Debiased Curriculum Adaptation for Safe Transfer Learning in Chest X-ray Classification

Mingyang Liu[1], Xinyang Chen[1]✉, Yang Shu[2]✉, Xiucheng Li[1], Weili Guan[3], Liqiang Nie[1]
[1]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[2]School of Data Science and Engineering, East China Normal University
[3]School of Information Science and Technology, Harbin Institute of Technology (Shenzhen)
{mingyangliu1024, chenxinyang95, nieliqiang}@gmail.com
yshu@dase.ecnu.edu.cn, {lixiucheng,guanweili}@hit.edu.cn

Abstract

Chest X-ray classification is extensively utilized within the field of medical image analysis. However, manually labeling chest X-ray images is time-consuming and costly. Domain adaptation, which is designed to transfer knowledge from related domains, could offer a promising solution. Existing methods employ feature adaptation or self-training for knowledge transfer. Nonetheless, negative transfer is observed due to the entanglement of class imbalance and distribution shift in chest X-ray classification. In this paper, we propose Debiased Curriculum Adaptation framework to mitigate negative transfer in two aspects: (1) Curriculum Adaptation, which is designed to transfer knowledge in an easy-to-hard way, is proposed to alleviate confirmation bias in self-training. (2) Spectral Debiasing is introduced to harmonize the feature space between the source and target domains, as well as balance the feature space of positive and negative samples. Extensive experiments on 72 transfer tasks (including 6 diseases and 4 domains) demonstrate our superiority over state-of-the-art methods. In comparison to advanced methods, our approach effectively mitigates negative transfer, ensuring safe knowledge transfer.

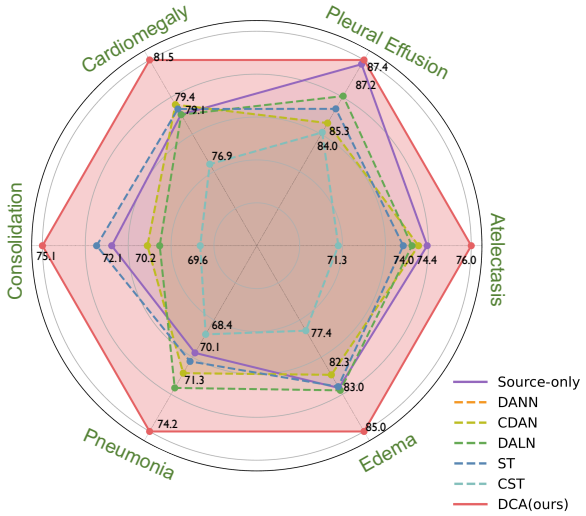


Figure 1. Performance comparison with Source-only method (without knowledge transfer). The label shows 6 common chest diseases. DANN, CDAN and DALN are feature adaptation methods, while ST and CST are self-training methods.

1. Introduction

Chest X-ray classification is crucial for diagnosing diseases in healthcare and medical systems [10]. In recent years, the advent of deep learning has led to the application of deep models in chest X-ray classification. Nevertheless, the persistent challenges remain in the scarcity of labeled data, due to the high cost of human annotation. Even though developing diagnostic models by fine-tuning limited labeled X-ray images is effective, when deploying models in real clinical scenarios, performance drops happen. This is attributed to the distribution difference between the development scenario (source domain) and the deployment scenario (target domain). Unsupervised domain adaptation (**UDA**), leveraging knowledge transfer from a related source domain to the unlabeled target domain, can be a promising solution.

The mainstream paradigms for UDA are feature adaptation and self-training. Feature adaptation methods aim to reduce the distance between the source and target feature distributions, learning domain-invariant representations to facilitate knowledge transfer across domains [8, 16, 33, 37, 49]. Self-training methods generate pseudo-labels for unlabeled data and jointly train the model using both source labels and target pseudo-labels [27, 31]. These two paradigms are effective across various scenarios in general domain

adaptation benchmarks [21, 41].

However, different from general domain adaptation benchmarks, chest X-ray datasets are naturally imbalanced. The class imbalance issue, entangled with the distribution shift, introduces new challenges for transfer learning in chest X-ray classification. For feature adaptation methods, impossibility theories [2] have revealed intrinsic limitations in learning invariant representations, especially in the presence of label shift, which is frequently encountered in different chest X-ray datasets, often characterized by varying degrees of class imbalance. For self-training methods, deep models become biased toward the majority class [26], producing unreliable pseudo-labels. Incorporating these pseudo-labels into training will lead the model into a vicious cycle, known as confirmation bias [1]. As shown in Fig. 1, with the Source-only method (training the model exclusively on source data) as a baseline, existing feature adaptation and self-training methods underperform, *i.e.* introducing UDA algorithms undesirably hinders models in the target domain, which is termed as negative transfer [43].

To mitigate negative transfer and achieve safe transfer learning, we opt to abandon the feature adaptation paradigm and instead embrace the self-training paradigm with the inspiration of impossibility theories [2]. To alleviate confirmation bias in self-training and calibrate the model, we propose Curriculum Adaptation, which achieves knowledge transfer in an easy-to-hard way. During this process, trustworthy pseudo labels are selected by mixing cross-domain samples. To further analyze the influence of class imbalance and distribution shift, we employ singular value decomposition (SVD) to analyze the properties of representation. Representation bias is observed: feature space of majority class is larger than feature space of minority class, and source feature space is larger than target feature space. To this end, we introduce Spectral Debiasing (SD), comprising Class-wise Spectral Debiasing (CSD) and Domain-wise Spectral Debiasing (DSD), as a solution to mitigate representation bias. The application of Debiased Curriculum Adaptation (DCA), integrating the aforementioned learning modules, achieves state-of-the-art performance across 72 transfer tasks and successfully mitigates negative transfer compared to advanced methods, achieving safe knowledge transfer. Our contributions are summarized as follows:

- We propose Curriculum Adaptation, which transfers knowledge in an easy-to-hard way and reduces the risk of learning incorrect pseudo-labels by incorporating a mixing strategy with cross-domain samples, to alleviate confirmation bias and boost safe knowledge transfer.
- We propose Spectral Debiasing, which balances the feature space of positive and negative samples as well as harmonizes the feature space between the source and target domains, to alleviate representation bias between classes and across domains.
- Debiased Curriculum Adaptation, consisting of the above two learning modules, yields state-of-the-art performance on 72 transfer tasks. Notably, our approach excels in mitigating negative transfer in comparison to advanced methods, ensuring safe knowledge transfer.

