MetaTeacher: Coordinating Multi-Model Domain Adaptation for Medical Image Classification

Zhenbin Wang[1], **Mao Ye**[1]*, **Xiatian Zhu**[2], **Liuhan Peng**[3], **Liang Tian**[1], **Yingying Zhu**[4]
[1]University of Electronic Science and Technology of China, Chengdu, China
[2]University of Surrey, Guildford, UK
[3]Xinjiang University, Ürümqi, China
[4]University of Texas, Arlington, US
zhenbinwang@foxmail.com, cvlab.uestc@gmail.com, xiatian.zhu@surrey.ac.uk

Abstract

In medical image analysis, we often need to build an image recognition system for a target scenario with the access to small labeled data and abundant unlabeled data, as well as multiple related models pretrained on different source scenarios. This presents the combined challenges of multi-source-free domain adaptation and semi-supervised learning *simultaneously*. However, both problems are typically studied independently in the literature, and how to effectively combine existing methods is non-trivial in design. In this work, we introduce a novel MetaTeacher framework with three key components: (1) A learnable coordinating scheme for adaptive domain adaptation of individual source models, (2) A mutual feedback mechanism between the target model and source models for more coherent learning, and (3) A semi-supervised bilevel optimization algorithm for consistently organizing the adaption of source models and the learning of target model. It aims to leverage the knowledge of source models adaptively whilst maximize their complementary benefits collectively to counter the challenge of limited supervision. Extensive experiments on five chest x-ray image datasets show that our method outperforms clearly all the state-of-the-art alternatives. The code is available at https://github.com/wongzbb/metateacher.

1 Introduction

Despite the great stride made by existing deep learning methods on medical image classification results [32, 53, 67], their performances often degrade drastically when applied to a new unseen scenario. This is mainly due to the domain shift challenge between the training and test data, caused by different environments, different instruments, and different acquisition protocols. Unlike natural images, annotating medical images requires special clinical expertise. It is hence more difficult to obtain large-scale medical image datasets with high-quality labels at every single scenario. Domain adaptation is a feasible solution, but comes with several limitations. Firstly, medical data is often under strict privacy and license constraints. That means the source domain data is usually inaccessible during domain adaptation. Secondly, medical data is typically multi-labeled which means that there are multiple labels for a sample, and the multiple categories are not mutually exclusive. It has more prominent different characteristics in different scenarios. Considering these practical constraints, we propose a new *Semi-supervised Multi-source-free Domain Adaptation* (SMDA) problem setting in the context of medical image classification. Our proposed setting has three key conditions: (1) There are multiple source domain models trained on respective multi-label medical image datasets; (2) All

[*] The corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the source domain data is inaccessible for adaptation; and (3) The target domain data has only a small number of labelled samples along with abundant unlabeled data.

In medical image classification, there are limited domain adaptation works, with a need of accessing the source domain data [5, 19, 24, 34, 45, 55, 57, 62]. Further, they usually consider a single source domain. On the other hand, for employing multiple source domains, existing Multi-Source Domain Adaptation (MSDA) methods typically learn a common feature space for all source and target domains [58] or use ensemble methods combined with source classifiers [8]. However, all of these MSDA methods require access to the source domain data. Regarding multi-label medical image classification, there exists a solution which extends the standard classifier network by conditional adversarial discriminator networks [46]. But it is still not source-free. Indeed, there have been extensive study on Source-Free Domain Adaptation (SFDA) [35, 64]. However, they are not directly applicable to our problem. Firstly, most of them assume a single source domain [35, 64]. Using a SFDA method to transfer each source domain model to the target domain separately and average their predictions, this strategy cannot reveal the complementary information between different source domains. Secondly, the source model is often domain biased. Different hospitals are featured with different populations, leading to a situation that the source datasets focus on a specific set of class labels. The existing SFDA methods can not assess the credibility of a source domain model with different labels.

To address the above SMDA's limitations, employing knowledge distillation from multi-source models to the target domain can be considered [18, 42, 65, 69, 70]. This forms a multi-teacher and one-student scheme. In our problem setting, a few labels of the target domain are provided to judge the credibility of multi-source models in different labels. In reality, it is common to exploit a few labeled data in the target domain. Recent works [25, 29, 50, 51] have shown that a few labeled data from the target domain can significantly improve the performance of the model. Inspired by meta-learning approaches [40, 47, 49], we consider a bilevel optimization strategy to update both the teachers and students. This is because different models vary in reliability and there is a need for optimizing the update direction for each source model. This offers an opportunity of leveraging the complementary and collaboration of different source models during model optimization, critical for solving the low-supervision challenge.

Based on the above analysis and consideration, we propose a novel framework, namely **MetaTeacher**. Specifically, it is based on multi-teacher and one-student models. Each teacher model is pre-trained on a specific labeled source data. The student model is initialized by a randomly chosen teacher. In order to provide different update directions for multiple teachers, a coordinating weight learning method is proposed to determine the contribution of each teacher for each target sample. In addition to knowledge transfer from multiple teachers, when adapting a specific teacher model, we also explore the feedback from the student and other teachers in a semi-supervised meta learning manner [16, 47]. Unlike the previous MSDA approaches, MetaTeacher can adapt each teacher in different directions according to the learned coordinating weight. This enables us to fully use different characteristics of source models, whilst avoiding the problem of insufficient training samples for multi-label classification to some extent.

Our **contributions** are summarized as follows: (1) We propose a new problem setting, i.e., semi-supervised multi-source-free domain adaptation for multi-label medical image classification. To our best knowledge, our work is the first attempt at multi-source-free and semi-supervised domain adaptation in the field of transfer learning. (2) A novel framework, MetaTeacher, based on a multi-teacher and one-student scheme is introduced to solve the proposed SMDA problem. A mutual feedback mechanism is designed based on meta-learning between the target model and the source models for more coherent learning and adaptation. The knowledge from multiple source models are sufficiently leveraged. (3) A coordinating weight learning method is derived for dynamically revealing the performance differences of different source models over different classes. It is integrated with the semi-supervised bilevel optimization algorithm for consistently updating the teacher and student models. Extensive experiments on five well-known chest radiography datasets show that our approach outperforms state-of-the-art alternatives clearly, along with in-depth ablation studies for verifying the design of our model components.

2 Relate Works

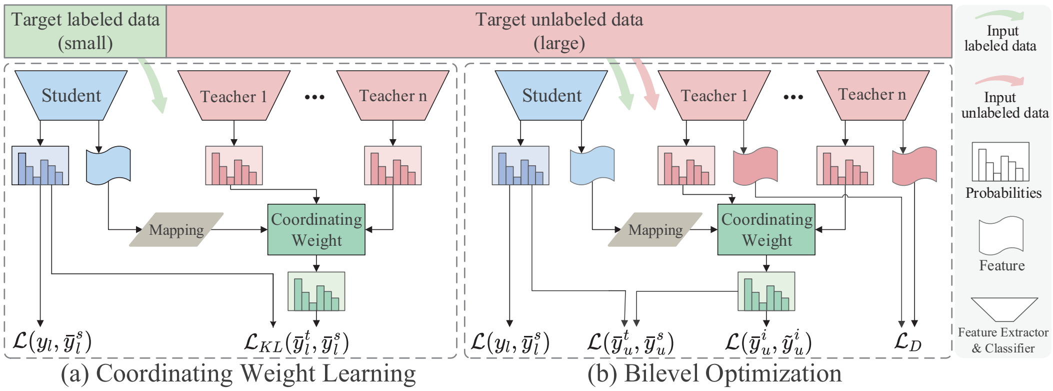
Unsupervised domain adaptation for medical image classification. There exist shallow UDA and deep UDA approaches in the literature. Shallow UDA approach adapts two routes, i.e., source domain instance weighting [55, 57] and feature transformation [24, 34]. All of these methods need to access source domain data. Similarly, there are also two routes for deep UDA approach. They are domain alignment based [19, 62] and pseudo-labeling based [5]. The first strategy solves the UDA problem by minimizing the domain difference between the source domain and target domain, and is currently the most popular method. Gao et al. [19] used the central moment difference matching to perform adaptation of classifying brain MRI data. The second strategy generates dummy data to retrain target model. For multi-label medical image classification, there exists a work based on domain alignment with a multi-label regularization term [46]. Bermúdez Chacón et al. [5] used the normalized cross-correlation to generate soft labels for the target domain. The above UDA methods do not update the source domain model, and they are all based on single-source domain. However, the situation of multi-source domains is very common in practical situations.

Source-free domain adaptation. Source-free domain adaptation methods can be roughly divided into two routes, i.e., generative approach [27, 28, 33, 63] and pseudo-label approach [4, 26, 35, 56]. The generation approach generates target-style training samples to train the prediction model. Since learning to generate features is difficult, this approach is extremely limited. The pseudo-label approach generates pseudo-labels through the source domain model, which is simple and general and has recently achieved good results in the machine learning community. The research of source-free domain adaptation in the medical image analysis field mainly focuses on image segmentation. Bateson et al. [4] maximized the mutual information between the target images and their label predictions to perform spine, prostate and cardiac segmentation. Vibashan et al. [56] implemented source-free domain adaptive image segmentation by generating pseudo-labels and applied self-training methods for task-specific representation. These works are all conducted in the single-source domain case. Currently, the research on multi-source-free domain adaptation is extremely limited, and most of the works adapt the method of generating trusted pseudo-labels [1, 14].

