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

Anonymous Author(s)
Affiliation
Address
email

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.

1 Introduction

Despite great stride made by existing deep learning methods on medical image classification results [24, 41, 51], their performance often degrade drastically when applied to a new 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 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 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.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In medical image classification, there are limited domain adaptation works, with a need of accessing the source domain data [3, 14, 19, 25, 33, 42, 43, 47]. 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 [44] or use ensemble methods combined with source classifiers [5]. 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 [34]. But it is still not source-free. Indeed, there have been extensive study on Source-Free Domain Adaptation (SFDA) [26, 48]. However, they are not directly applicable to our problem. Firstly, most of them assume a single source domain [26, 48]. Using the 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 [13, 30, 49, 52, 53]. 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 [20, 21, 38, 39] have shown that a few labeled data from the target domain can significantly improve the performance of the model. Inspired by meta-learning approaches [28, 35, 37], 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, we propose a novel framework termed as MetaTeacher. It is based on multi-teacher and one-student model. 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 student and other teachers in a semi-supervised meta learning manner [12, 35]. Unlike the previous MSDA approaches, MetaTeacher can adapt each teacher in different directions according to the learned coordinating weight. This enable 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 exploration of 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-detail ablation studies for verifying the design choices of our model components.

2 Relate Works

Unsupervised domain adaptation for medical image classification. There are shallow UDA and deep UDA approaches respectively. Shallow UDA approach adapts two routes, i.e., source domain instance weighting [42, 43] and feature transformation [19, 25]. 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 [14, 47] and pseudo-labeling based [3]. 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. [14] 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 [34]. Bermúdez Chacón et al. [3] 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.

Multi-source domain adaptation for medical image classification. In machine learning community, MSDA works mainly have two strategies, i.e. distribution alignment [31, 56] and adversarial learning [46, 54, 55]. 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. [44] proposed to map multiple source and target data to a common latent space for autism spectrum disorder classification. Cheng et al. [5] 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 extending it to the multi-source domain, it will face a challenging multi-objective optimization problem [7, 29].

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 [23, 32, 50]. 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 [32, 50], entropy minimization [15, 38], label smoothing [9, 36] and active learning [36, 40]. 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.

3 The Proposed Method

Problem statements. 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 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 source classifiers f_{T_i} for $i = 1, \cdots, n$ and the target data D_T, the task is to find a mapping $f_S : X_U^t \to Y_U^t$ where Y_U^t denotes the predicted labels for target domain sample X_U^t that can work well in the target domain.

Overview. As shown in Fig.1, our framework is based on 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 based on Resnet50 [16] and a multi-label classifier. The classifier consists of a fully connected layer,

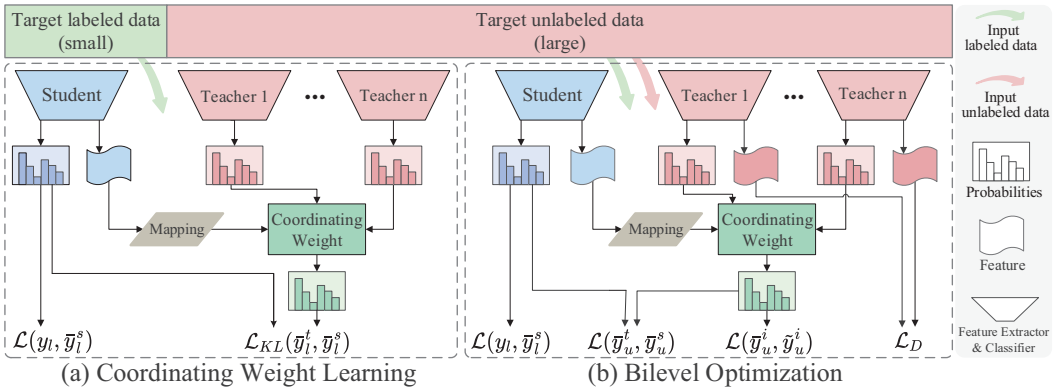

Figure 1: Overview of MetaTeacher. (a) learns coordinating weight mapping which will be used later to provide guidance for updating teacher model. (b) alternately updates the teacher and student models. Each teacher is updated with feedback signals from student and other teachers.