2. Related Work

Unsupervised domain adaptation leverages the labeled source domain to predict the labels of a different but related, unlabeled target domain. Mainstream methods can be divided into two categories: (1) Feature adaptation focuses on transferring knowledge by reducing the distance between the source and target feature distributions. Widely used methods include DANN [16], CDAN [33], MCD [37], MDD [49] and DALN [8]. However, impossibility theories [2] have revealed intrinsic limitations in learning invariant representations, especially in the presence of label shift. (2) Self-training, *i.e.* pseudo-labeling, produces pseudo-labels on unlabeled data and uses confident pseudo-labels together with the labeled data to train the model [29, 32, 50]. Many efforts are devoted to boosting performance by utilizing selective pseudo-labeling [7], a teacher-student network [22, 29], voting scheme [13, 45] or clustering the unlabeled target data [46]. However, the pseudo-labels can be unreliable even with expensive threshold tweaking [30, 31], leading to confirmation bias and negative transfer.

UDA for medical image analysis can be categorized into shallow methods and deep methods. The shallow UDA approach typically adapts two strategies, *i.e.* instance weighting [9, 20] and feature transformation [25, 28]. Similarly, two main strategies are commonly used in the deep UDA approach: feature adaptation methods [17, 39, 48] and self-training methods [5, 44]. However, class imbalance is highly skewed in medical image datasets, which can lead to negative transfer of existing classic UDA methods.

Negative transfer, which is first defined as transferring knowledge from the source can have a negative impact on the target learner [34], has been explored in many transfer learning scenarios. In this paper, we use the Source-only method, *i.e.* the model trained only on the source data and with no access to target data, as the baseline. We define negative transfer as introducing UDA algorithms undesirably leads to a performance drop compared with the Source-only method. In chest X-ray classification, due to the entanglement of class imbalance and distribution shift, negative transfer is frequently encountered. We reveal two types of biases that lead to negative transfer in chest X-ray classification: confirmation bias and representation bias.

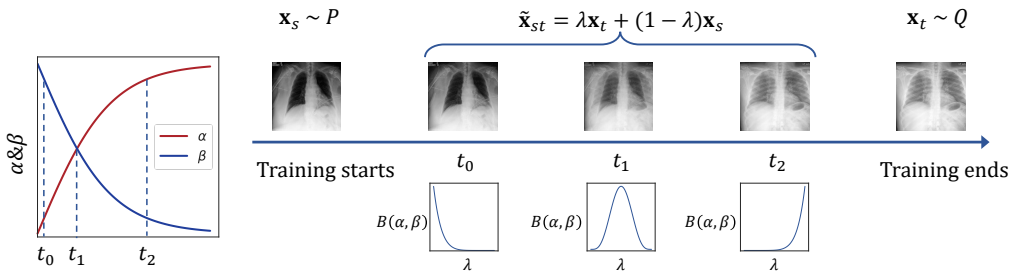


Figure 2. The variation process of α_p, β_p, λ and $\tilde{\mathbf{x}}^{st}$. In the early stage of adaptation, *e.g.* t_0, $\alpha_p < \beta_p$, the mean of $\lambda \sim B(\alpha_p, \beta_p)$ is close to 0. As the adaptation progresses, α_p gradually increases while β_p decreases, making the mean of λ close to 0.5, *e.g.* t_1. In the later stage, *e.g.* t_2, $\alpha_p > \beta_p$, making the mean of λ close to 1.

3. Approach

3.1. Problem Setup

The standard chest X-ray classification problem can be formulated as the problem of supervised learning. The model receives a set of n labeled points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is an output space. This work focuses on binary classification, where $\mathcal{Y}$ is $\{0, 1\}$. We define samples with labels $y = 1$ as positive samples with diseases and $y = 0$ as negative samples or healthy samples. In unsupervised domain adaptation (UDA), there are two domains with different distributions: the source P and the target Q. The model is trained on a labeled sample set $\hat{P} = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^n$ drawn from the source distribution, and an unlabeled set $\hat{Q} = \{\mathbf{x}_i^t\}_{i=1}^m$ drawn from the target. The goal of UDA is to learn a classification model with the labeled source data $\hat{P}$ and the unlabeled target data $\hat{Q}$, which can make accurate predictions on the target domain.

3.2. Curriculum Adaptation Framework

Adaptation with Self-training The deep learning model for chest X-ray classification constitutes a feature extractor and a category classifier, where the feature extractor $F : \mathcal{X} \to \mathbb{R}^d$ outputs the feature representation of the input data $\mathbf{f} = F(\mathbf{x})$ and the classifier $G : \mathbb{R}^d \to \mathbb{R}$ outputs a score $g = G(\mathbf{f})$ representing the confidence that the given data sample is a positive one (i.e. the sample with certain diseases). The fundamental loss function used in this case is the Binary Cross-Entropy (BCE) loss, denoted by ℓ. Given labeled data in the source, BCE loss can be used as the objective function to optimize the model:

$$\min_{F,G} \mathcal{L}_s = \mathbb{E}_{(\mathbf{x}_i^s, y_i^s) \sim \hat{P}} [\ell \left(G \left(F \left(\mathbf{x}_i^s \right) \right), y_i^s \right)]. \tag{1}$$