Multi-source domain adaptation for medical image classification. In machine learning community, MSDA works mainly have two strategies, i.e. distribution alignment [43, 74] and adversarial learning [61, 71, 72]. The first strategy computes the statistical discrepancy between multi-source domains and target domain, and then combines all predictions. The second strategy trains a domain discriminator and forces the feature extraction network to learn domain-invariant features to confuse the domain discriminator. For medical image classification, there only exist several shallow DA models. Wang et al. [58] proposed to map multiple source and target data to a common latent space for autism spectrum disorder classification. Cheng et al. [8] constructed a multi-domain transfer classifier for the early diagnosis of Alzheimer's disease. All of these strategies require to access source domain data and are not suitable for solving the proposed SMDA problem. To the best of our knowledge, current teacher-student domain adaptation methods in the medical and machine learning communities only consider the single-source domain case. When extended to the multi-source domain, it will face a challenging multi-objective optimization problem [10, 41].

Semi-supervised domain adaptation (SSDA). Our problem is also related to SSDA which assumes a small number of labeled samples in the target domain. Compared to UDA, using a few labeled samples of the target domain allows to further achieve better domain alignment [31, 44, 66]. Due to the shift of domain distribution, directly applying classical semi-supervised learning methods to the SSDA problem will lead to sub-optimal performance. Representative SSDA works are based on subspace learning [44, 66], entropy minimization [20, 50], label smoothing [13, 48] and active learning [48, 52]. However, all of these methods assume a single source domain with the source domain data accessible. Unlike these works, our method incorporates meta-learning and uses the performance on the labeled target data as a feedback signal.

Teacher-student domain adaptation models in medical image analysis. Usually, teacher-student domain adaptation model proposes multiple consistencies to solve UDA problem. To the best of our knowledge, teacher-student based domain adaptation methods have received little attention on medical image analysis. Perone et al. [45] proposed a semi-supervised learning based UDA method for medical image segmentation, which minimizes the consistency loss of the predicted results between the student model and the teacher model for unlabeled samples in the target domain during the training process. The network is updated by the exponential moving average of the student


Target labeled data (small)
Target unlabeled data (large)
Student
Teacher 1
•••
Teacher n
Mapping
Coordinating Weight
$\mathcal{L}(y_l, \bar{y}_l^s)$
$\mathcal{L}_{KL}(\bar{y}_l^t, \bar{y}_l^s)$
$\mathcal{L}(\bar{y}_u^t, \bar{y}_u^s)$
$\mathcal{L}(\bar{y}_u^i, \tilde{y}_u^i)$
$\mathcal{L}_D$
Input labeled data
Input unlabeled data
Probabilities
Feature
Feature Extractor & Classifier

(a) Coordinating Weight Learning

(b) Bilevel Optimization

Figure 1: Overview of MetaTeacher. (a) Learning the coordinating weight mapping which will be used subsequently to provide guidance for updating the teacher models. (b) Alternately updating the teacher and student models. Each teacher is updated with feedback signals from the student and other teachers.

network weights (mean-teacher [54]). The method is experimentally performed on the SCAM (Spinal Cord Anatomy MR Image) dataset to demonstrate its effectiveness. There are several approaches on teacher-student domain adaptation in the field of machine learning. French et al. [17] made some modifications to the mean-teacher scheme for the challenging domain adaptation of natural image classification. Cai et al. [7] proposed multiple consistency regulations to solve cross-domain detection problem. Deng et al. [12] combined the idea of feature alignment and data augmentation based on mean-teacher scheme. These methods all assume single-source domain, and to our knowledge, there is currently no work on multi-source teacher-student domain adaptation. Additionally, the mean-teacher approach does not sufficiently use the feedback signal from the target domain, so the performance improvement is limited.

3 Methodology

Problem statement. Suppose $D_T = \{(X_L^t, Y_L^t), X_U^t\}$ where Y_L^t denotes label annotations for a small amount of target domain samples X_L^t and X_U^t for target domain samples without any label annotations. The dimension of label vector is m. $D_{S_i} = \left\{\left(X_L^i, Y_L^i\right)\right\}$ where Y_L^i denotes label annotations for i-th source domain samples X_L^i. For the proposed semi-supervised multi-source-free domain adaptation problem, when the pretrained source classifiers f_{T_i} is applied to the target domain, the source dataset D_{S_i} is not accessible for $i = 1, \cdots, n$. Given the source classifiers f_{T_i} for $i = 1, \cdots, n$ and the target data D_T, the *objective* is to find a target-domain mapping $f_S : X_U^t \to Y_U^t$ where Y_U^t denotes the predicted labels for target domain samples X_U^t.

Overview. As shown in Fig.1, our framework is based on a *multi-teacher and one-student scheme*. First, multiple teacher models are pretrained according to each source domain, and then the student model is initialed using a randomly chosen teacher model. They are all composed of a feature extractor (e.g., Resnet50 [21]) and a multi-label classifier. The classifier consists of a fully connected layer, where the input is an one-dimensional expanded feature, and the output is the probability of each label. The objective function is the error loss between the predicted output and the ground truth.

Compared with traditional teacher-student models, our method is featured with two unique parts: (1) *Coordinating weight learning*; (2) *Bilevel optimization*. For the first part, a mapping is trained based on labeled target domain samples, which fuses the multi-teacher predictions adaptively for each target sample. This mapping will be used in the second part. In the initial iteration, the mapping and student model are trained based on labeled target samples. In the subsequent iterations, this part will only optimize the mapping while the student model will be updated by bilevel optimization. In the bilevel optimization part, the student and teacher models are updated alternately in a meta-learning manner. Specifically, for an unlabeled target sample, a coordinating weight is generated, which provides optimization direction for each teacher model. Finally, these two parts will be iteratively undated until convergence.

3.1 Coordinating Weight Learning

As mentioned earlier, the teacher models are trained on different source domain data. Due to different distributions, they often present different characteristics. Therefore, for a target domain sample, the classification probability of each teacher model could be inconsistent. When we want to optimize a teacher model based on the target domain samples, the optimization direction of each teacher model should be different. So it is necessary to obtain the contribution weight of each teacher model to the final classification results. We call this *coordinating weight*. Fortunately, we can obtain the weight mapping with the labeled samples in the target domain.

As shown in Fig.1(a), for obtaining the coordinating weight, we first input the labeled target sample x_l^t into the student network, and get the output $B = g(x_l^t)$ from feature extraction network g, where $B \in R^{c \times h \times w}$, with c, h, and w the number of channels, height, and width respectively. Then, we perform a maximum pool operation on the feature map B to get $\psi \in R^{1 \times c}$ which retains the most important information of each channel. Our mapping consists of two learnable variables μ and ν, where $\mu \in R^{n \times 1}$, $\nu \in R^{c \times m}$. Then, we define a mapping $\phi = \mu \psi \nu \in R^{n \times m}$ for the target sample x_l^t. After normalizing, we get the coordinating weight matrix W where

$$W_{j,k} = \frac{exp(\phi_{j,k})}{\sum_{z=1}^{n} exp(\phi_{z,k})}. \tag{1}$$

Suppose for the sample x_l^t, the predictions of all teachers are formed as a matrix $P \in R^{n \times m}$. By taking the Hadamard product between the teacher predictions and the coordinating weight matrix, we obtain the fused prediction as the following,

$$\bar{y}_l^t = Sum(P \circ W) \tag{2}$$

where $Sum(\cdot)$ means adding by rows. Denoting $\bar{y}_l^s = f_S(x_l^t; \theta_S)$ as the student prediction on the target sample x_l^t, we train the weight mapping and initialize student network using the following loss,

$$\mathcal{L}_W = \mathcal{L}\left(\bar{y}_l^s, y_l\right) + \alpha \mathcal{L}_{KL}\left(\bar{y}_l^t, \bar{y}_l^s\right) + \beta\left(\|\mu\| + \|\nu\|\right) \tag{3}$$

where $\mathcal{L}\left(\bar{y}_l^s, y_l\right) = \frac{1}{m} \sum_{i=1}^{m} [y_{l,i} log(\bar{y}_{l,i}^s) + (1 - y_{l,i}) log(1 - \bar{y}_{l,i}^s)]$ represents the BCE (Binary Cross Entropy) loss, y_l is the ground truth, θ_S is the parameter of student network. $\mathcal{L}_{KL}\left(\bar{y}_l^t, \bar{y}_l^s\right) = \sum_{i=1}^{m} \bar{y}_{l,i}^t \log(\bar{y}_{l,i}^t / \bar{y}_{l,i}^s)$ represents the KL (Kullback-Leibler divergence) loss which measures the distribution difference between the fused teacher prediction and student prediction. α and β are two balance parameters.

Remark. The mapping ϕ generates coordinating weight with Eq.(1). It not only reveals the complementarity of different teachers on different instances, but also, more interestingly, participates in the derivation of the update formula of teacher models in the bilevel optimization process (see Appendix), providing a reference for the update direction of different teachers.

