Subtype-aware Unsupervised Domain Adaptation for Medical Diagnosis

Xiaofeng Liu,[1,2] **Xiongchang Liu,** [2,3] **Bo Hu,** [2,4] **Wenxuan Ji,**[5] **Fangxu Xing,** [1] **Jun Lu*,** [2] **Jane You,** [6]
C.-C. Jay Kuo,[7] **Georges El Fakhri,** [1] **Jonghye Woo** [1]

[1] Dept. of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[2] Dept. of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[3] China University of Mining and Technology, China
[4] Beijing University of Posts and Telecommunications, China
[5] School of Artificial Intelligence, Nankai University, China
[6] Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong
[7] Dept. of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA

Abstract

Recent advances in unsupervised domain adaptation (UDA) show that transferable prototypical learning presents a powerful means for class conditional alignment, which encourages the closeness of cross-domain class centroids. However, the cross-domain inner-class compactness and the underlying fine-grained subtype structure remained largely underexplored. In this work, we propose to adaptively carry out the fine-grained subtype-aware alignment by explicitly enforcing the class-wise separation and subtype-wise compactness with intermediate pseudo labels. Our key insight is that the unlabeled subtypes of a class can be divergent to one another with different conditional and label shifts, while inheriting the local proximity within a subtype. The cases with or without the prior information on subtype numbers are investigated to discover the underlying subtype structure in an online fashion. The proposed subtype-aware dynamic UDA achieves promising results on a medical diagnosis task.

Introduction

The goal of unsupervised domain adaptation (UDA) is to transfer knowledge learned from a label-rich domain to new unlabeled target domains (Saito et al. 2017; Liu et al. 2020c; Liu 2020; Zou et al. 2019). The conventional adversarial training and maximum mean discrepancy (MMD) based methods propose to align the marginal distribution of sample x in the feature space, i.e., $p(f(x))$, where $f(\cdot)$ is a feature extractor. Given the Bayes' theorem $p(f(x)|y) = \frac{p(y|f(x))p(f(x))}{p(y)}$, suppose that there are no concept and label shifts (i.e., $p(y|f(x))$ and $p(y)$ are the same for two domains), then the conditional distribution $p(f(x)|y)$ can be aligned by aligning $p(f(x))$ (Liu et al. 2021a). Nonetheless, the label shift $p(y)$ is quite common in real-world applications, which indicates the label proportion is different (Zhao et al. 2019).

Recently, transferable prototypical networks (TPN) (Pan et al. 2019) is proposed to promote the source domain class separation with a cross-entropy (CE) loss, and match the class centroids of source and target samples to perform
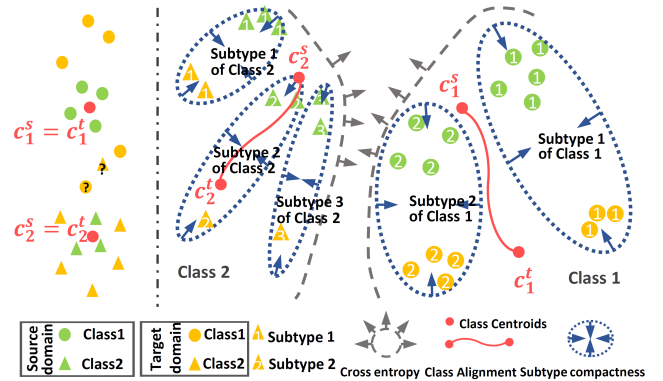
Figure 1: Illustration of the failure case of prototypical UDA (left), and the idea of our subtype-aware UDA (right).

class-wise alignment. However, the CE loss in the source domain cannot minimize the inner-class variation (Liu et al. 2016). The class-wise separation in the to-be tested target domain cannot be well supported by the centroid closeness objective. As shown in Fig. 1 left, although the class centroids are well aligned, the sparsely distributed target samples can be easily misclassified.