To further enhance knowledge transfer from source data and adaptation on target data, we leverage self-training on the target data together with source supervised training. With the model being trained on the source, pseudo-labels can be made on the target data as $\hat{y}^{t*} = \mathbf{1} \left(G \left(F \left(\mathbf{x}^t \right) \right) > 0.5 \right)$. As outputs with higher confidence imply more trustworthy predictions, we continually select target samples with high confidence ($G(F(\mathbf{x}^t)) = \hat{y}^t > \delta$) into the self-training process:

$$\min_{F,G} \mathcal{L}_t = \mathbb{E}_{(\mathbf{x}_i^t) \sim \hat{Q}} \mathbf{1}(\hat{y}_i^t > \delta) \cdot [\ell \left(G \left(F \left(\mathbf{x}_i^t \right) \right), \hat{y}_i^{t*} \right)]. \tag{2}$$

Unreliable pseudo-labels caused by confirmation bias may raise the risk of negative transfer [40] in the above self-training paradigm. Let $\epsilon_P(h) = \mathbb{E}_{(\mathbf{x},y) \sim P}[h(\mathbf{x}) \neq y]$ denote the risk of the hypothesis h w.r.t. the certain distribution P, $h^* = \arg\min_h \epsilon_P(h) + \epsilon_Q(h)$ be the ideal hypothesis that embodies the notion of adaptability, $\epsilon_P(h_1, h_2) = \mathbb{E}_{(\mathbf{x},y) \sim P}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$ be the disagreement between hypotheses. As proposed by Ben-David *et al.* [3], the target risk ϵ_Q of hypothesis h is bounded by the source risk ϵ_P plus the distribution discrepancy:

$$\begin{aligned} \epsilon_Q(h) \leq & \epsilon_P(h) + [\epsilon_P(h^*) + \epsilon_Q(h^*)] \\ & + |\epsilon_P(h, h^*) - \epsilon_Q(h, h^*)|. \end{aligned} \tag{3}$$

This indicates that a larger distribution discrepancy may induce less reliable predictions on the target data, improving the difficulty of adaptation and the risk of negative transfer.

Safe Transfer with Curriculum Adaptation To facilitate safe knowledge transfer, inspired by the idea of curriculum learning [4, 12, 38], we propose an easy-to-hard way to transfer from source to target with curriculum adaptation. The general idea is to generate and adapt to intermediate distributions with a gradual transition from the source domain to the target domain, forming the curriculum from smaller discrepancies to larger discrepancies. We adapt Mixup [47] to realize this idea in chest X-ray classification. Given a sample pair $(\mathbf{x}^s, y^s)$ and $(\mathbf{x}^t, \hat{y}^{t*})$, where $(\mathbf{x}^s, y^s)$ refers to a source sample with its annotated label and $(\mathbf{x}^t, \hat{y}^{t*})$ refers to a target sample with its pseudo-label provided by the model, we generate the mixed sample $(\tilde{\mathbf{x}}^{st}, \tilde{y}^{st})$ from the mixup of the sample pair, i.e. the convex combinations of the source and target samples:

$$\begin{aligned} \tilde{\mathbf{x}}^{st} &= \lambda \mathbf{x}^t + (1-\lambda)\mathbf{x}^s, \\ \tilde{y}^{st} &= \lambda \hat{y}^{t*} + (1-\lambda) y^s, \end{aligned} \tag{4}$$

where $\lambda \in [0,1]$ is randomly sampled from a Beta distribution $B(\alpha_p, \beta_p)$ and controls mixing degree. The generated samples from the mixup process form the sample set $\hat{I} = \{(\widetilde{\mathbf{x}}_i^{st}, \widetilde{y}_i^{st})\}_{i=1}^{k}$, which can be regarded as drawn from a dynamically changing intermediate distribution I. Inspired by the risk bound in Eq. (3), the design of the curriculum encourages adaptation on distributions with smaller discrepancies from the source domain first (i.e., with a smaller λ in statistics), which gradually transition to the target domain (with a larger λ). Thus, as the curriculum adaptation goes on, it dynamically increases α_p and decreases β_p with α_p-β_p schedule:

$$\begin{aligned} &\alpha_p = \beta_0 t(p), \beta_p = \beta_0 \left(1 - t\left(p\right)\right), \\ &\mathbb{E}_{\lambda \sim B(\alpha_p, \beta_p)} = \frac{\alpha_p}{\alpha_p + \beta_p} = t(p), \end{aligned} \tag{5}$$

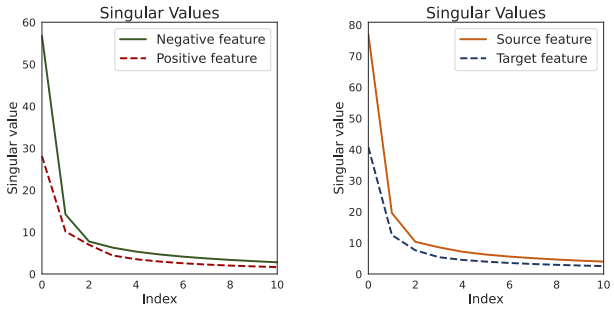
where $p \in [0,1]$ is the training progress, and $t(p)$ can be any monotonic function that varies from 0 to 1. We suggest setting $\beta_0 > 1$ to make the variation of $B(\alpha_p, \beta_p)$ intuitive. The α_p-β_p schedule ensures that the mean of λ continues to increase during the adaptation process. Fig. 2 illustrates the variation process of α_p, β_p, λ and $\widetilde{\mathbf{x}}^{st}$.