3.2 Bilevel Optimization

The bilevel optimization problem [6, 9] was first proposed in the field of game theory. It includes an *upper-level optimization task* and a *lower-level optimization task*, where the former contains the latter as a constraint. Here, the upper-level optimization task (student) provides feedback signals to the lower-level optimization tasks (teachers) through the performance on labeled data and the coordinating weight mapping. For an unlabeled target sample x_u^t, suppose the pseudo-label based on the learned coordinating weight mapping ϕ from multi-teachers Eq.(2) is $\bar{y}_u^t$ and the corresponding coordinating weight matrix is W_u, we can define a loss function Γ_u as follows,

$$\Gamma_u(\theta_{T_1}, \cdots, \theta_{T_n}, \theta_S) = \mathcal{L}(\bar{y}_u^t, \bar{y}_u^s) \tag{4}$$

where $\bar{y}_u^s = f_S(x_u^t; \theta_S)$, θ_{T_i} is the parameter of the i-th teacher network. Similarly, a loss function $\Gamma_l\left(\theta_{T_1}, \cdots, \theta_{T_n}, \theta_S\right) = \mathcal{L}\left(y_l, \bar{y}_l^s\right)$ is defined for a labeled target samples x_l^t. In the bilevel optimization task, updating θ_S is the upper-level optimization task objective, while updating $\theta_{T_1}, \cdots, \theta_{T_n}$ is the lower-level optimization task objective. The upper-level optimization task and the lower-level optimization task are mutually constrained. To reach the lower-level optimization task objective, the performance of the upper-level optimization task objective on the labeled target data is utilized as *feedback signal*. We formulate the objective function in lower-level optimization task as the following,

$$\underset{\theta_{T_1},\cdots,\theta_{T_n}}{\operatorname{argmin}}\ \Gamma_l\left(\theta_{T_1},\cdots,\theta_{T_n},\theta_S^{OP}\right) \quad \text{s.t.} \quad \theta_S^{OP}=\underset{\theta_S}{\operatorname{argmin}}\ \Gamma_u\left(\theta_{T_1},\cdots,\theta_{T_n},\theta_S\right). \tag{5}$$

Eq.(5) cannot be optimized simply by gradient descent algorithm, because the teacher's parameters can not be updated until θ_S reaches the optimum. To overcome this issue, we resort to the idea of meta-learning [16, 38, 47] by making a one-step approximation of the problem,

$$\theta_S^{OP} \approx \theta_S - \eta_S \cdot \nabla_{\theta_s}\Gamma_u\left(\theta_{T_1},\theta_{T_2},\cdots,\theta_{T_n},\theta_S\right) \tag{6}$$

where η_S is the learning rate of the student network. Substituting Eq. (6) into Eq. (5), we obtain a new optimization objective function

$$\Gamma_l\left(\theta_{T_1},\cdots,\theta_{T_n},\theta_S-\eta_S\cdot\nabla_{\theta_s}\Gamma_u\left(\theta_{T_1},\theta_{T_2},\cdots,\theta_{T_n},\theta_S\right)\right). \tag{7}$$

By optimizing Eq. (7) (see Appendix), we get the following update rules,

$$\theta'_S = \theta_S - \eta_S \cdot \nabla_{\theta_s}\Gamma_u, \tag{8}$$

$$\theta'_{T_i} = \theta_{T_i} - \eta_{T_i} \cdot \left[(\nabla_{\theta'_S}\Gamma_l)^T \cdot \nabla_{\theta_S}\Gamma_u\right]^T \cdot \nabla_{\theta_{T_i}}\mathcal{L}\left(\bar{y}_u^i,\tilde{y}_u^i\right) \tag{9}$$

for $i=1,\cdots,n$, where θ'_S and θ'_{T_i} are the updated parameters corresponding to the student and teachers respectively. $\bar{y}_u^i = f_{T_i}\left(x_u^t;\theta_{T_i}\right)\cdot W_u^i$ with W_u^i the ith-row coordinating weight vector of W_u *w.r.t* the i-th teacher. We obtain the pseudo label $\tilde{y}_u^i$ by binarizing $\bar{y}_u^i$ as: $\tilde{y}_{u,j}^i = 0$ when $\bar{y}_{u,j}^i < 0.5$ and $\tilde{y}_{u,j}^i = 1$ for the other cases.

Additionally, in order to prevent optimizing teachers in the same direction, the predictions of the updated multiple teachers should be as far away from each other as possible. To that end, we define a divergence loss as follows,

$$\mathcal{L}_D = -ln\sum_{j=1,j\neq i}^{n}\mathcal{L}_2\left(B_{T_i}\left(x_u^t;\theta_{T_i}\right),B_{T_j}\left(x_u^t;\theta_{T_j}\right)\right) \tag{10}$$

where $B_{T_i}\left(x_u;\theta_{T_i}\right)$ represents the max-pooled results of the output feature map of the i-th teacher network. Here, we apply a max-pooling operation to the output features of multiple teachers and calculate the distance with L_2 norm. By requiring these feature maps to be far away each other, the optimization direction of teachers will be effectively adjusted. Finally, we update the i-th teacher network by the following rule,

$$\theta'_{T_i} = \theta_{T_i} - \eta_{T_i} \cdot \left(\left[(\nabla_{\theta_{S'}}\Gamma_l)^T \cdot \nabla_{\theta_S}\Gamma_u\right]^T \cdot \nabla_{\theta_{T_i}}\mathcal{L}\left(\bar{y}_u^i,\tilde{y}_u^i\right) + \gamma\nabla_{\theta_{T_i}}\mathcal{L}_D\right) \tag{11}$$

where γ is a hyperparameter.

Remark. Eq.(11) reveals that the update direction of θ_{T_i} is determined by three factors: (1) Coordinating weight confuses the feedback signals from different teachers; (2) Student network parameters provide the feedback signals and generate coordinating weight; (3) Diversity constraint emphasizes the characteristic of different teacher networks. Interestingly, these three factors change over time during the meta-learning process. In addition to alternating updates of the student and teacher models, we also update the mapping periodically.

3.3 Optimization Process

The optimization process alternates between upper-level optimization and lower-level optimization. Since the coordinating weight mapping is fixed in the process of bilevel optimization, with the progress of meta-learning, the coordinating weight W_u can not reflect the internal relationship between the source domains and the target domain gradually. Therefore, we use the learned teachers to update the mapping at regular intervals. The training process is summarized in Algorithm 1.

4 Experiments

Datasets. Five publicly available chest x-ray datasets are used to construct our multi-domain adaptation scenarios. *NIH-CXR14* [59] is a large public dataset of chest x-ray, which contains 108,948

Algorithm 1 Our proposed MetaTeacher method.

```
Require: Student network parameters $S^{(0)}$, teacher network parameters $T_1^{(0)} \sim T_n^{(0)}$, labeled data $(x_l^t, y_l)$, unlabeled data $(x_u^t)$, hyperparameters $\alpha, \beta, \gamma$, mapping updating interval $\mathcal{T}$.
Ensure: Optimized student model $S^{(N)}$.
function MetaTeacher($S^{(0)}, T_1^{(0)} \sim T_n^{(0)}, \alpha, \beta, \gamma, \mathcal{T}$)
    $S^{(0)}$, mapping $\leftarrow$ Coordinating Weight Learning
    for $t = 0 \rightarrow N-1$ do
        Upper-level optimization:
        Compute gradient $\nabla_{\theta_{S^{(t)}}} \mathcal{L}_u$
        Update the student: $\theta_{S^{(t+1)}} \leftarrow \theta_{S^{(t)}} - \eta_S \nabla_{\theta_{S^t}} \mathcal{L}_u$        ▷ Eq.(8)
        Lower-level optimization:
        Compute gradient $\nabla_{\theta_{S^{(t+1)}}} \mathcal{L}_l$
        for all $T_1^{(t)} \sim T_n^{(t)}$ do
            Compute gradient $\nabla_{\theta_{T_i^{(t)}}} \mathcal{L}\left(\bar{y}_u^i, \tilde{y}_u^i\right)$
            Compute gradient $\nabla_{\theta_{T_i^{(t)}}} \mathcal{L}_D$        ▷ Eq.(10)
            Update the $i$-th teacher: $\theta_{T_i^{(t+1)}} \leftarrow \theta_{T_i^{(t)}} - \eta_{T_i} \cdot$        ▷ Eq.(11)
                $([(\nabla_{\theta_{S^{(t+1)}}} \Gamma_l)^T \cdot \nabla_{\theta_{S^{(t)}}} \Gamma_u]^T \cdot \nabla_{\theta_{T_i^{(t)}}} \mathcal{L}\left(\bar{y}_u^i, \tilde{y}_u^i\right) + \gamma \nabla_{\theta_{T_i^{(t)}}} \mathcal{L}_D)$
        end for
        if $t \% \mathcal{T} = 0$ then
            mapping $\leftarrow$ Coordinating Weight Learning
        end if
    end for
    return $S^{(N)}$
end function
```

front view x-ray images of 32,717 patients collected from NIH Clinical Center, with a total of 14 disease labels. *MIMIC-CXR* [23] contains 377,110 images and text reports, corresponding to 227,835 radiological studies conducted by Beth Israel Deaconess Medical Center in Boston, Massachusetts. *CheXpert* [22] consists of 224,316 chest x-ray of 65,240 patients. The dataset collected chest x-ray examinations and related radiology reports performed at inpatient and outpatient centers at Stanford Hospital from October 2002 to July 2017. *Open-i* [11] is collected by Indiana University Hospital through the network from open source literature and biomedical image collection. It contains 3955 radiology reports, corresponding to 7470 frontal and lateral chest films. To be consistent with other datasets, we filter out the side chest x-ray in Open-I, leaving only 3955 frontal images. *Google-Health-CXR* [3] is manually labeled by medical experts for CXR images with high accuracy and contains about 4000 images. We follow the traditional UDA setting, and choose the disease closed set in these five datasets as multi classification labels, i.e., Atelectasis, Cardiomegaly, Effusion, Consolidation, Edema and Pneumonia. Four transfer scenarios are constructed, which are *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*; *NIH-CXR14, CheXpert, MIMIC-CXR* to *Google-Health-CXR*; *CheXpert, MIMIC-CXR to NIH-CXR14* and *NIH-CXR14, CheXpert* to *Open-i*.

Implementation details. In order to make a compromise between images in different datasets, we scale the images to 128*128 before feeding them into the network. To expand the training set, several data augmentation techniques are used, including random cropping and horizontal flipping. SGD with momentum of 0.9 is used as the optimizer. For the student model, the initial learning rate is 0.01 and the weight decay is 5e-4. The learning rate for coordinating weight mapping is 0.001; For the teacher models, the initial learning rate is 0.001 and the weight decay is 5e-6. The values of α, β and γ are set as 0.5, 0.01 and 0.01 respectively. For the case when the target domains datasets are small-scale, such as *Open-i* and *Google-Health-CXR*, we assume that there are 200 labeled data in the target domains, and in order to give a good initial condition for training, we randomly select a source model to initialize the target model. For the case when the target domains datasets are large-scale, such as *NIH-CXR14*, we assume that there are 500 labeled data in the target domains. Unless otherwise specified, the interval for updating coordinating weight mapping is set as 100 iterations. Following the setting of multi-label medical image classification problems, the evaluation criterion is Area Under the Receiver Operating Characteristic (AUROC) [15] curve score.