One way to tackle this is by simply enforcing the cross-domain inner-class feature distribution compactness. However, in many cases, the unlabeled subtypes in a class can be diverse, and form an underlying local distribution. For instance, different cancer subtypes may have significantly diverse patterns (Yeoh et al. 2002). In such circumstances, the shared pattern among two different subtypes may not be exclusive for class-level discrimination. In these applications, it would be more reasonable and effective to exploit the subtype-wise patterns. Moreover, unsupervised deep clustering (Caron et al. 2018) empirically assigns 10× more clusters of the class to contain diverse subtypes. Recent works also show the fine-grained label can be helpful for the coarse classification (Chen et al. 2019b).

Moreover, domain shifts can be different w.r.t. subtypes, which leads to *subtype conditional shift*. Besides, the incidence of disease subtypes is usually varied across differ-

ent regions, which leads to *subtype label shift*. The proportion difference at the subtype-level can usually be more significant than the class-level (Wu et al. 2019). This motivates us to extend the concept of class conditional and label shifts (Kouw 2018) to the fine-grained subtype-level (i.e., $p(f(x)|k)$ and $p(k)$ vary across domains for subtype k). Therefore, a more realistic presumption of UDA can be both the class and subtype conditional and label shifts.

In this work, we resort to the feature space metric learning with intermediate pseudo labels to adaptively achieve both class-wise separation and cross-domain subtype-wise compactness. We first propose an online clustering scheme to explore the underlying subtype structure in an unsupervised fashion. With the prior knowledge of subtype numbers, concise k-means clustering can be simply applied, by assigning k to the subtype numbers. However, the subtype can be challenging to define, due to different taxonomy. We thereby further expand on our framework to unknown subtype numbers by capturing the underlying subtype structure with an adaptive sub-graph scheme using a reliability-path. With a few meta hyperparameters shared between clusters, the sub-graph scheme can be scalable to several classes and subtypes. To explicitly enforce the subtype-wise distribution compactness, the involved samples of a subtype are expected to be close to their subtype centroid in the feature space.

Our main contribution can be summarized as follows:

- We propose to adaptively explore the subtype-wise conditional and label shifts in UDA without the subtype labels, and explicitly enforce the subtype-aware compactness.

- We systematically investigate the cases with or without the prior information on subtype numbers. Our reliability-path based sub-graph scheme can effectively explore the underlying subtype local distribution with a few meta hyperparameters in an online fashion.

- We empirically validate its effectiveness on a multi-view congenital heart disease (CHD) diagnosis task with an efficient multi-view processing network and achieve promising performance.

Related Work

In recent years, big data drives the fast development of deep learning, which has transformed many fields, such as computer vision and medical image analysis (Han et al. 2020; Liu et al. 2019a, 2020f). Deep learning has drastically transformed the way in which features are extracted and then fed into a prediction model into simultaneously learning both features and a prediction model in an end-to-end fashion (Che et al. 2019; Liu et al. 2021c; Xing et al. 2021). The effectiveness of deep learning has been demonstrated in many computer vision tasks, such as classification, detection, and segmentation (Liu et al. 2019f, 2020d, 2018a, 2019c). In addition, to date, conventional machine learning research in medical image analysis has relied on hand-crafted features (Maraci et al. 2017; Liu et al. 2018c, 2017a) and expert decision rules (De Fauw et al. 2018). End-to-end deep learning approaches have also shown promising performance in many disease diagnosis tasks (Liu et al. 2019d, 2018e). For

example, (Litjens et al. 2019) attempts to analyze ultrasonic data with deep learning and provides a diagnosis suggestion.

Compared with over one million images of the ImageNet dataset, the collection of large-scale medical data is challenging for clinical applications (Liu et al. 2020b, 2018d; He et al. 2020a). To counter this, UDA has gradually become popular (Zou et al. 2019; Liu et al. 2020c, 2021b, 2020e), which aims to match covariate shift (i.e., only $p(x)$ shift). Discrepancy-based methods (Long et al. 2015), such as minimizing MMD, address the dataset shift by mitigating specific discrepancies defined on different layers of a shared model between domains. Recently, adversarial training utilizes a discriminator to classify the domain to encourage domain confusion (Tzeng et al. 2017; Liu et al. 2020a; He et al. 2020b). These methods assume that the label proportion is invariant for all of the involved domains (Moreno-Torres et al. 2012). Yet, the conditional shift (Magliacane et al. 2018) (i.e., only $p(x|y)$ shift) can be more realistic than the covariate shift (Zhao et al. 2019), and the class label shift (i.e., only $p(y)$ shift) also widely exists in the most of real-world applications (Kouw 2018). Furthermore, the subtype-wise conditional and label shifts are a realistic assumption in many applications, which can be more challenging than the class-wise shifts, due to the unavailability of subtype labels in both source and target domains.

