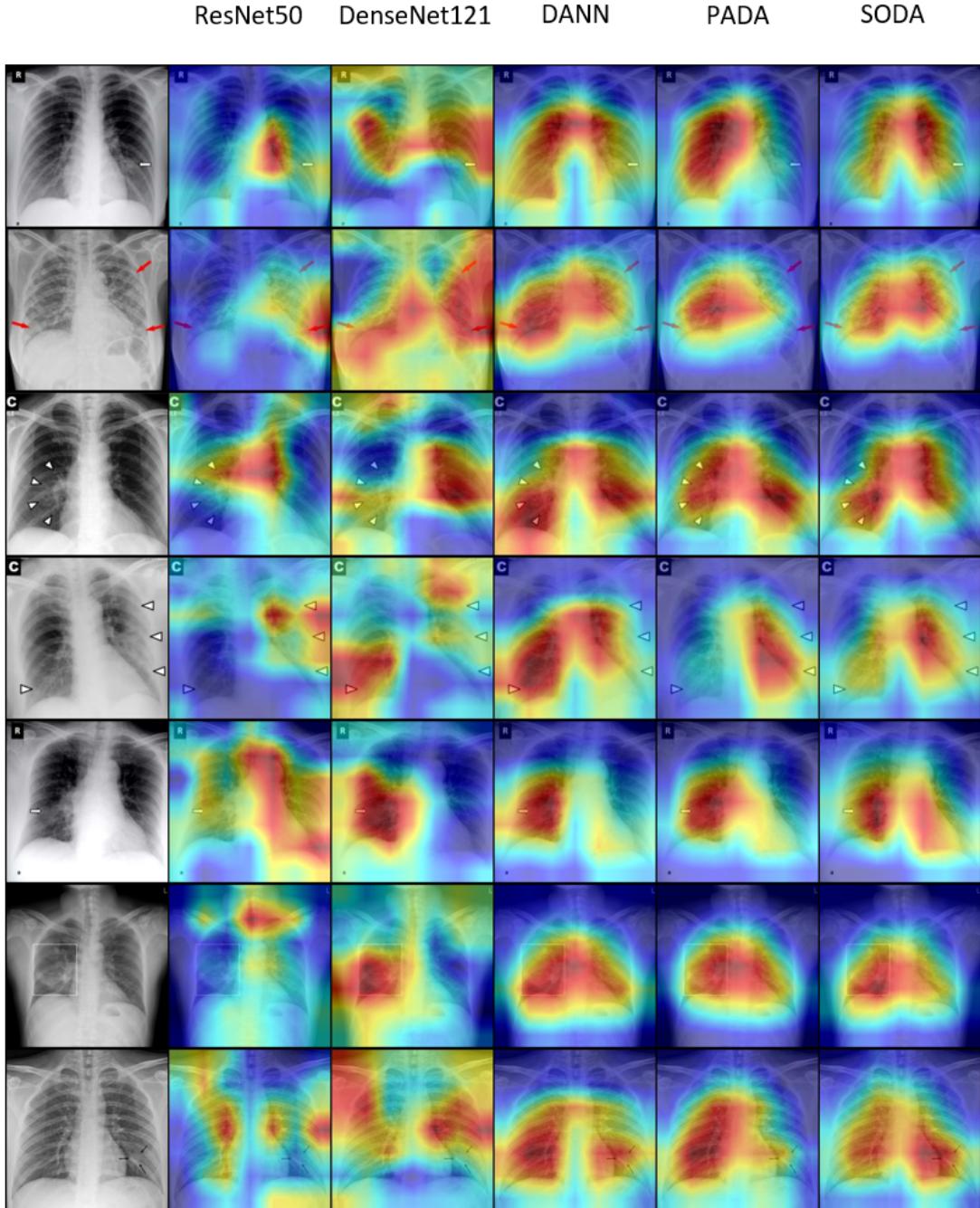
Semi-supervised learning is a very important task for image classification, which can make use of both labeled and unlabeled data at the same time [27]. Recently it has been used to solve image classification problems on a very large (1 billion) set of unlabelled images [35]. In spite of many progresses that have been made with unsupervised domain adaptation methods, the domain adaptation with semi-supervised learning has not yet been fully explored.