Table 1: Comparing the state-of-the-art methods on the transfer from *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. Metric: AUROC.

Method	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edema	Pneumonia	*Average*
DECISION [1]	**83.27**	91.55	96.18	97.02	92.74	89.24	91.67
CAiDA [14]	82.45	92.16	95.12	95.92	89.89	90.37	90.99
SHOT-best [35]	81.48	91.22	94.19	95.10	88.96	89.58	90.09
MME [50]	82.44	90.82	95.46	96.07	90.26	87.20	90.38
ECACL [30]	82.60	92.18	96.32	95.97	90.70	89.61	91.23
Source Only(N)	83.09	87.20	96.11	95.10	86.87	77.40	87.63
Source Only(C)	82.26	87.64	94.71	96.61	90.22	75.12	87.76
Source Only(M)	80.63	91.31	94.87	94.53	84.91	82.78	88.05
Fine-tune(*average*)	82.14	88.71	95.32	95.52	88.77	78.48	88.16
MetaTeacher(w/o *mapping*)	79.99	**92.64**	**98.22**	93.64	**95.50**	84.54	90.76
MetaTeacher(w/o *update*)	81.98	90.72	95.76	95.51	89.40	82.53	89.32
MetaTeacher(*all*)	81.72	92.59	96.25	**97.64**	94.52	**94.33**	**92.84**

Table 2: Comparing the state-of-the-art methods on the transfer from *NIH-CXR14, CheXpert, MIMIC-CXR* to *Google-Health-CXR*. Metric: AUROC.

Method	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edema	Pneumonia	*Average*
DECISION [1]	77.24	81.71	85.94	79.03	83.48	83.68	81.85
CAiDA [14]	76.90	81.82	87.55	79.62	85.10	82.72	82.29
SHOT-best [35]	75.43	80.28	86.63	77.88	82.37	81.22	80.64
MME [50]	77.34	**84.93**	86.17	78.65	85.33	71.28	80.62
ECACL [30]	76.27	84.54	87.06	79.95	85.82	72.66	81.05
Source Only(N)	76.54	84.48	86.36	75.66	83.94	62.59	78.26
Source Only(C)	72.09	76.45	84.55	79.07	68.25	58.39	73.13
Source Only(M)	68.04	79.38	84.17	72.41	68.71	52.60	70.88
Fine-tune(*average*)	73.48	80.14	85.96	74.17	74.74	60.20	74.78
MetaTeacher(w/o *mapping*)	75.62	83.91	85.40	**80.27**	75.13	81.77	80.35
MetaTeacher(w/o *update*)	76.75	84.30	86.67	78.59	82.31	65.84	79.08
MetaTeacher(all)	**77.65**	79.52	**88.73**	78.74	**86.73**	**84.78**	**82.69**

4.1 Comparisons to State-of-the-Art

At present, there does not exist any experimental report on our problem setting. So we choose four category of methods for compare. The first category is Source only which means directly applying a teacher model to the target domain. The second category is Fine-tune(*average*) which fine-tune each teacher network using labeled target domain data, then average their predicted values. The third category is the state-of-the-art multi-source-free domain adaptation methods, which are DECISION [1], CAiDA [14], and SHOT-best. The SHOT-best refers to adapting each source domain separately through the SHOT [35] method. The model with the best performance on the validation set is selected. The final category is semi-supervised domain adaptation methods, including MME [50] and ECACL [30]. For the semi-supervised domain adaptation methods, we assume that the labeled target data are the same as our method. Since they are single-source based methods, we perform domain adaptation for each source model and take the best result.

Tables 1-4 show the comparison results on four transfer scenarios. MetaTeacher(all) is our proposed method. Source Only(N), Source Only(C) and Source Only(M) are the teacher models respect to the *NIH-CXR14*, *CheXpert* and *MIMIC-CXR* datasets respectively. For the scenario from *CheXpert*, *MIMIC-CXR* to *NIH-CXR14*, since the dataset *NIH-CXR14* contains 108,948 x-ray images, different from other scenarios, this time we do not need to initialize the target model with the source models. It can be observed that our method achieves the best performance. The extensive experiments on four different transfer scenarios verify the adaptability of our method under multi-label chest x-ray dataset transfer cases. For the scenario from *NIH-CXR14*, *CheXpert* to *Open-i*, as show in Table 4, the

Table 3: Comparing the state-of-the-art methods on the transfer from *CheXpert, MIMIC-CXR* to *NIH-CXR14*. Metric: AUROC .

Method	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edema	Pneumonia	Average
DECISION [1]	72.99	80.73	79.37	**75.52**	82.30	71.38	77.05
CAiDA [14]	72.64	81.12	80.25	74.73	81.02	70.44	76.70
SHOT-best [35]	70.79	79.62	79.24	72.25	80.79	69.65	75.39
MME [50]	72.90	81.73	81.01	73.11	81.03	71.52	76.88
ECACL [30]	72.41	81.98	**82.07**	72.92	80.82	**71.65**	76.98
Source Only(N)	72.31	80.52	79.42	69.66	77.95	67.37	74.54
Source Only(C)	70.45	79.66	79.98	68.26	78.01	70.82	73.86
Fine-tune(*average*)	71.52	80.29	80.08	68.97	78.02	69.05	74.66
MetaTeacher(w/o *mapping*)	72.05	81.58	78.36	72.94	82.19	69.82	76.16
MetaTeacher(w/o *update*)	72.24	80.69	79.56	69.80	78.13	70.55	75.16
MetaTeacher(all)	**73.63**	**86.64**	80.86	72.24	**86.68**	66.37	**77.74**

Table 4: Comparing the state-of-the-art methods on the transfer from *NIH-CXR14, CheXpert* to *Open-i*. Metric: AUROC.

Method	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edema	Pneumonia	*Average*
DECISION [1]	83.15	90.86	96.12	96.32	92.33	88.79	91.26
CAiDA [14]	82.38	91.97	94.89	95.30	89.81	90.44	90.80
SHOT-best [35]	81.48	91.22	94.19	95.10	88.96	89.58	90.09
MME [50]	81.46	90.40	94.86	**97.73**	89.79	87.31	90.26
ECACL [30]	82.22	88.76	96.04	96.85	**92.43**	87.90	90.70
Source Only(N)	83.09	87.20	96.11	95.10	86.87	77.40	87.63
Source Only(C)	82.26	87.64	94.71	96.61	90.22	75.12	87.76
Fine-tune(*average*)	82.66	87.98	95.85	95.67	88.58	77.02	87.96
MetaTeacher(w/o *mapping*)	**83.73**	**93.37**	96.04	97.30	91.51	82.34	90.72
MetaTeacher(w/o *update*)	82.70	88.91	95.47	95.48	88.96	78.85	88.40
MetaTeacher(all)	82.11	92.42	**96.80**	97.07	**92.20**	**91.27**	**91.98**

performance of two source domains is 0.86% lower than that of three source domains. Furthermore, MetaTeacher also has moderate training time and more clearer background (see Appendix).

4.2 Ablation Analysis and Discussion

Component analysis. In Tables 1-4, MetaTeacher(w/o *mapping*) represents that our proposed method removes the part of coordinating weight learning and optimization substituted by average. MetaTeacher(w/o *update*) means to remove the bilevel optimization process. In this situation, the weighted output of teachers is used to supervise the learning of student network. The results in the last three rows of Tables 1-4 show that these two parts are indispensable. It is worth mentioning that MetaTeacher(w/o *mapping*) still obtains promising performance due to the following reasons. First, for student updating, averaging predictions from multiple teachers is beneficial for student performance, consistent with the finding by [68]. Second, the fixed W is also involved in the teacher optimization. It means bilevel optimization contributes more gain to the overall performance than the coordinating weight learning. However, the coordinating weight learning can judge which disease category the teacher is good at by weight, knowledge with different weights can be learned from different teachers. Therefore, the results in each disease category are close to the predictions of the best teachers, such as Pneumonia in Table 1 and Atelectasis in Table 2 (also see Appendix).

Effects of proportion of labeled target data. Table 5 shows the influence of the amount of labeled data in the target domain on the transfer scenario of *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. The experimental results show that the performance slowly improves as the amount of labeled data increases; a small number of labeled target domain samples can achieve good results.

Table 5: Effect of the size of labeled target data on the transfer from *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. Metric: AUROC.

Number(propotion)	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edema	Pneumonia	*Average*
50(1.4%)	82.48	92.22	95.19	96.10	89.96	90.58	91.09
100(2.8%)	82.19	92.50	96.83	97.02	92.43	91.20	92.03
200(5.6%)	81.72	92.59	96.25	97.64	94.52	94.33	92.84
300(8.4%)	82.21	92.97	96.83	97.42	94.07	94.33	92.97

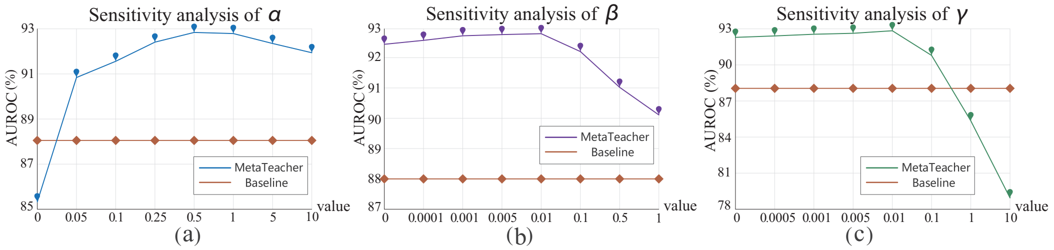


Figure 2: Effect of different hyperparameters on the transfer from *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. Baseline: source only(M).