Contrary to the original self-training, in curriculum adaptation, the unlabeled target samples are only used to generate mixed samples and are not directly involved in the backpropagation of training. Instead, we use the generated mixed samples from $\hat{I}$ and the source labeled samples from $\hat{P}$ to train the classifier. Thus, the self-training loss $\mathcal{L}_t$ in Eq. (2) is replaced by $\mathcal{L}_{st}$ on the mixed samples:

$$\mathcal{L}_{st} = \mathbb{E}_{(\widetilde{\mathbf{x}}_i^{st}, \widetilde{y}_i^{st}) \sim \hat{I}} [\ell(G(F(\widetilde{\mathbf{x}}_i^{st})), \widetilde{y}_i^{st})]. \tag{6}$$

In the early stage of adaptation, the generated mixed samples are dominated by the source samples. Therefore, even if the target pseudo-labels contain significant noise, they do not mislead the self-training. As the training progresses, the model gradually adapts to the intermediate distribution $I = (1-\lambda)P + \lambda Q$, which is closer to the target distribution Q than the source distribution P. According to the expected error bound on the target samples in Eq. (3), the pseudo-labels on the target domain will be more reliable. Therefore, we gradually empower the target samples to take the lead in mixing, fully leveraging the implicit label information in the target domain.

Remark on Confidence Calibration. Medical diagnosis models are sensitive to prediction confidence, where incorrect predictions with high confidence pose significant clinical risks, particularly under severe class imbalances between positive and negative cases. To mitigate the issue, implicit confidence calibration is incorporated into Curriculum Adaptation Framework with the supervision of uncertainty-aware cross-class samples. Concretely, the mixup between positive and negative samples generates cross-class samples with soft labels, providing fine-grained class and confidence information for chest X-ray classification.



(a) Positive vs. Negative (b) Source vs. Target

Figure 3. Singular values of the feature matrix. (a) Features trained on NIH-CXR. (b) Features trained on the transfer task from NIH-CXR to MIMIC-CXR for Consolidation.

3.3. Spectral Debiasing

Apart from the challenge of confirmation bias on target data adaptation, representation bias is another crucial issue in UDA chest X-ray classification, which also leads to negative transfer during the adaptation process. Tab. 1 shows the incidence r of six diseases on four different datasets. From the perspective of classes, samples with specific diseases (positive samples) are rarer and harder to collect than those without diseases (negative samples) in practice, which can be shown from the low incidence of diseases in these datasets. From the perspective of domains, there exists domain shifts between source and target domains, which consist of both covariate shifts (different statistics of input data) and label shifts (different incidences of diseases across domains and datasets). These two issues may cause class bias and domain bias in the learned representations, making it insufficient to extract disease-relevant and target-specific features. In order to enhance safe knowledge transfer, we further propose debiased learning with spectral regularization to overcome such representation biases.

Table 1. Incidence r of Atelectasis, Pneumonia, Cardiomegaly, Consolidation, Pleural Effusion and Edema in four chest X-ray datasets.More information about these datasets is in Sec. 4.1.

Disease	CheXpert	MIMIC-CXR	NIH-CXR	Open-i
Atelectasis	30.8%	24.3%	10.4%	7.0%
Pneumonia	10.7%	15.0%	1.2%	0.9%
Cardiomegaly	15.9%	22.3%	2.5%	7.2%
Consolidation	19.9%	6.9%	4.2%	0.7%
Pleural Effusion	46.0%	26.2%	11.2%	11.2%
Edema	32.0%	17.5%	2.1%	0.7%

Class-wise Spectral Debiasing. Neural collapse [35] describes the geometric properties of the learned embeddings (i.e. the penultimate layer features) in deep neural networks, where the embeddings of each class tend to concentrate around their class means. Fang *et al.* [14] discovered the phenomenon of minority collapse, which reveals

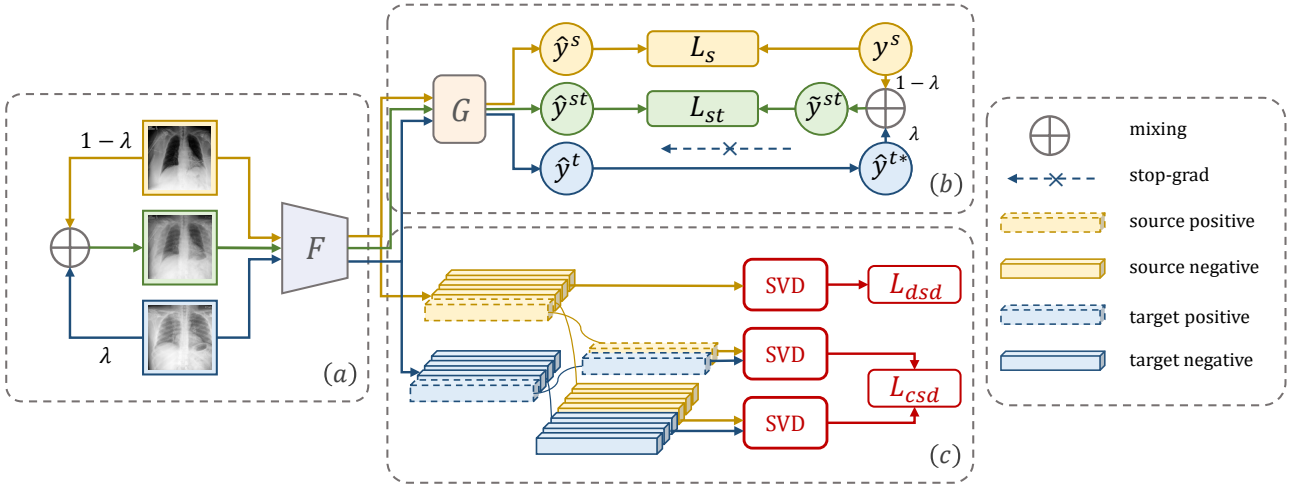


Figure 4. The architecture of the proposed debiased curriculum adaptation framework. (a) Source (yellow) and target (blue) data sampling, mixup (green) data generation; (b) Curriculum adaptation with source data and the dynamically generated mixup data; (c) Spectral debiasing on the split source feature matrix (above) and positive-negative feature matrix (below).

the inconsistency between the majority class and the minority class in terms of collapse in the presence of class imbalance. To be specific, as the level of imbalance increases, the majority classes become the predominant part of the risk function. As a result, the training process pays too much emphasis on the majority classification, encouraging the feature spaces of majority classes to grow and meanwhile shrinking those of the minority ones. Inspired by this, we propose class-wise debiased learning by regularizing the feature spaces of different classes.