Recently, pseudo labels of a target domain have been widely used in UDA (Zou et al. 2019; Liu et al. 2020c). The pseudo labels are used to estimate target class centers (Chen et al. 2019a), and the results are enforced to match the source class centers. Contrastive Adaptation Network (Kang et al. 2019) is proposed to estimate contrastive domain discrepancy with the target pseudo labels. Considering that the pseudo labels can be noisy, the Gaussian-uniform mixture model is proposed to measure the correctness (Gu, Sun, and Xu 2020). In this work, rather than using a sophisticated noisy model, we propose a noise-robust sub-graph scheme with a simple online semi-hard mining method (Liu et al. 2017b, 2019e).

The class-wise conditional alignment (Pan et al. 2019) is also proposed by matching the source and target class centers as in (Chen et al. 2019a). Moreover, source centers can be regarded as class protocols for classification. Although the source class-wise separation can be enforced by its CE loss (Liu et al. 2016), the center matching does not encourage the compactness of source and target samples, and can lead to sparse target distribution and considerable inner-class variation.

The center loss (Wen et al. 2016) is proposed to encourage the compact distributed representation feature for face identification. Following this line of research, numerous works (Liu et al. 2017b, 2018b; Xu et al. 2020; Liu et al. 2019e,b) adapt the center loss to metric learning and optimal transport methods. However, the underlying subtype distribution and shifts are largely ignored. Moreover, the online exploring of the subtype proposed in this work is also closely related to the unsupervised deep clustering (Caron et al. 2018). Not limited by using k-means with the prior knowledge of the subtype numbers, we further propose a scalable sub-graph scheme without the need for the subtype number. Notably,

the dynamic memory framework is robust to the pseudo label noise and subtype undersampling.

Methodology

In this work, we consider the UDA task, where we have a source domain $p^s(x, y)$ and a target domain $p^t(x, y)$, and our learning framework has access to a labeled source set $\{(x_i^s, y_i^s)\}$ drawn from $p^s(x, y)$ and an unlabeled target set $\{(x_i^t)\}$ drawn from $p^t(x, y)$. The class label space of $y_i \in \{1, 2, \cdots, N\}$ is shared for both source and target domains. We use n to index N classes. For class n, we assume that there are K_n underlying subtypes indexed with $k \in \{1, 2, \cdots, K_n\}$. UDA aims to build a good classifier in the target domain $p^t(x, y)$ with the following theorem:

Theorem 1 For a hypothesis h drawn from $\mathcal{H}$, $\epsilon^t(h) \leq \epsilon^s(h) + \frac{1}{2}d_{\mathcal{H}\triangle\mathcal{H}}\{s, t\} + \min_{h \in \mathcal{H}}[\epsilon^s(h, l_s) + \epsilon^t(h, l_t)]$.

Here, $\epsilon^s(h)$ and $\epsilon^t(h)$ denote the expected loss with hypothesis h in the source and target domains, respectively. Considering that the disagreement between labeling function l_s and l_t, i.e., $\min_{h \in \mathcal{H}}[\epsilon^s(h, l_s) + \epsilon^t(h, l_t)]$, can be small by optimizing h with the source data (Ben-David et al. 2007), the UDA focuses on minimizing the cross-domain divergence $d_{\mathcal{H}\triangle\mathcal{H}}\{s, t\}$ in the feature space of $f(x_i^s)$ and $f(x_i^t)$.

Instead of aligning $p^s(f(x))$ and $p^t(f(x))$ (Kouw 2018), the prototypical networks propose to match the class centroids (Pan et al. 2019). However, the decision boundary of the low-density distributed target sample can be difficult to define, and the inherent subtype structure is underexplored.

On the embedding space, we expect the class separation in the target domain can be achieved when the source domain classes are well-separated, and the class-wise source-target distribution compactness is enforced. Moreover, under the fine-grained subtype local structures and their conditional and label shifts, the subtype-wise tight clustering can be a good alternative to achieve class-wise alignment and compactness. Accordingly, we have the following proposition:

Proposition 1. The class-wise compactness can be a special case of subtype-wise compactness by assigning the subtype number of this class to 1.