Parameter analysis. We conduct parameter analysis experiment on the transfer scenario of *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. The basic strategy is to change a parameter while other parameters are fixed. Our method MetaTeacher has three hyperparameters, i.e., α and β in Eq. (3), and γ in Eq.(11). Fig.(2)(a) shows performance changing with the parameter α. When $\alpha = 0$, the coordinating weight mapping is not trained effectively resulting in the inability to determine the optimization direction of each teacher. When α gradually increases to around 0.5, the result achieve optimal performance. Fig.(2)(b) shows the influences of the parameters β. When the β is too large, it means that the coordinating weight learning part is ineffective and cannot express the relationship between the source domains. When β is set to 0, coordinating weight learning may overfit, which may cause coordinating weights to work well on some instances but poorly on other instances; for this case, the performance is 92.49% about 0.35% lower than the result 92.84% in Table 1. Fig.(2)(c) shows the influences of the parameter γ on divergence loss. When γ is set to 0.01, the performance reaches the best, but with the continuous increase of γ, the performance decreases obviously. When $\gamma = 0$, the result is 92.29%, which is 0.55% lower. We can also see that our method is also quite stable for the parameters α, β and γ in a large interval.

5 Conclusion

In this paper, we have proposed a novel framework, termed as MetaTeacher, for semi-supervised multi-source-free domain adaptation for medical image classification. The transfer learning process is modeled as a multi-teacher and one-student scheme. We not only optimize the student, but also optimize the teachers through the student's feedback in the target domain. Our optimization is based on meta-learning with two main parts: coordinating weight learning, and bilevel optimization. The first part obtains the coordinating weight mapping which is then used to coordinate the teacher outputs and updates. Bilevel optimization updates the student based on the pseudo-labeled data produced by the teachers and updates each teacher based on the feedback signal generated by the student and other teachers. Extensive experiments on multi-label chest x-ray datasets empirically demonstrated the superiority of our method over many state-of-the-art approaches.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Key R&D Program of China (2018YFE0203900), National Natural Science Foundation of China (62276048), Sichuan Science and Technology Program (2020YFG0476).

References

[1] Ahmed, S. M., Raychaudhuri, D. S., Paul, S., Oymak, S., and Roy-Chowdhury, A. K. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10103–10112.

[2] Aviles-Rivero, A. I., Papadakis, N., Li, R., Sellars, P., Fan, Q., Tan, R. T., and Schönlieb, C.-B. Graphxnet-chest x-ray classification under extreme minimal supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), Springer, pp. 504–512.

[3] Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., and Saalbach, A. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports 9*, 1 (2019), 1–10.

[4] Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., and Ayed, I. B. Source-free domain adaptation for image segmentation. *arXiv preprint arXiv:2108.03152* (2021).

[5] Bermúdez-Chacón, R., Márquez-Neila, P., Salzmann, M., and Fua, P. A domain-adaptive two-stream u-net for electron microscopy image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), IEEE, pp. 400–404.

[6] Bracken, J., and McGill, J. T. Mathematical programs with optimization problems in the constraints. *Operations Research 21*, 1 (1973), 37–44.

[7] Cai, Q., Pan, Y., Ngo, C.-W., Tian, X., Duan, L., and Yao, T. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11457–11466.

[8] Cheng, B., Liu, M., Shen, D., Li, Z., and Zhang, D. Multi-domain transfer learning for early diagnosis of alzheimer's disease. *Neuroinformatics 15*, 2 (2017), 115–132.

[9] Colson, B., Marcotte, P., and Savard, G. An overview of bilevel optimization. *Annals of operations research 153*, 1 (2007), 235–256.

[10] Deb, K. Multi-objective optimization. In *Search methodologies*. Springer, 2014, pp. 403–449.

[11] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association 23*, 2 (2016), 304–310.

[12] Deng, J., Li, W., Chen, Y., and Duan, L. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4091–4101.

[13] Donahue, J., Hoffman, J., Rodner, E., Saenko, K., and Darrell, T. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013), pp. 668–675.

[14] Dong, J., Fang, Z., Liu, A., Sun, G., and Liu, T. Confident anchor-induced multi-source free domain adaptation. *Advances in Neural Information Processing Systems 34* (2021).

[15] Fawcett, T. An introduction to roc analysis. *Pattern recognition letters 27*, 8 (2006), 861–874.

[16] Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (2017), PMLR, pp. 1126–1135.

[17] French, G., Mackiewicz, M., and Fisher, M. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208* (2017).

[18] Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. In *International Conference on Machine Learning* (2018), PMLR, pp. 1607–1616.

[19] Gao, Y., Zhang, Y., Cao, Z., Guo, X., and Zhang, J. Decoding brain states from fmri signals by using unsupervised domain adaptation. *IEEE Journal of Biomedical and Health Informatics 24*, 6 (2019), 1677–1685.

[20] Grandvalet, Y., and Bengio, Y. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems 17* (2004).

[21] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[22] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, 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* (2019), vol. 33, pp. 590–597.

[23] Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).

[24] Kamphenkel, J., Jäger, P. F., Bickelhaupt, S., Laun, F. B., Lederer, W., Daniel, H., Kuder, T. A., Delorme, S., Schlemmer, H.-P., König, F., et al. Domain adaptation for deviating acquisition protocols in cnn-based lesion classification on diffusion-weighted mr images. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 73–80.

[25] Kim, T., and Kim, C. Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. In *European conference on computer vision* (2020), Springer, pp. 591–607.

[26] Kim, Y., Cho, D., Han, K., Panda, P., and Hong, S. Domain adaptation without source data. *arXiv preprint arXiv:2007.01524* (2020).

[27] Kundu, J. N., Venkat, N., Babu, R. V., et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 4544–4553.

[28] Kurmi, V. K., Subramanian, V. K., and Namboodiri, V. P. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), pp. 615–625.

[29] Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., and Zhao, H. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1104–1113.

[30] Li, K., Liu, C., Zhao, H., Zhang, Y., and Fu, Y. Ecacl: A holistic framework for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 8578–8587.

[31] Li, L., and Zhang, Z. Semi-supervised domain adaptation by covariance matching. *IEEE transactions on pattern analysis and machine intelligence 41*, 11 (2018), 2724–2739.

[32] Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)* (2014), IEEE, pp. 844–848.

[33] Li, R., Jiao, Q., Cao, W., Wong, H.-S., and Wu, S. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9641–9650.

[34] Li, W., Zhao, Y., Chen, X., Xiao, Y., and Qin, Y. Detecting alzheimer's disease on small dataset: A knowledge transfer perspective. *IEEE journal of biomedical and health informatics 23*, 3 (2018), 1234–1242.

[35] Liang, J., Hu, D., and Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning* (2020), PMLR, pp. 6028–6039.

[36] Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., and Carneiro, G. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 20697–20706.

[37] Liu, F., Tian, Y., Cordeiro, F. R., Belagiannis, V., Reid, I., and Carneiro, G. Self-supervised mean teacher for semi-supervised chest x-ray classification. In *International Workshop on Machine Learning in Medical Imaging* (2021), Springer, pp. 426–436.

[38] Liu, H., Simonyan, K., and Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).

[39] Liu, Q., Yu, L., Luo, L., Dou, Q., and Heng, P. A. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging 39*, 11 (2020), 3429–3440.

[40] Madhawa, K., and Murata, T. Metal: Active semi-supervised learning on graphs via meta-learning. In *Asian Conference on Machine Learning* (2020), PMLR, pp. 561–576.

[41] Marler, R. T., and Arora, J. S. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization 26*, 6 (2004), 369–395.

[42] Park, S., and Kwak, N. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI 2020*. IOS Press, 2020, pp. 1411–1418.

[43] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 1406–1415.

[44] Pereira, L. A., and da Silva Torres, R. Semi-supervised transfer subspace for domain adaptation. *Pattern Recognition 75* (2018), 235–249.

[45] Perone, C. S., Ballester, P., Barros, R. C., and Cohen-Adad, J. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage 194* (2019), 1–11.

[46] Pham, D., Koesnadi, S., Dovletov, G., and Pauli, J. Unsupervised adversarial domain adaptation for multi-label classification of chest x-ray. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (2021), IEEE, pp. 1236–1240.

[47] Pham, H., Dai, Z., Xie, Q., and Le, Q. V. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 11557–11568.

[48] Prabhu, V., Chandrasekaran, A., Saenko, K., and Hoffman, J. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 8505–8514.

[49] Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676* (2018).

[50] Saito, K., Kim, D., Sclaroff, S., Darrell, T., and Saenko, K. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8050–8058.

[51] Singh, A. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems 34* (2021).

[52] Su, J.-C., Tsai, Y.-H., Sohn, K., Liu, B., Maji, S., and Chandraker, M. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 739–748.

[53] Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., and Lippert, C. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems 33* (2020), 18158–18172.

[54] Tarvainen, A., and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems 30* (2017).

[55] van Opbroek, A., Vernooij, M. W., Ikram, M. A., and de Bruijne, M. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. *Medical image analysis 24*, 1 (2015), 245–254.

[56] VS, V., Valanarasu, J. M. J., and Patel, V. M. Target and task specific source-free domain adaptive image segmentation. *arXiv preprint arXiv:2203.15792* (2022).

[57] Wachinger, C., Reuter, M., Initiative, A. D. N., et al. Domain adaptation for alzheimer's disease diagnostics. *Neuroimage 139* (2016), 470–479.

[58] Wang, J., Zhang, L., Wang, Q., Chen, L., Shi, J., Chen, X., Li, Z., and Shen, D. Multi-class asd classification based on functional connectivity and functional correlation tensor via multi-source domain adaptation and multi-view sparse representation. *IEEE transactions on medical imaging 39*, 10 (2020), 3137–3147.