Consider an $n \times n$ matrix $\mathbf{A}$, the matrix can be written in terms of its column vectors $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$. From the geometric view, the determinant of the matrix, $\det(\mathbf{A})$ gives the n-dimensional volume of the parallelotope defined by its column vectors. A larger determinant $\det(\mathbf{A})$ indicates that the column vectors span a larger volume in the n-dimensional space. Let $\sigma_1, \sigma_2, \cdots, \sigma_n$ denote the eigenvalues of the matrix $\mathbf{A}$, we further have $\det(\mathbf{A}) = \prod_{i=1}^{n} \sigma_i$. Thus, the volume spanned by a set of vectors is controlled by the eigenvalues of the matrix. In the batch-training practice, we consider the feature matrices of positive samples $\mathbf{F}^p = [\mathbf{f}_1^p, \ldots, \mathbf{f}_{b_p}^p]$ and negative samples $\mathbf{F}^n = [\mathbf{f}_1^n, \ldots, \mathbf{f}_{b_n}^n]$ partitioned from the training batch, and propose to regularize their singular values to balance positive and negative feature spaces, and thus to overcome minority collapse caused by class bias.

For the positive and negative feature matrix $\mathbf{F}^p$ and $\mathbf{F}^n$ in mini-batches, we apply singular value decomposition (SVD) [18] to compute their singular values:

$$\begin{aligned} \mathbf{F}^p &= \mathbf{U}^p \mathbf{\Sigma}^p (\mathbf{V}^p)^\top, \\ \mathbf{F}^n &= \mathbf{U}^n \mathbf{\Sigma}^n (\mathbf{V}^n)^\top. \end{aligned} \tag{7}$$

Fig. 3a shows the singular values of the positive and negative feature matrix respectively. We observe that the negative feature matrix generally has larger singular values, which is consistent with the phenomenon of minority collapse. Furthermore, the distribution of singular values is sharp, with the top singular values much larger than others. Based on this, to grasp the key point in feature space regularization, we propose to regularize the largest singular values of the class-wise feature matrix, i.e., their spectral norms. Let $\sigma_{\max}^p$ and $\sigma_{\max}^n$ denote the spectral norms of the positive and negative feature matrix $\mathbf{F}^p$ and $\mathbf{F}^n$ respectively, the class-wise spectral debiasing loss to regularize these spectral norms is computed as:

$$\mathcal{L}_{csd} = (\sigma_{\max}^p)^2 + (\sigma_{\max}^n)^2. \tag{8}$$

We regularize both positive and negative features to avoid over-regularization of negative ones. With a larger spectral norm, the regularization term has a larger effect on the negative features, which balances the class-wise features and helps overcome the class bias in the learned representations.

Domain-wise Spectral Debiasing. Similar representation bias happens during the adaptation stage when samples from the source and target domains are trained together due to the distribution shifts between them. Since source domain data are fully annotated with trustworthy labels and trained with supervised learning, while target data are unlabeled and adapted indirectly with curriculum self-training, this causes domain bias of the learned representations towards source distributions. As the source domain may have different statistics of input data and different disease incidences from the target, the resulting covariate shift and label

shift make domain bias a factor of negative transfer, which may prevent the learning of target-specific representations.

We consider the feature matrix of source samples $\mathbf{F}^s = [\mathbf{f}_1^s \dots \mathbf{f}_b^s]$ and those of target samples $\mathbf{F}^t = [\mathbf{f}_1^t \dots \mathbf{f}_b^t]$ in the training batches to evaluate and overcome such domain bias. We compute singular values of the source and target feature matrix in Fig. 3b. A similar representation bias phenomenon is observed, with singular values of source domain features generally larger than those of the target domain, indicating a dominant state of source distributions in the feature space. Let σ_{max}^s denote the spectral norm of the source domain matrix, the proposed domain-wise spectral debiasing loss to harmonize the feature space between the source and target domains is computed as:

$$\mathcal{L}_{dsd} = (\sigma_{\text{max}}^s)^2. \tag{9}$$

Since the goal of domain adaptation is accurate classification on the target domain, we regularize the source features to overcome the domain bias, which helps mitigate negative transfer and enhance adaptation on the target domain.

3.4. Debiased Curriculum Adaptation

We incorporate the proposed curriculum adaptation and spectral debiasing in the debiased curriculum adaptation framework for safe knowledge transfer.

The whole framework is illustrated in Fig. 4. In each iteration, a mini-batch of source and target data is sampled. Target samples with more trustworthy pseudo-labels are selected and participate in the mixup with source domain data to generate data from intermediate distributions according to Eq. (4). Then, the model is trained with source domain data and generated mixup data using the supervised learning loss $\mathcal{L}_s$ in Eq. (1) and the self-training loss $\mathcal{L}_{st}$ in Eq. (6). The mixup process is controlled by a dynamically changing weight distribution, which enables curriculum adaptation from the source to the target domain.

Spectral debiasing is performed at the same stage of curriculum adaptation to further mitigate the representation bias. The extracted feature matrix $\mathbf{F}$ is split from two perspectives. One is to split the matrix into positive samples $\mathbf{F}^p$ and negative samples $\mathbf{F}^n$, where class-wise spectral debiasing $\mathcal{L}_{csd}$ is performed on the singular values as in Eq. (8). Another is to split the matrix into source samples $\mathbf{F}^s$ and target samples $\mathbf{F}^t$, where domain-wise spectral debiasing $\mathcal{L}_{dsd}$ is performed on the singular value as in Eq. (9). The overall optimization problem can be written as:

$$\min_{F,G} \mathcal{L} = \mathcal{L}_s + \mathcal{L}_{st} + \gamma(\mathcal{L}_{csd} + \mathcal{L}_{dsd}), \tag{10}$$

where γ is a hyperparameter to trade off the importance of curriculum adaptation and spectral debiasing. The whole framework integrates the above two learning modules to enable safe transfer learning. The implementation details are included in supplementary material.