Targeting the conditional and label shifts in both class-level ($p^s(f(x)|y) \neq p^t(f(x)|y), p^s(y) \neq p^t(y)$) and subtype-level ($p^s(f(x)|k) \neq p^t(f(x)|k), p^s(k) \neq p^t(k)$), we propose a novel subtype-aware alignment framework based on an adaptive clustering scheme as shown in Fig. 1.

Class-wise Source Separation and Matching

Forcing the separation of classes in the source domain, $p^s(x, y)$, can be achieved by the conventional CE loss (Liu et al. 2016). With the extracted features, we carry out the classification via a remold of the distance to each class cluster center (Chen et al. 2019a; Pan et al. 2019).

For the labeled source data $\{(x_i^s, y_i^s)\}$, we represent the feature distribution of class n with a class centroid $c_n^s = \frac{1}{M_c^s} \sum_{i=1}^{M_c^s} f(x_i^s)$, where M_c^s is the involved source sample number. For an input source sample x_i^s, we can directly produce a probability histogram with the softmax normalized distance between x_i^s and the centroids c_n^s. Specifically, the probability of x_i^s belonging to class n can be formulated as

$$p(y_i^s = n|x_i^s) = \frac{e^{-||f(x_i^s)-c_n^s||_2^2}}{\sum_{n=1}^{N} e^{-||f(x_i^s)-c_n^s||_2^2}}. \quad (1)$$

With the one-hot encoding of true class n, we define the class-wise CE loss for the source domain samples as $\mathcal{L}_{CE}^{class} = -logp(y_i^s = n|x_i^s)$.

Following the self-labeling scheme (Zou et al. 2019; Pan et al. 2019), each target sample x_i^t is assigned with a pseudo class label $\hat{y}_i^t$ to its nearest source centroids, i.e., $\hat{y}_i^t = n$ if $\min_{\forall n}||f(x_i^t) - c_n^s||_2^2$. With the pseudo class label $\hat{y}_i^t$, we calculate the target domain class-level centroids $c_n^t = \frac{1}{M_c^t} \sum_{i=1}^{M_c^t} f(x_i^t)$, where M_c^t is the involved target sample number. We expect the close proximity of c_n^s and c_n^t with $\mathcal{L}^{class} = \frac{1}{N} \sum_{n=1}^{N} ||c_n^s - c_n^t||_2^2$, which is not sensitive to label shift, since it only chooses the representative centroids of the source and target distribution[1]. However, neither $\mathcal{L}_{CE}^{class}$ nor $\mathcal{L}^{class}$ considers the inner-class compactness (Wen et al. 2016) and the fine-grained subtype structure.

Subtype-aware Alignment with K_n Prior

If the subtype numbers K_n of class n is known (e.g., 4 subtypes in the CHD disease dataset), we can achieve feature space class-independent clustering with the concise K-means, by defining K to be K_n.

We denote K_n clustered subtypes with $k \in \{1, 2, \cdots, K_n\}$, and calculate the source and target subtype centroids μ_k^s and μ_k^t, respectively. However, K-means does not assign the specific class label to each cluster. To correlate the source and target clusters, we rank the distance of K_n^2 subtype centroid pairs and link the smallest rank first.

Because of the imbalance distribution of subtypes and possible label shift, we assign the subtype centroids of both the source and target samples with $\mu_k^{st} = \frac{\mu_k^s + \mu_k^t}{2}$ instead of averaging all of the source and target samples in subtype k. Therefore, each subtype in both source and target domains contributes equally to μ_k^{st}. Then, we enforce all of the samples in subtype k to be close to the subtype centroid $\mu_k^{s,t}$. For the sake of simplicity, we omit the class notation. The subtype compactness objective $\mathcal{L}_k^{sub}$ can be

$$\frac{1}{M_k^s} \sum_{i=1}^{M_k^s} ||f(x_i^s) - \mu_k^{st}||_2^2 + \frac{1}{M_k^t} \sum_{i=1}^{M_k^t} ||f(x_i^t) - \mu_k^{st}||_2^2, \quad (2)$$

where M_k^s and M_k^t are the numbers of source and target samples in subtype k, respectively, to balance the subtype