[59] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and 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), pp. 2097–2106.

[60] Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning 8*, 3 (1992), 229–256.

[61] Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3964–3973.

[62] Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L., and Tao, Q. The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), Springer, pp. 623–631.

[63] Yang, C., Guo, X., Chen, Z., and Yuan, Y. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis* (2022), 102457.

[64] Yang, S., Wang, Y., van de Weijer, J., Herranz, L., and Jui, S. Unsupervised domain adaptation without source data by casting a bait. *arXiv e-prints* (2020), arXiv–2010.

[65] Yang, Z., Shou, L., Gong, M., Lin, W., and Jiang, D. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (2020), pp. 690–698.

[66] Yao, T., Pan, Y., Ngo, C.-W., Li, H., and Mei, T. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2015), pp. 2142–2150.

[67] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? *Advances in neural information processing systems 27* (2014).

[68] You, S., Xu, C., Xu, C., and Tao, D. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), pp. 1285–1294.

[69] Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., and Jiang, D. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'21)* (2021).

[70] Zhao, H., Sun, X., Dong, J., Chen, C., and Dong, Z. Highlight every step: Knowledge distillation via collaborative teaching. *IEEE Transactions on Cybernetics* (2020).

[71] Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. *Advances in neural information processing systems 31* (2018).

[72] Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., and Keutzer, K. Multi-source distilling domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 12975–12983.

[73] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2921–2929.

[74] Zhu, Y., Zhuang, F., and Wang, D. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 5989–5996.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the Implementation details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the Implementation details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Updating Rules of Bilevel Optimization

We follow the derivation route in [47] except the coordinating weight part. For expression clarity, let $|T_i|$ and $|S|$ denote the dimensions of θ_{T_i} and θ_S respectively, where $\theta_S \in R^{|S|\times 1}$ and $\theta_{T_i} \in R^{|T_i|\times 1}$. Suppose there is a batch of unlabeled target samples x_u^t, the j-th teacher T_j samples the pseudo labels $\hat{y}_u^j \sim f_{T_i}(x_u^t;\theta_{T_j})$ for $j = 1, \cdots, n$ and

$$\bar{y}_u^t = \sum_{j=1}^{n} \bar{y}_u^j = \sum_{j=1}^{n} \hat{y}_u^j \circ W_u^j \tag{12}$$

where W_u^j is the j-th row of coordinating weight matrix W_u *w.r.t* x_u^t and $\circ$ is the Hadamard product. So we can use $(x_u^t, \bar{y}_u^t)$ to update the parameter θ_S according to Eq.(6) in expectation as follows,

$$\theta'_S = E_{\bar{y}_u^t}\left[\theta_S - \eta_S \cdot \nabla_{\theta_s} \mathcal{L}\left(\bar{y}_u^t, f_S\left(x_u^t;\theta_S\right)\right)\right]. \tag{13}$$

According to Eq.(7), we update θ_{T_i} based on a batch of labeled target data (x_l^t, y_l) by optimization the following objective function

$$\underset{\theta_{T_1},\cdots,\theta_{T_n}}{\operatorname{argmin}} \mathcal{L}\left(y_l, f_S\left(x_l^t, \theta'_S\right)\right). \tag{14}$$

For end-to-end optimization with gradient descent, the partial derivative respect to the above objective function R is

$$\underbrace{\frac{\partial R}{\theta_{T_i}}}_{1\times|T_i|} = \frac{\partial \mathcal{L}\left(y_l, f_S\left(x_l^t;\theta'_S\right)\right)}{\partial \theta_{T_i}}, \tag{15}$$

for $i = 1, \cdots n$. According to the chain rule, Eq.(15) can be written as:

$$\underbrace{\frac{\partial R}{\theta_{T_i}}}_{1\times|T_i|} = \underbrace{\frac{\partial \mathcal{L}\left(y_l, f_S\left(x_l^t;\theta'_S\right)\right)}{\partial \theta'_S}}_{1\times|S|} \cdot \underbrace{\frac{\partial \theta'_S}{\partial \theta_{T_i}}}_{|S|\times|T_i|}. \tag{16}$$

For the right part of Eq.(16), it follows that

$$\underbrace{\frac{\partial \theta'_S}{\partial \theta_{T_i}}}_{|S|\times|T_i|} = \underbrace{\frac{\partial E\left[\theta_S - \eta_S \cdot \left(\frac{\partial \mathcal{L}\left(\bar{y}_u^t, f_S\left(x_u^t;\theta_S\right)\right)}{\partial \theta_S}\right)^{\mathsf{T}}\right]}{\partial \theta_{T_i}}}_{|S|\times|T_i|} = \underbrace{\frac{\partial E\left[-\eta_S \cdot \left(\frac{\partial \mathcal{L}\left(\bar{y}_u^t, f_S\left(x_u^t;\theta_S\right)\right)}{\partial \theta_S}\right)^{\mathsf{T}}\right]}{\partial \theta_{T_i}}}_{|S|\times|T_i|} \tag{17}$$

where $(\cdot)^T$ is the transpose notation. Suppose that

$$\underbrace{G(\theta_S, \bar{y}_u^t)}_{|S|\times 1} = \underbrace{\left(\frac{\partial \mathcal{L}\left(\bar{y}_u^t, f_S\left(x_u^t;\theta_S\right)\right)}{\partial \theta_S}\right)^{\mathsf{T}}}_{|S|\times 1}, \tag{18}$$

it follows that

$$\begin{aligned}
\underbrace{\frac{\partial \theta'_S}{\partial \theta_{T_i}}}_{|S|\times|T_i|} &= -\eta_S \cdot \underbrace{\frac{\partial E\left[G\left(\theta_S, \bar{y}_u^t\right)\right]}{\partial \theta_{T_i}}}_{|S|\times|T_i|} \\
&= -\eta_S \cdot \sum \underbrace{\frac{\partial \left[G\left(\theta_S, \bar{y}_u^t\right) \cdot P\left(\bar{y}_u^t \mid x_u^t : \theta_{T_1}, \theta_{T_2}, \cdots, \theta_{T_n}, W_u\right)\right]}{\partial \theta_{T_i}}}_{|S|\times|T_i|} \\
&= -\eta_S \cdot \sum \underbrace{G\left(\theta_S, \bar{y}_u^t\right)}_{|S|\times 1} \cdot \underbrace{\frac{\partial P\left(\bar{y}_u^t \mid x_u^t : \theta_{T_1}, \theta_{T_2}, \cdots, \theta_{T_n}, W_u\right)}{\partial \theta_{T_i}}}_{1\times|T_i|}.
\end{aligned} \tag{19}$$

Since $\bar{y}_u^t = \sum_{j=1}^n \bar{y}_u^j = \sum_{j=1}^n \hat{y}_u^j \circ W_u^j$, Eq.(19) can be further resolved

$$\begin{aligned}\underbrace{\frac{\partial \theta'_S}{\partial \theta_{T_i}}}_{|S|\times|T_i|} &= -\eta_S \cdot \sum \underbrace{G\left(\theta_S, \bar{y}_u^t\right)}_{|S|\times 1} \cdot \underbrace{\frac{\partial\left(\sum_{j=1}^n P\left(\bar{y}_u^j \mid x_u^t; \theta_{T_j}, W_u^j\right)\right)}{\partial \theta_{T_i}}}_{1\times|T_i|} \\ &= -\eta_S \cdot \sum \underbrace{G\left(\theta_S, \bar{y}_u^t\right)}_{|S|\times 1} \cdot \underbrace{\frac{\partial\left(P\left(\bar{y}_u^i \mid x_u^t; \theta_{T_i}, W_u^i\right)\right)}{\partial \theta_{T_i}}}_{1\times|T_i|}.\end{aligned} \tag{20}$$

From the REINFORCE equation [60], we can get

$$\begin{aligned}&\sum \underbrace{\frac{\partial\left(P\left(\bar{y}_u^i \mid x_u^t; \theta_{T_i}, W_u^i\right)\right)}{\partial \theta_{T_i}}}_{1\times|T_i|} \\ =&\sum \left(P\left(\bar{y}_u^i \mid x_u^t; \theta_{T_i}, W_u^i\right)\right) \cdot \underbrace{\frac{\partial \log\left(P\left(\bar{y}_u^i \mid x_u^t; \theta_{T_i}, W_u^i\right)\right)}{\partial \theta_{T_i}}}_{1\times|T_i|} \\ =&-E \underbrace{\left[\frac{\partial \mathcal{L}\left(\left(f_{T_i}\left(x_u^t; \theta_{T_i}\right) \cdot W_u^i\right), \tilde{y}_u^i\right)}{\partial \theta_{T_i}}\right]}_{1\times|T_i|}\end{aligned} \tag{21}$$

where $\tilde{y}_u^i$ is the pseudo labels after normalizing the values of $f_{T_i}\left(x_u^t; \theta_{T_i}\right) \cdot W_u^i$ to 0 or 1, i.e., $\tilde{y}_{u,j}^i = 0$ when $\hat{y}_{u,j}^i < 0.5$ and $\tilde{y}_{u,j}^i = 1$ for other cases. After substituting Eq.(21) into Eq.(19), Eq. (18) into Eq.(19), and Eq.(19) into Eq.(16), we obtain the following gradient,

$$\underbrace{\frac{\partial R}{\theta_{T_i}}}_{1\times|T_i|} = \eta_S \cdot \underbrace{\frac{\partial \mathcal{L}\left(y_l, f_S\left(x_l^t; \theta'_S\right)\right)}{\partial \theta'_S}}_{1\times|S|} \cdot E\left[\underbrace{G\left(\theta_S, \bar{y}_u^t\right)}_{|S|\times 1} \cdot \underbrace{\frac{\partial \mathcal{L}\left(f_{T_i}\left(x_u^t; \theta_{T_i}\right) \cdot W_u^i, \tilde{y}_u^i\right)}{\partial \theta_{T_i}}}_{1\times|T_i|}\right]. \tag{22}$$