4. Experiments

In this section, we evaluate the proposed DCA and compare it with several advanced UDA methods on four publicly available chest X-ray datasets.

4.1. Setup

Datasets We conduct extensive experiments using four public chest X-ray datasets. (1) CheXpert [23] is a public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients. The dataset collected chest X-ray examinations and related radiology reports from Stanford Hospital. (2) MIMIC-CXR [24] is a publicly available dataset of chest radiographs with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. (3) NIH-CXR14 [42] is a large public dataset of chest X-ray, which contains 108,948 frontal X-ray images of 32,717 patients collected from NIH Clinical Center, with a total of 14 disease labels. (4) Open-i [11] is collected by Indiana University Hospital. It contains 3955 radiology reports, corresponding to 7470 frontal and lateral chest films. We establish benchmarks for six diseases (Atelectasis, Pneumonia, Cardiomegaly, Consolidation, Pleural Effusion, and Edema), treating each dataset as a distinct domain. Hence we have 4 domains, which are represented as CheXpert (**C**), MIMIC-CXR (**M**), NIH-CXR14 (**N**) and Open-i (**O**). For each disease, we have 12 transfer tasks.

Baselines We compare the proposed DCA with three categories of advanced methods for comparison. A common approach within the first category involves training the model exclusively on source data, commonly referred to as "Source-only". The second category is the advanced feature adaptation methods, which are **DANN** [16], **MCD** [37], **CDAN** [33], **MDD** [49] and **DALN** [8]. The third category is self-training methods, which are the vanilla Self-Training (**ST**) [27] and Cycle Self-Training (**CST**) [31].

Implementation We use **PyTorch** to implement our methods and fine-tune ResNet-50 pre-trained on ImageNet [36]. We adopt mini-batch SGD with momentum of 0.9 and use the learning rate schedule of DANN [16]. The evaluation criterion is Area Under the Receiver Operating Characteristic (AUROC) [15] curve score. More experimental details are provided in the appendix. Code is available at https://github.com/mingyangliu1024/DCA.

4.2. Results

Due to the limited pages, we report the average AUROC results and results across 4 domains for 3 diseases: Atelectasis, Pneumonia and Cardiomegaly. Each disease has 12

Table 2. Average improvement compared to Source-only method of 12 transfer tasks for each disease.

Method	Atelectasis	Pneumonia	Cardiomegaly	Consolidation	Effusion	Edema	Average
DANN [16]	-0.242%	+1.302%	+0.545%	-1.046%	-3.355%	-0.746%	-0.590%
CDAN [33]	-0.507%	+1.748%	+0.365%	-2.880%	-2.250%	-2.884%	-1.068%
MCD [37]	-2.113%	-3.410%	-3.450%	-1.327%	-3.359%	-6.091%	-3.292%
MDD [49]	+0.886%	+2.163%	+0.089%	+0.766%	-1.108%	-0.166%	+0.438%
DALN [8]	-0.408%	+1.394%	-0.021%	-1.606%	-1.842%	+0.328%	-0.408%
ST [27]	-0.889%	+0.834%	+0.347%	+0.698%	-2.495%	+0.139%	-0.228%
CST [31]	-4.386%	-2.325%	-2.885%	-3.498%	-3.762%	-6.268%	-3.854%
DCA (ours)	**+2.227%**	**+6.504%**	**+3.110%**	**+4.412%**	**+0.234%**	**+2.691%**	**+3.196%**

Table 3. AUROC (%) on 12 transfer tasks across four domains for Atelectasis.

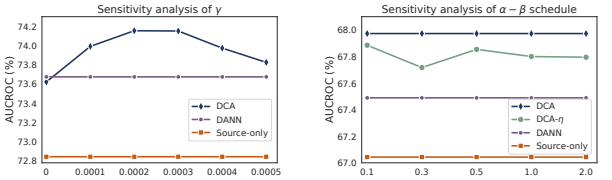
Method	C→M	C→N	C→O	M→C	M→N	M→O	N→C	N→ M	N→O	O→C	O→M	O→N	Average
Source-only	77.846	78.198	82.672	68.530	77.541	86.534	67.043	75.359	86.971	57.196	67.588	67.735	74.434
DANN [16]	77.946	77.681	79.296	67.989	78.008	84.343	66.858	76.566	84.626	58.791	69.876	67.482	74.122
CDAN [33]	77.978	78.080	80.740	68.047	78.073	83.986	66.748	76.168	84.807	59.280	69.728	63.835	73.956
MCD [37]	78.041	76.853	79.832	67.339	74.946	85.489	63.705	72.311	81.177	58.754	67.353	67.156	72.746
MDD [49]	**78.643**	78.873	83.459	69.097	78.653	86.441	67.420	76.882	85.513	59.361	69.121	66.837	75.025
DALN [8]	77.977	76.995	82.054	68.006	77.478	86.073	66.387	74.828	85.468	57.696	67.522	66.176	73.888
ST [27]	77.356	77.339	**85.050**	67.412	77.527	**87.245**	66.067	74.487	83.587	57.957	67.300	64.084	73.784
CST [31]	78.295	66.109	81.625	68.761	76.196	82.720	57.804	**77.204**	86.598	55.252	58.681	66.150	71.283
DCA (ours)	78.208	**79.306**	84.839	**69.216**	**79.240**	87.240	**67.973**	77.123	**86.992**	**60.013**	**70.978**	**70.800**	**75.994**

transfer tasks. More results for Consolidation, Pleural Effusion and Edema are included in the supplementary material.

We calculate average improvement compared to Source-only on 72 transfer tasks for each method respectively, presented in Tab. 2. It can be observed that the average improvements of existing advanced methods are either less than 0% or close to 0%. Hence it is essential to develop safe transfer learning methods tailored for chest X-ray classification, addressing the persistent challenge of highly imbalanced class distribution. In contrast, DCA achieves improvement in a wide range of tasks with an average improvement of 3.196%, which demonstrates our superiority.