---

[1]The source & target center $c_n^{st} = \frac{\sum_{i=1}^{M_c^s + M_c^t} f(x_i^{st})}{M_c^s + M_c^t}$ used in (Pan et al. 2019), and its objective of close proximity of $c_n^s \leftrightarrow c_n^{st}$, or $c_n^t \leftrightarrow c_n^{st}$ are not robust to label shift. Note that c_n^{st} will change if we simply double the involved source/target samples.
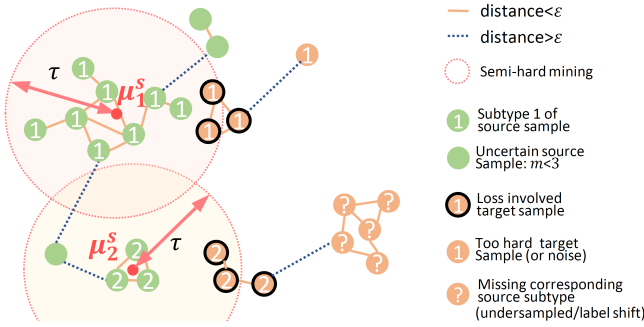
Figure 2: Illustration of the reliability-path based sub-graph construction and alignment with $m = 3$. ❓ will be assigned to subtype 2, which will then be rejected by τ.

label shift. $\mathcal{L}_k^{sub}$ is traversed for N classes and their K_n subtypes to calculate the sum of normalized subtype compactness loss $\mathcal{L}^{sub} = \frac{1}{N}\sum^N(\frac{1}{K_n}\sum^{K_n}\mathcal{L}_k^{sub})$. Note that an image that does not belong to class n does not belong to any of its K_n subtypes. The class-wise matching and subtype-wise compactness objectives can be aggregated as a hierarchical alignment loss $\frac{1}{N}\sum^N(\alpha\mathcal{L}^{class} + \beta\frac{1}{K_n}\sum^{K_n}\mathcal{L}_k^{sub})$, where α and β are the balancing parameters. In the feature space, the learned representations are expected to form K_n compact clusters for class n, while each cluster does not need to be far away from one another.

Reliability-path based Sub-graph Construction

Defining or estimating K_n can be difficult in many applications. Setting K_n of all classes as N hyper-parameters requires costly trails considering the diverse value range of different classes. Note that with a deterministic encoder f, the distribution protocol can be similar among different subtypes and classes (Fahad et al. 2014).

Therefore, we propose constructing the sub-graph with the reliability-path to achieve the online source domain subtype clustering. We assume that similar samples are likely to be distributed closely in the feature space with a deterministic encoder f and form a high-density region (Carlucci et al. 2019). Given M^s samples from the class n in the source domain, there are $(M^s)^2$ possible edges in a graph. To explore the local structure of the feature space, the two nodes $\{f(x_i^s), f(x_j^s)\}$ are connected by the reliability-path, if $||f(x_i^s) - f(x_j^s)||_2^2 \leq \epsilon$. The directly or indirectly linked nodes are combined to form a sub-graph.

To further eliminate the effect of noise and undersampled subtypes on a batch, we only select the sub-graphs with more than m nodes as the valid subtype clusters.

After exploring the K_n subtypes in the source domain and calculating their centroids μ_k^s, we assign each target sample with the pseudo label of class n to the subtype with the most similar centroid (i.e., $\min_{\forall k}||f(x_i^t) - \mu_k^s||_2^2$).

Considering the relatively low confidence or reliability of pseudo target labels (Zou et al. 2019; Gu, Sun, and Xu 2020), we adopt a simple online semi-hard mining scheme to select the target sample in a subtype. The cross-domain

margin τ is used to define a circle at the center of μ_k^s. For the target sample with the initial pseudo subtype label k, we choose these samples to distribute within the circle. We note that some target samples may be densely distributed around the circle boundary, and it is not reasonable to cut them apart simply. Therefore, we also resort to the reliability-path to involve the closely distributed neighboring target samples. The sub-graph construction can be robust to missing subtypes in the source or target domains caused by undersampling, since m filters the unreliable source cluster out, and the self-labeling with semi-hard mining of $f(x_i^t)$ rejects the additional subtypes in the sampled target domain. The operation is illustrated in Fig. 2.