By Monte Carlo approximation, we use the sampled $\hat{y}_u^j$ for $j = 1, \cdots, n$ to obtain the update rules,

$$\theta'_S = \theta_S - \eta_S \cdot \nabla_{\theta_s} \Gamma_u \tag{23}$$

$$\begin{aligned}\theta'_{T_i} = \theta_{T_i} - \eta_{T_i} \cdot \underbrace{\frac{\partial \mathcal{L}\left(y_l, f_S\left(x_l^t; \theta'_S\right)\right)}{\partial \theta_{S'}}}_{1\times|S|} \cdot \underbrace{\left(\frac{\partial \mathcal{L}\left(\bar{y}_u^t, f_S\left(x_u^t; \theta_S\right)\right)}{\partial \theta_S}\right)^{\mathsf{T}}}_{|S|\times 1} \cdot \underbrace{\frac{\partial \mathcal{L}\left(f_{T_i}\left(x_u^t; \theta_{T_i}\right) \cdot W_u^i, \tilde{y}_u^i\right)}{\partial \theta_{T_i}}}_{1\times|T_i|} \\ = \theta_{T_i} - \eta_{T_i} \cdot \left[\left(\nabla_{\theta_{S'}} \Gamma_l\right)^{\mathsf{T}} \cdot \nabla_{\theta_S} \Gamma_u\right]^{\mathsf{T}} \cdot \nabla_{\theta_{T_i}} \mathcal{L}\left(f_{T_i}\left(x_u^t; \theta_{T_i}\right) \cdot W_u^i, \tilde{y}_u^i\right).\end{aligned} \tag{24}$$

B Visualization

B.1 Visualization of ablation analysis

We visualize the domain adaptation performance on the transfer scenario *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. The visualization sample in the *Open-i* is suffering from Atelecsis and Effusion disease. The comparison models are 1) Source only(N): the model trained on the *NIH-CXR14* dataset; 2) Source only(C): the model trained on the *Chexpert* dataset; 3) Source only(M): the model trained on *MIMIC-CXR* dataset; 4) MetaTeacher(*w/o update*): our approach only containing coordinating weight learning part; 5) MetaTeacher(*all*): our full approach MetaTeacher containing both coordinating weight learning and bilevel optimization.

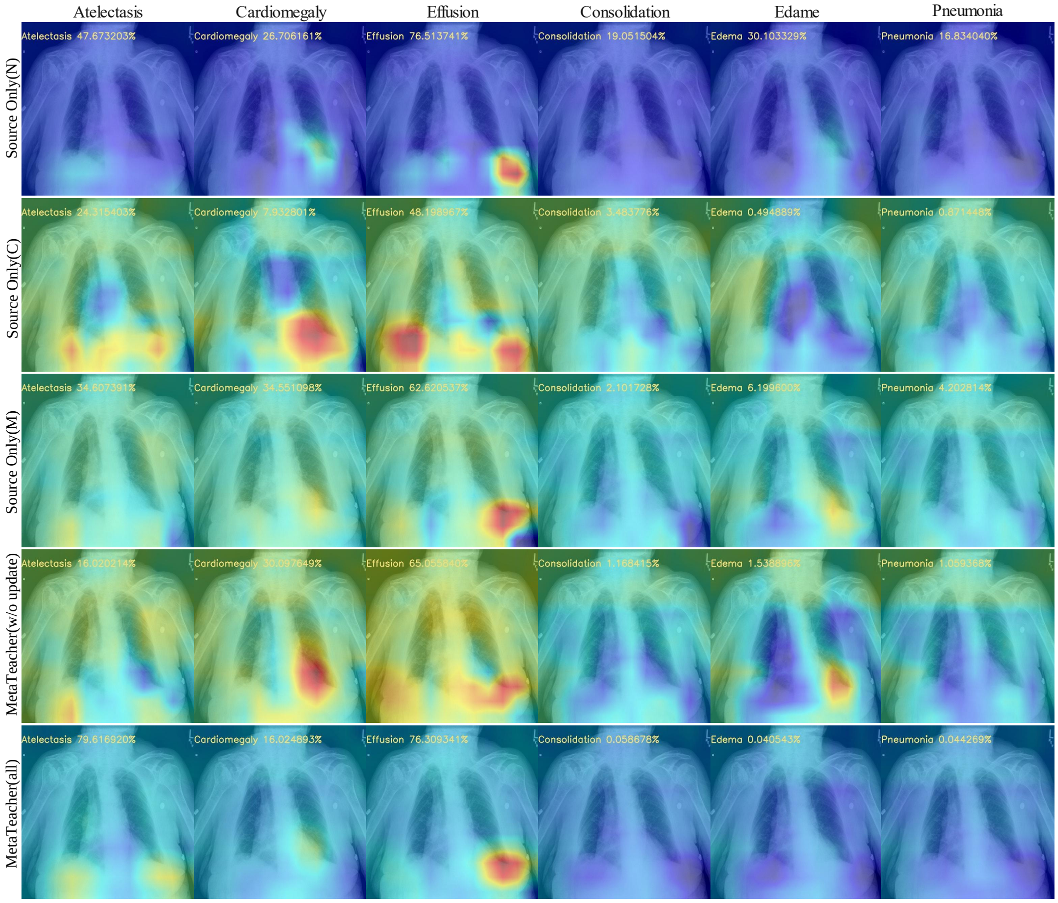
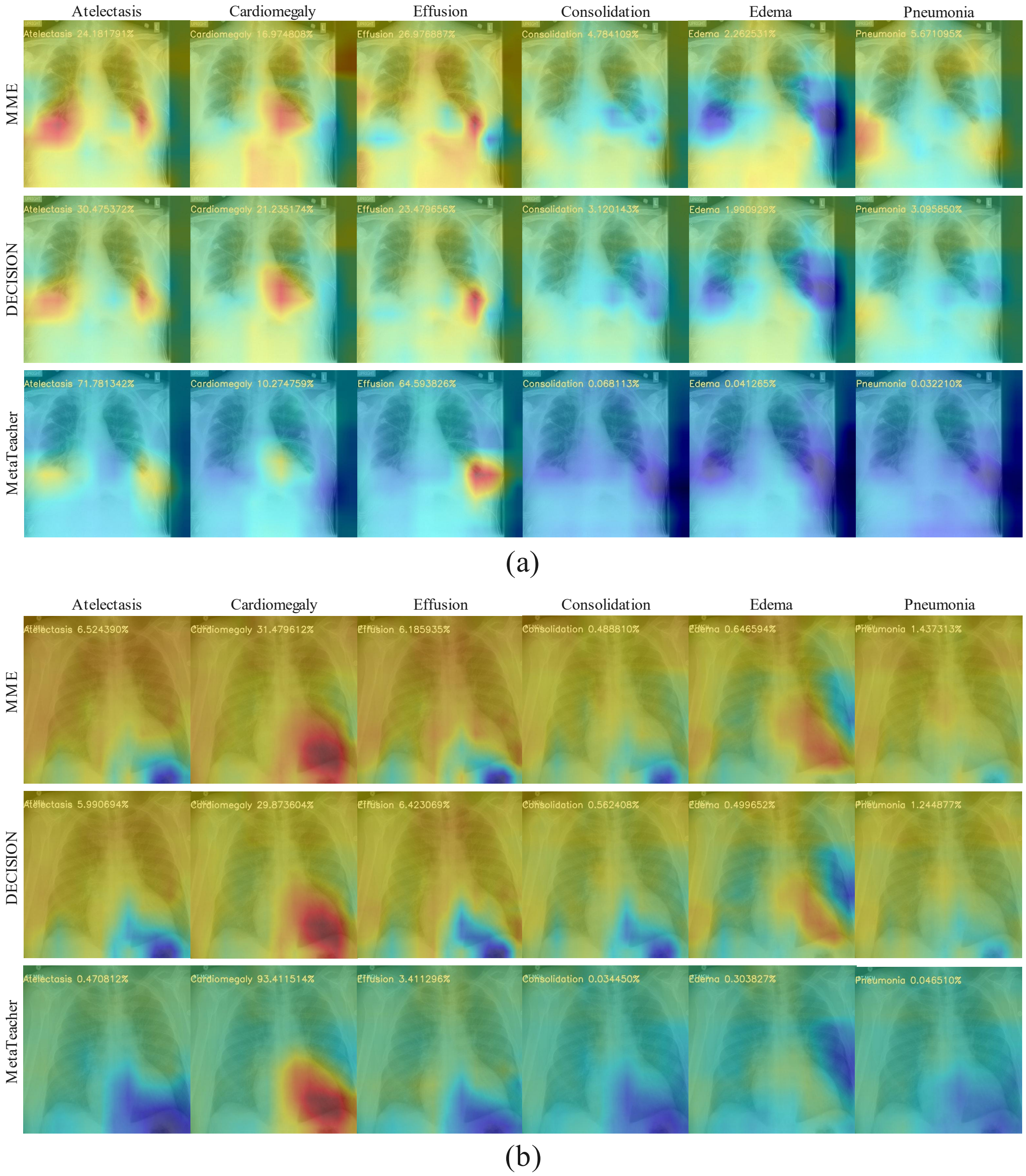


Figure 3: The Class Activation Map (CAM) [73] is used to perform visual ablation analysis on a chest x-ray image in *Open-i* dataset. The background color is blue, with red or yellow representing the disease location. The number on the top left corner of each image is the predicted probability for the corresponding disease. *Zoom in for best view.*

From the visualization results, we have the following observations. 1) The source models trained on different datasets have different concerns about different diseases. It can be seen that Source only(N) and Source only(M) can identify patients with Effusion disease, with probabilities of 76.513741% and 62.620537%, respectively. However, Source only(C) shows that the patient has only a 48.198967% probability of Effusion disease. 2) Simply fusing multiple teacher predictions does not work in the target domain. MetaTeacher(*w/o update*) is a distillation learning with coordinating weights, which can coordinate the knowledge of each teacher about each disease category. As shown in Fig.3, if most of the source models can detect some disease, the fused model can also detect this disease, and its probability is slightly lower than the maximum value such as Effusion disease. Conversely, if the disease cannot be detected by most of the source models, the fused model can not detected it too such as Atelectasis, Consolidation and Edame. 3) Collaboratively updating teacher and student models works in the target domain. MetaTeacher(*all*) can learn knowledge that the source model does not have. None of these three source domain models can accurately detect the Atelectasis disease, but MetaTeacher(*all*) can identify it, and the output probability is as high as 79.616920%. In addition, for Consolidation, Edame and Pneumonia diseases, MetaTeacher(*all*) predictions for them are close to 0, which shows that our method has a more clear judgment ability for non-existing diseases.