### 5.3 Chest X-Ray Image Analysis

There has been substantial progress in constructing publicly available databases for chest x-ray images as well as a related line of works to identify lung diseases using these images. The largest public datasets of chest x-ray images are CheXpert [14] and ChestXray14 [32], which respectively include more than 200,000 and 100,000 chest x-ray images collected by Stanford University and National Institute of Healthcare. The creation of these datasets have also motivated and promoted the multi-label chest x-ray classification for helping the screening and diagnosis of various lung diseases. The problems of disease detection [14, 32, 33] and report generation using chest x-rays [4, 16, 17, 21] are fully investigated and have achieved much-improved results upon recently. However, there have been very few attempts for studying the domain adaptation problems with the multi-label image classification problem using chest x-rays.

## 6 CONCLUSION

In this paper, in order to assist and complement the screening and diagnosing of COVID-19, we formulate the problem of COVID-19 chest x-ray image classification in a semi-supervised open set domain adaptation framework. Accordingly, we propose a novel deep domain adversarial neural network, Semi-supervised Open set Domain Adversarial network (SODA), which is able to align the data distributions across different domains at both domain level and common label level. Through evaluations of the classification accuracy, we show that SODA achieves better AUC-ROC scores than the recent state-of-the-art models. We further demonstrate that the features extracted by SODA is more tightly related to the lung pathology locations, and get initial positive feedback from an experienced radiologist. In practice, SODA can be generalized to any semi-supervised open set domain adaptation settings where there are a large well-annotated dataset and a small newly available dataset. In conclusion, SODA can serve as a pilot study in using techniques and methods from domain adaptation to radiology imaging classification problems.



**Figure 4:** Grad-CAM[28] visualization for ResNet50, DenseNet121, DANN, PADA and SODA. From left to right, the first column is the chest x-ray images from COVID-ChestXray dataset, the second and third columns are visualizations of the weights on last layers from ResNet50 and DenseNet121, the fourth and fifth columns are visualizations of weights from the last layers of the domain adaptation models DANN and PADA. The last column is the visualization of our model. SODA has the best pathology localization among all models.

## REFERENCES

- [1] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. 2020. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology* (2020), 200642.
- [2] Ioannis D Apostolopoulos and Tzani A Mpesiana. 2020. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine* (2020), 1.
- [3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*. 137–144.
- [4] Siddharth Biswal, Cao Xiao, Lucas Glass, Brandon Westover, and Jimeng Sun. 2020. Clinical Report Auto-completion. In *Proceedings of The Web Conference 2020*. 541–550.
- [5] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. 2018. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 135–150.
- [6] Joseph Paul Cohen, Paul Morrison, and Lan Dao. 2020. COVID-19 image data collection. *arXiv 2003.11597* (2020). <https://github.com/ieee8023/covid-chestxray-dataset>
- [7] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
- [8] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. 2020. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* (2020), 200432.
- [9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning*. 1180–1189.
- [10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 590–597.
- [15] Baoyu Jing, Chenwei Lu, Deqing Wang, Fuzhen Zhuang, and Cheng Niu. 2018. Cross-Domain Labeled LDA for Cross-Domain Text Classification. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 187–196.
- [16] Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6570–6580.
- [17] Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195* (2017).
- [18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [19] Paras Lakhani and Baskaran Sundaram. 2017. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 284, 2 (2017), 574–582.
- [20] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. 2020. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* (2020), 200905.
- [21] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in neural information processing systems*. 1530–1540.
- [22] Zhong Qiu Lin Linda Wang and Alexander Wong. 2020. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest Radiography Images. *arXiv:cs.CV/2003.09871*
- [23] Shervin Minaee, Rahele Kafeh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamaliour Soufi. 2020. Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. *arXiv preprint arXiv:2004.09363* (2020).
- [24] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.
- [25] Pau Panareda Busto and Juergen Gall. 2017. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*. 754–763.
- [26] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [27] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*. 8050–8058.
- [28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [29] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial Discriminative Domain Adaptation. In *CVPR IEEE*. 2962–2971.
- [30] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).
- [31] Deqing Wang, Baoyu Jing, Chenwei Lu, Junjie Wu, Guannan Liu, Chenguang Du, and Fuzhen Zhuang. 2020. Coarse Alignment of Topic and Sentiment: A Unified Model for Cross-Lingual Sentiment Classification. *IEEE Transactions on Neural Networks and Learning Systems* (2020).
- [32] Xaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadhi Bagheri, and Ronald M Summers. 2017. 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*. 2097–2106.
- [33] Xaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tinet: 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*. 9049–9058.
- [34] Zeya Wang, Baoyu Jing, Yang Ni, Nanqing Dong, Pengtao Xie, and Eric P Xing. 2019. Adversarial Domain Adaptation Being Aware of Class Relationships. *arXiv preprint arXiv:1905.11931* (2019).
- [35] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).
- [36] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2019. Universal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2720–2729.