Specifically, as shown in Tabs. 3 to 5, DCA outperforms feature adaptation and self-training methods significantly on **9 out of 12** tasks for Atelectasis, **10 out of 12** tasks for Pneumonia and **8 out of 12** tasks for Cardiomegaly. It is worth noting that these feature adaptation and self-training methods commonly encounter negative transfer, while DCA achieves improvement in **all** tasks. From Tab. 1, we observe that Open-i is an extremely imbalanced dataset, which makes transfer tasks from Open-i to other domains highly challenging. On the tasks where Open-i is the source domain, our model achieves an obvious boost with an average improvement of **8.623%**, while the largest improvement of other methods is only **1.361%**.

4.3. Analysis



(a) Effect of different γ (b) Effect of different schedules

Figure 5. Sensitivity analysis: (a) Effect of γ. (b) Effect of different α_p-β_p schedules. DCA-η: Employ the Charbonnier penalty function with different η as $\alpha_p - \beta_p$ schedules.

Sensitivity Analysis We conduct **parameter analysis** on the transfer task from NIH-CXR to MIMIC-CXR for Consolidation. Fig. 5a shows performance changing with the parameter γ in Eq. (10). $\gamma = 0$ means that we only use CA. When $\gamma = 0.0002$, the result achieves optimal performance. We conduct α_p-β_p **schedule analysis** on the transfer task from Open-i to NIH-CXR for Atelectasis. In our experiments, we set the α_p-β_p schedule in Eq. (5):

$$t(p) = 2(1 + \exp(-p))^{-1} - 1. \tag{11}$$

To investigate the sensitivity with respect to the choice of α_p-β_p schedule, we employ the Charbonnier penalty func-

Table 4. AUROC (%) on 12 transfer tasks across four domains for Pneumonia.

Method	C→M	C→N	C→O	M→C	M→N	M→O	N→C	N→ M	N→O	O→C	O→M	O→N	Average
Source-only	70.339	73.226	86.395	70.921	74.102	96.197	59.245	64.052	84.125	50.941	53.132	58.500	70.098
DANN [16]	70.437	71.394	88.988	**71.648**	73.950	94.967	63.401	64.903	85.412	49.817	54.427	61.623	70.914
CDAN [33]	70.976	73.844	86.660	71.557	73.612	96.216	63.198	64.080	90.823	52.423	54.086	58.077	71.296
MCD [37]	66.750	71.630	88.855	62.669	72.313	90.974	57.260	62.363	75.043	52.088	53.882	56.229	67.505
MDD [49]	**71.085**	73.107	89.309	71.629	74.611	95.459	59.465	65.019	91.466	51.892	54.358	61.998	71.617
DALN [8]	69.462	72.963	92.772	70.806	74.398	97.048	60.577	64.415	84.219	50.918	53.634	62.016	71.102
ST [27]	68.715	73.468	87.663	71.030	73.960	96.821	61.990	64.168	78.486	52.953	55.375	60.621	70.438
CST [31]	65.816	72.203	79.357	66.987	72.476	90.066	52.185	61.476	**93.188**	52.647	53.694	60.223	68.360
DCA (ours)	70.766	**75.267**	**93.113**	71.500	**75.498**	**97.578**	**65.279**	**65.603**	90.653	**60.489**	**56.582**	**68.529**	**74.238**

Table 5. AUROC (%) on 12 transfer tasks across four domains for Cardiomegaly.

Method	C→M	C→N	C→O	M→C	M→N	M→O	N→C	N→ M	N→O	O→C	O→M	O→N	Average
Source-only	76.015	86.375	93.488	79.743	79.606	89.272	76.831	71.493	83.631	69.083	70.371	73.017	79.077
DANN [16]	76.577	87.784	92.786	79.365	**84.334**	87.030	78.961	71.561	85.838	69.830	70.768	69.229	79.505
CDAN [33]	76.674	87.297	93.590	79.230	80.079	88.424	78.818	71.847	86.121	70.165	69.880	70.396	79.377
MCD [37]	71.041	86.411	93.871	77.395	73.531	87.666	69.717	69.444	71.047	70.535	70.104	74.854	76.301
MDD [49]	**76.690**	86.190	93.463	**80.755**	80.049	86.885	76.762	**74.021**	84.793	68.483	70.383	71.039	79.126
DALN [8]	75.874	87.006	93.251	78.931	80.106	87.214	77.045	71.029	87.214	69.527	69.785	71.870	79.071
ST [27]	73.933	86.745	93.302	77.588	82.407	88.631	77.594	73.054	83.777	69.604	69.652	75.568	79.321
CST [31]	75.836	82.206	93.156	79.189	74.951	83.430	74.743	71.603	**90.416**	59.611	66.967	70.742	76.904
DCA (ours)	76.204	**89.086**	**94.027**	80.701	82.662	**90.968**	**79.286**	73.221	88.064	**73.383**	**73.421**	**76.624**	**81.471**

Table 6. Ablation analysis: From Open-i to NIH-CXR for Atelectasis (A), Pneumonia (P), Consolidation (C).

L_{st}	L_{csd}	L_{dsd}	A	P	C	Avg
-	-	-	67.735	58.500	59.986	62.074
-	✓	-	68.850	60.191	67.582	65.541
-	-	✓	69.064	63.747	69.282	67.364
✓	-	-	69.155	64.683	69.665	67.834
✓	✓	-	69.512	68.518	70.019	69.350
✓	-	✓	70.066	67.110	69.987	69.054
✓	✓	✓	**70.800**	**68.529**	**70.332**	**69.887**

Table 7. Calibration analysis: ECE of different confidence bins.

Confidence	0∼0.2	0.2∼0.4	0.4∼0.6	0.6∼0.8	0.8∼1.0
Source-only	0.31	0.23	0.34	0.62	0.89
DCA	0.28	0.20	0.22	0.07	0.20

tion [6] as new α_p-β_p schedules:

$$t(p) = (p^2 + 0.001^2)^{\eta}. \tag{12}$$

By altering η in $[0.1, 0.3, 0.5, 1.0, 2.0]$, we obtain concave, linear, and convex functions respectively, while α_p varies from 0 to 1. Fig. 5b show performance changing with η. It can be observed that our method is quite stable for the parameter γ and the choice of α_p-β_p schedules.