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 model, our method has two differences: (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 another part. In the initial iteration, the mapping and student model are trained based on labeled target samples. In subsequent iterations, this part will only optimize the mapping while the student model will be updated based on bilevel optimization. Then, in the bilevel optimization part, the student and teacher models are updated alternately based on meta-learning. 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 have different characteristics. Therefore, for a target domain sample, the classification probability of each teacher model is 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 it 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 $B \in R^{c \times h \times w}$. c, h, and w are 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 can get 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 $\mathcal{L}\left(\bar{y}_l^s, y_l\right) = \frac{1}{m}\sum_{i=1}^{m}[y_{l,i} log(\bar{y}_{l,i}^s) + (1 - y_{l,i}) log(1 - \bar{y}_{l,i}^s)]$ represent the BCE (Binary Cross Entropy) loss, y_l is the groud 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 generates coordinating weight by Eq.(1), which 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 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 [4, 6] was first proposed in the field of game theory. It includes an upper-level optimization task and a lower-level optimization task, where upper-level optimization task contains lower-level optimization task 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 pesudo-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 $\bar{y}_u^s = f_T(x_u^t; \theta_S)$, 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. So we get the objective function in lower-level optimization task as the following,

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

Obviously, Eq.(5) cannot be optimized simply by gradient descent method, because the teacher models parameters can not be updated until θ_S reaches the optimum. We refer to the idea of meta-learning [12, 27, 35] and make 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. Substitute Eq. (6) into Eq. (5) to 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,

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

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

for $i = 1, \cdots, n$, where θ_S' and θ_{T_i}' are the updated parameters corresponding to the student and teachers respectively. $\bar{y}_u^i = f_{T_i}\left(x_u^t; \theta_{T_i}\right) \cdot W_u^i$ and W_u^i is the ith-row coordinating weight vector of W_u respect to the i-th teacher. $\tilde{y}_u^i$ is the pseudo labels after normalizing the values of $\bar{y}_u^i$ to 0 or 1, i.e., $\tilde{y}_{u,j}^i = 0$ when $\bar{y}_{u,j}^i < 0.5$ and $\tilde{y}_{u,j}^i = 1$ for 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. So, 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 feedback signals from different teachers; (2) student network parameters provide 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.

4 Experiments

Datasets. Five publicly available chest x-ray datasets are used to construct our multi-domain adaptation scenarios. *NIH-CXR14* [45] is a large public dataset of chest x-ray, which contains 108,948 front view x-ray images of 32,717 patients collected from NIH Clinical Center, with a total of 14 disease labels. *MIMIC-CXR* [18] contains 377,110 images and text reports, corresponding to 227,835 radiological studies conducted by Beth Israel Deaconess Medical Center in Boston, Massachusetts. *CheXpert* [17] 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* [8] 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* [2] 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

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 [10]	82.45	92.16	95.12	95.92	89.89	90.37	90.99
SHOT-best [26]	81.48	91.22	94.19	95.10	88.96	89.58	90.09
MME [38]	82.44	90.82	95.46	96.07	90.26	87.20	90.38
ECACL [22]	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
Ours(w/o *mapping*)	79.99	**92.64**	**98.22**	93.64	**95.50**	84.54	90.76
Ours(w/o *update*)	81.98	90.72	95.76	95.51	89.40	82.53	89.32
Ours(*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 [10]	76.90	81.82	87.55	79.62	85.10	82.72	82.29
SHOT-best [26]	75.43	80.28	86.63	77.88	82.37	81.22	80.64
MME [38]	77.34	**84.93**	86.17	78.65	85.33	71.28	80.62
ECACL [22]	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
Ours(w/o *mapping*)	75.62	83.91	85.40	**80.27**	75.13	81.77	80.35
Ours(w/o *update*)	76.75	84.30	86.67	78.59	82.31	65.84	79.08
Ours(all)	**77.65**	79.52	**88.73**	78.74	**86.73**	**84.78**	**82.69**

the setting of multi-label medical image classification problems, the evaluation criterion is Area Under the Receiver Operating Characteristic (AUROC) [11] curve score.

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 [10], and SHOT-best. The SHOT-best refers to adapting each source domain separately through the SHOT [26] method. The model with the best performance on the validation set is selected. The final category is semi-supervised domain adaptation methods, which are MME [38] and ECACL [22]. 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. Ours(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

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 [10]	72.64	81.12	80.25	74.73	81.02	70.44	76.70
SHOT-best [26]	70.79	79.62	79.24	72.25	80.79	69.65	75.39
MME [38]	72.90	81.73	81.01	73.11	81.03	71.52	76.88
ECACL [22]	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
Ours(w/o *mapping*)	72.05	81.58	78.36	72.94	82.19	69.82	76.16
Ours(w/o *update*)	72.24	80.69	79.56	69.80	78.13	70.55	75.16
Ours(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 [10]	82.38	91.97	94.89	95.30	89.81	90.44	90.80
SHOT-best [26]	81.48	91.22	94.19	95.10	88.96	89.58	90.09
MME [38]	81.46	90.40	94.86	**97.73**	89.79	87.31	90.26
ECACL [22]	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
Ours(w/o *mapping*)	**83.73**	**93.37**	96.04	97.30	91.51	82.34	90.72
Ours(w/o *update*)	82.70	88.91	95.47	95.48	88.96	78.85	88.40
Ours(all)	82.11	92.42	**96.80**	97.07	**92.20**	**91.27**	**91.98**

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*, compared with the first scenario, *MIMIC-CXR* is removed. As show in Table 4, the effect of two source domains is 0.86% lower than that of three source domains.