B.2 Visualization of different methods

The comparison models are DECISION [1] and MME [50]. The first method is a multi-source-free domain adaptation approach, which works by learning a set of weight values corresponding to each



(a)

(b)

Figure 4: The Class Activation Map (CAM) [73] is used to perform visual ablation analysis on two chest x-ray images in *Open-i* dataset. The background color is blue, with red or yellow representing the disease location. The number on the top left corner of each image is the predicted probability for the corresponding disease. *Zoom in for best view.*

source domain model, while learning these weights by using unlabeled target data, then combining the predictions from each source domain as the final prediction. To fit the problem setting, the performance of DECISION is visualized on the transfer scenario *NIH-CXR14, CheXpert,MIMIC-CXR* to *Open-i*. The second method is a single-source semi-supervised domain adaptation approach, which alternately maximizing the conditional entropy of unlabeled target data with respect to the classifier and minimizing it with respect to the feature encoder. Similarly to fit the problem setting, we visualize the MME performance on the transfer scenario *MIMIC-CXR* to *Open-i*. The visualization sample in Fig.4(a) is suffering from Atelecsis and Effusion disease while the sample in Fig.4(b) is suffering from Cardiomegaly disease.

Table 6: Training time comparison. Metric: minutes.

Method	DECISION [1]	MME [50]	MetaTeacher
NIH-CXR14, CheXpert, MIMIC-CXR to Open-i	32	41	36
NIH-CXR14, CheXpert, MIMIC-CXR to Google-Health-CXR	33	43	38

As shown in Fig.4, both of MME and DECISION cannot detect the corresponding diseases. From the visualization results, it can be seen that MME and DECISION contain a widely distributed yellow color, mixed with the red part, which affects their judgments of the disease. For example, Atelectasis and Effusion diseases in Fig.4(a), or Cardiomegaly disease in Fig.4(b), although MME and DECISION can mark the disease location in red, they also contain a lot of yellow color in other places, which confuse their attentions to the right diseases. Unlike them, MetaTeacher contains more blue background color, which can more clearly distinguish the background color from the disease location. Therefore, the disease can be judged more accurately. Additionally, for diseases that are clearly not present in the figure, such as Consolidation, Edema and Pneumonia diseases in Fig.4(a), or Atelectasis and Pneumonia in Fig.4(b). The widespread yellow color makes MME and DECISION more conservative in their predicted probabilities, while the predicted probabilities by MetaTeacher are closer to 0 compared to them. From the visualization results, it can be seen that MetaTeacher is more accurate.

C More Discussions

C.1 Training Time Comparison

The training runtime of MetaTeacher is compared with DECISION [1] and MME [50] on a single NVIDIA 3090Ti GPU over the transfer scenarios *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i* and *NIH-CXR14, CheXpert, MIMIC-CXR* to *Google-Health-CXR*. The results are shown in Table 6. It can be observed that our method MetaTeacher is slightly slower than the approach of multi-source-free domain adaptation (e.g., DECISION). Although multi-source-free domain adaptation methods do not need to update a student model, they involve other complex designs (e.g., DECISION need to do k-means clustering, and CAiDA [14] has a searching process of Semantic-Nearest Confident Anchor). Instead, MetaTeacher only involves some simple matrix calculation. The second observation is that MetaTeacher is slightly faster than the approach of semi-supervised domain adaptation (MME). This is because semi-supervised domain adaptation needs to train a model for each source domain, whilst other complex computations are involved in their optimization process. Overall, our method has similar running speed as existing alternative methods.

C.2 Probing the Behavior of Coordinating Weight

The coordinating weight is critical in our MetaTeacher framework. Firstly, for the upper-level optimization objective, it combines the predictions of multiple teachers to provide the updating direction for the student model. Secondly, for the lower-level optimization objective, we split coordinating weight into multiple vectors to provide different updating directions for each teacher. On the training of the transfer scenario *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*, we choose a sample labeled with Atelectasis and Effusion disease classes to inspect the behavior of coordinating weight. As shown in Table 7, initially, each teacher, as well as their joint predictions ($\bar{y}_u^t$ in Eq. (12)), failed to predict the Atelectasis disease. During training, each teacher was updated, with teacher 1 acquiring the ability to predict Atelectasis disease (0.371739 to 0.756016). Meanwhile, the coordinating weight was also accordingly updated and assigned more weight to the Atelectasis class for teacher 1 (0.214039 to 0.950953). This process is summarized in Table 7.

C.3 Performance Comparisons from Two Sources to Three Sources

Compared with multi-source-free domain adaptation methods, our MetaTeacher yields more significant gains in the multi-source transfer situation. On the two-teacher (Table 4) and three-teacher (Table 1) scenario, DECISION [1] method increases from 91.26% to 91.67% (a gain of 0.41%), CAiDA [14] method from 90.80% to 90.99% (0.19%), *vs.* our MetaTeacher from 91.98% to 92.84% (0.86%). This clearly shows that our performance gain is more significant than those by prior art methods.

Table 7: For a sample labeled as Atelectasis and Effusion classes, the weight changes before and after training.

	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edame	Pneumonia
Predictions for each teacher (pre-train)	0.476732	0.267061	0.765137	0.190515	0.301033	0.168340
	0.243154	0.079328	0.481989	0.034837	0.400948	0.087144
	0.346073	0.345510	0.626205	0.021017	0.061996	0.042028
Coordinating weight (pre-train)	**0.214039**	0.149377	0.371503	0.733469	0.404418	0.445663
	0.022341	0.458379	0.377687	0.042648	0.404455	0.017554
	0.763619	0.392244	0.250810	0.223883	0.191127	0.536783
Joint prediction (pre-train)	**0.371739**	0.211779	0.623350	0.145928	0.295758	0.099113
Predictions for each teacher (after-train)	**0.770673**	0.327727	0.776080	0.161243	0.437272	0.377679
	0.430125	0.154535	0.124078	0.069540	0.047292	0.043537
	0.554779	0.255637	0.631193	0.052395	0.122772	0.199748
Coordinating weight (after-train)	**0.950953**	0.280765	0.984358	0.019614	0.000677	0.025918
	0.032633	0.704228	0.011930	0.196210	0.995239	0.943713
	0.016413	0.015007	0.003711	0.784177	0.004084	0.030369
Joint prediction (pre-train)	**0.756016**	0.204678	0.767763	0.057894	0.047864	0.056941

Table 8: Two-teacher and three-teacher transfer scenarios for the target domain *Google-Health-CXR*. Metric: AUROC.

Method	DECISION [1]	MetaTeacher
CheXpert, MIMIC-CXR to Google-Health-CXR	81.16	81.34
NIH-CXR14, CheXpert, MIMIC-CXR to Google-Health-CXR	81.85	82.69

For further validation, we have added a two-teacher transfer experiment: *CheXpert, MIMIC-CXR* to *Google-Health-CXR*. From the two-teacher case to the three-teacher transfer scenario (*NIH-CXR14, CheXpert, MIMIC-CXR* to it Google-Health-CXR), the performance gain of DECISION is 0.69%, vs. 1.35% by our MetaTeacher (see Table 8). This suggests that our method is superior at leveraging the diversity and complementary effect of multiple teacher models.

C.4 Comparison with One-Teacher and One-Student Framework

Assuming no data privacy issue as discussed above, we experiment with a one-teacher one-student design where the component of adaptively training the teachers goes away naturally. We obtain the results of 89.97%/79.94%/75.38%/90.13%, inferior to 92.84%/82.69%/77.74%/91.98% by our MetaTeacher (corresponding to Tables 1,2,3,4 in the main paper). This is due to that each dataset presents unique characteristics in category imbalance and labeling error (e.g., false negatives), resulting in different per-category qualities. Aggregating such datasets into one would introduce negative interference. Besides, using multiple teachers would reduce the learning difficulties of the entire classification problem in a spirit of divide-and-conquer principle, in addition to the opportunity of modeling the confidence per teacher. It should be noted that our multi-teacher setup is underpinned by the nature of our problem where data privacy protection is fundamentally critical (i.e., data sharing across hospitals is typically banned). That being said, a specific teacher model would be trained using the training data of each individual hospital. This leads to the result of multiple teacher models in practice.

C.5 Comparison with Semi-supervised Methods

Compared to the existing semi-supervised methods, our method requires less labeled data. For example, in the transfer scenario *CheXpert, MIMIC- CXR* to *NIH-CXR14*, for achieving 77.74% classification rate, the existing semi-supervised methods [2, 36, 37, 39] require about 20,000 labeled samples (20% of the total), *vs.* our method only needing 500 labeled samples. Hence, our MetaTeacher is more data efficient and favored in practice.