Ablation Analysis We verify the effectiveness of all learning modules. As shown in Tab. 6, L_{st} means whether to use Curriculum Adaptation. L_{csd} means whether to use Class-wise Spectral Debiasing. L_{dsd} means whether to use Domain-wise Spectral Debiasing. We observe that all three modules can enhance adaptation performance and promote each other, verifying the effectiveness of each module.

Calibration Analysis DCA incorporates implicit confidence calibration. We use Expected Calibration Error (ECE) [19] to investigate this issue on the transfer task from CheXpert to NIH-CXR for Atelectasis. Smaller ECE means better calibration. In Tab 7, DCA reduces ECE, especially in high-confidence bins, improving prediction reliability. More analysis is included in the supplementary material.

5. Conclusion

The entanglement of class imbalance and distribution shift poses challenges for knowledge transfer algorithm in chest X-ray classification, leading to negative transfer. In this paper, Curriculum Adaptation is proposed to transfer knowledge in an easy-to-hard way, alleviating confirmation bias. Additionally, Spectral Debiasing is introduced to mitigate representation bias. DCA outperforms state-of-the-art UDA methods and successfully alleviates negative transfer.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62306085, 62406112, 62476071, and U24A20328), Shenzhen College Stability Support Plan (GXWD20231130151329002, and GXWD20220811173233001), Guangdong Basic and Applied Basic Research Foundation (2025A1515012932).

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020. 2

[2] Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory: 23rd International Conference*, 2012. 2

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 2010. 3

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *annual ICML*, 2009. 3

[5] Róger Bermúdez-Chacón, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. A domain-adaptive two-stream u-net for electron microscopy image segmentation. In *ISBI*, 2018. 2

[6] Andrés Bruhn and Joachim Weickert. Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In *ICCV*, 2005. 8

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, 2019. 2

[8] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *CVPR*, 2022. 1, 2, 6, 7, 8

[9] Veronika Cheplygina, Isabel Pino Pena, Jesper Holst Pedersen, David A Lynch, Lauge Sørensen, and Marleen De Bruijne. Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics*, 2017. 2

[10] Louke Delrue, Robert Gosselin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck. Difficulties in the interpretation of chest radiography. *Comparative interpretation of CT and standard radiography of the chest*, 2011. 1

[11] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 2015. 6

[12] Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 1993. 3

[13] Hehe Fan, Xiaojun Chang, Wanyue Zhang, Yi Cheng, Ying Sun, and Mohan Kankanhalli. Self-supervised global-local structure modeling for point cloud domain adaptation with reliable voted pseudo labels. In *CVPR*, 2022. 2

[14] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 2021. 4

[15] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 2006. 6

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 2016. 1, 2, 6, 7, 8

[17] Yufei Gao, Yameng Zhang, Zhiyuan Cao, Xiaojuan Guo, and Jiacai Zhang. Decoding brain states from fmri signals by using unsupervised domain adaptation. *IEEE journal of biomedical and health informatics*, 2019. 2

[18] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation: Volume II: Linear Algebra*. 1971. 5

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 8

[20] Tobias Heimann, Peter Mountney, Matthias John, and Razvan Ionasec. Learning without labeling: Domain adaptation for ultrasound transducer localization. In *MICCAI*, 2013. 2

[21] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[22] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 2

[23] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*, 2019. 6

[24] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 6

[25] Jennifer Kamphenkel, Paul F Jäger, Sebastian Bickelhaupt, Frederik Bernd Laun, Wolfgang Lederer, Heidi Daniel, Tristan Anselm Kuder, Stefan Delorme, Heinz-Peter Schlemmer, Franziska König, et al. Domain adaptation for deviating acquisition protocols in cnn-based lesion classification on diffusion-weighted mr images. In *MICCAI*, 2018. 2

[26] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decou-

pling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 2

[27] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 1, 6, 7, 8

[28] Wei Li, Yifei Zhao, Xi Chen, Yang Xiao, and Yuanyuan Qin. Detecting alzheimer's disease on small dataset: A knowledge transfer perspective. *IEEE journal of biomedical and health informatics*, 2018. 2

[29] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, 2022. 2

[30] Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *CVPR*, 2021. 2

[31] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *NeurIPS*, 2021. 1, 2, 6, 7, 8

[32] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013. 2

[33] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 1, 2, 6, 7, 8

[34] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009. 2

[35] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020. 4

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1, 2, 6, 7, 8

[38] Terence D Sanger. Neural network learning control of robot manipulators using gradually increasing task difficulty. *IEEE transactions on Robotics and Automation*, 1994. 3

[39] Yongheng Sun, Duwei Dai, and Songhua Xu. Rethinking adversarial domain adaptation: Orthogonal decomposition for unsupervised domain adaptation in medical image segmentation. *Medical Image Analysis*, 2022. 2

[40] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. 2010. 3

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[42] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. 6

[43] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *CVPR*, 2019. 2

[44] Zhenbin Wang, Mao Ye, Xiatian Zhu, Liuhan Peng, Liang Tian, and Yingying Zhu. Metateacher: Coordinating multi-model domain adaptation for medical image classification. In *NeurIPS*, 2022. 2

[45] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*, 2021. 2

[46] Zhongqi Yue, Qianru Sun, and Hanwang Zhang. Make the u in uda matter: Invariant consistency learning for unsupervised domain adaptation. In *NeurIPS*, 2023. 2

[47] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3

[48] Jing Zhang, Mingxia Liu, Yongsheng Pan, and Dinggang Shen. Unsupervised conditional consensus adversarial network for brain disease identification with structural mri. In *MICCAI*, 2019. 2

[49] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019. 1, 2, 6, 7, 8

[50] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 2