4.2 Ablation Analysis and Discussion

Component analysis. In Tables 1-4, Ours(w/o *mapping*) represents that our proposed method removes the part of coordinating weight learning and optimization substituted by average. Ours(w/o *update*) means to remove the bilevel optimization process. In this situation, updating teachers and student follows the knowledge distillation learning process, i.e, using the weighted output of teachers 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 the bilevel optimization part contributes more gain to the overall performance than the coordinating weight learning. This is because the bilevel optimization can greatly improve the predictive ability of some diseases by updating the teachers, such as Edema in Table 1 and Pneumonia in Table 2, which are limited by distillation methods. 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.

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

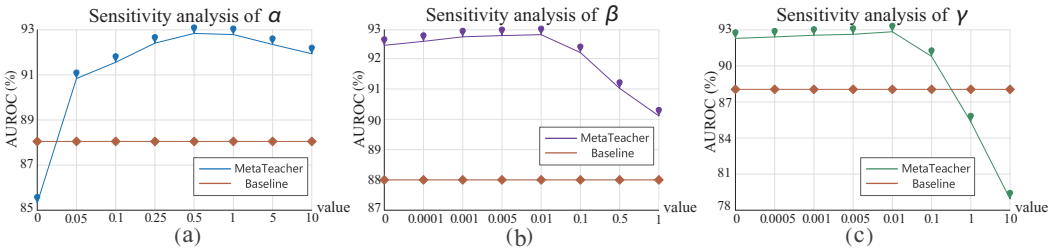

Figure 2: Effect of different hyperparameters on the transfer from *NIH-CXR14, CheXpert, MIMIC-CXR* to *Open-i*. Baseline: source only(M).

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.

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. Fig.(2)(c) shows the influences of the parameters γ. When γ is set to 0.01, the performance reaches the best, but with the continuous increase of γ, the performance decreases obviously. It is worth to note that our performance is about 0.55% better than the case not adding the divergence loss, which shows its rationality. 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 proposed a new framework, termed as MetaTeacher, for semi-supervised multi-source-free domain adaptation of medical image classification. The transfer learning process is modeled as a multi-teacher and one-student scheme. We not only optimize student, but also optimize teachers through student's feedback in the target domain. The optimization is based on meta-learning, which consists of two main part: coordinating weight learning, and bilevel optimization. The first part obtains the coordinating weight mapping which is used to coordinate the teacher outputs and updates. Bilevel optimization updates the student base on the pseudo-labeled data produced by teachers and updates each teacher base on the feedback signal generated by 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.

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] 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.

[3] 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.

[4] BRACKEN, J., AND MCGILL, J. T. Mathematical programs with optimization problems in the constraints. *Operations Research 21*, 1 (1973), 37–44.

[5] 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.

[6] COLSON, B., MARCOTTE, P., AND SAVARD, G. An overview of bilevel optimization. *Annals of operations research 153*, 1 (2007), 235–256.

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

[8] 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.

[9] 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.

[10] 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).

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

[12] 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.

[13] 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.

[14] 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.

[15] GRANDVALET, Y., AND BENGIO, Y. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems 17* (2004).

[16] 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.

[17] 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.

[18] 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).

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] LIU, H., SIMONYAN, K., AND YANG, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).

[28] 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.

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

[30] 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.

[31] 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.

[32] PEREIRA, L. A., AND DA SILVA TORRES, R. Semi-supervised transfer subspace for domain adaptation. *Pattern Recognition 75* (2018), 235–249.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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).

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] WACHINGER, C., REUTER, M., INITIATIVE, A. D. N., ET AL. Domain adaptation for alzheimer's disease diagnostics. *Neuroimage 139* (2016), 470–479.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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).

[52] 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).

[53] ZHAO, H., SUN, X., DONG, J., CHEN, C., AND DONG, Z. Highlight every step: Knowledge distillation via collaborative teaching. *IEEE Transactions on Cybernetics* (2020).

[54] 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).

[55] 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.

[56] 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]