With the reliability-path connected M_k^s source samples and the refined M_k^t target samples in subtype k, we calculate $\mu_k^{st} = \frac{\mu_k^s + \mu_k^t}{2}$, and enforce the subtype-wise compactness with $\mathcal{L}_k^{sub}$ as in Eq. (2).

The online sub-graph construction and alignment have three hyperparameters, including ϵ, τ, and m that are shared for all classes and their subtypes, which can be regarded as the meta-knowledge across clusters. Moreover, we can simplify ϵ to be the constant 1, and change it to any other positive value results only in the matrices being multiplied by corresponding factors (Liu et al. 2017b). The range of m can also be narrow and similar among different subtypes/classes.

Optimization and Implementation

The modern neural networks usually extract a high-dimensional vector, e.g., 4,096 or 2,048-dimensional features, as their representation, thereby demanding high memory and time complexity in subsequent clustering. To remedy this, deep clustering (Caron et al. 2018) proposes to perform dimension reduction via Principal Component Analysis (PCA) for the extracted features from all of the samples in a dataset. PCA, however, is not applicable anymore in our online SubUDA. The feature representations are extracted in different training iterations with different timestamps, which can have incompatible statistics. It is also computationally demanding to carry out PCA for all iterations. Accordingly, a non-linear head-layer with the structure of fc→bn→relu→dropout→fc→relu is adopted in order to reduce high dimensional features into 256 dimensions. As well, it is simultaneously optimized in each online SubUDA iteration. The non-linear head-layer is eliminated for the subsequent operations, e.g., calculating the L2 distance between features.

In order to prevent the subtype clustering from collapsing to a few subtype groups, (Caron et al. 2018) makes uniform sampling in all of the epochs, which is difficult in our online UDA setting, due to the missing subtype and target class labels. We thus propose a concise approach for SubUDA via re-weighting the loss with $\omega_k \propto \frac{1}{\sqrt{M_k^s + M_k^t}}$, according to the number of samples in the k-th subtype. Therefore, samples in smaller clusters are accounted more for the loss, which thereby pushes the classification boundary away to incorporate as more samples as possible. The optimization objective
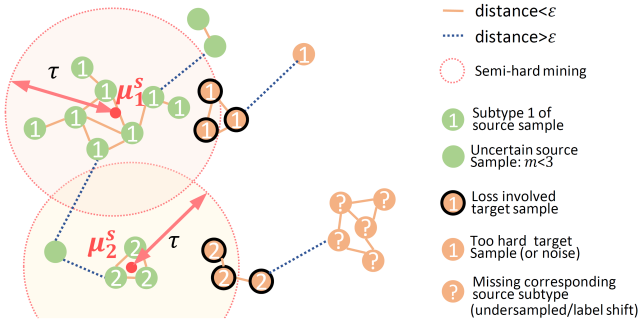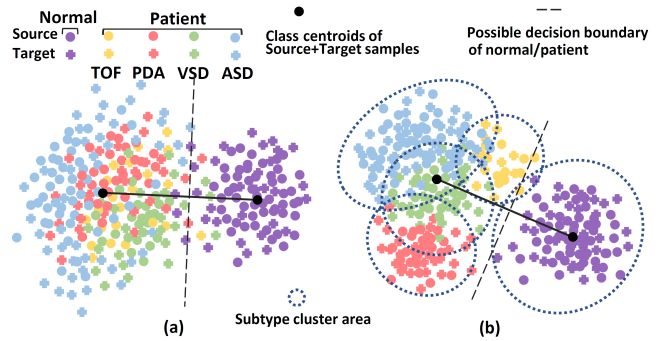
Figure 3: Data pre-processing flowchart.

Figure 4: T-SNE visualization of the sampled CHD features with (a) TPN (Pan et al. 2019) and (b) our SubUDA. We normalized the distance between class centroids to demonstrate the class separation and inner-class/subtype compactness.

can be summarized as

$$\mathcal{L} = \mathcal{L}_{CE}^{class} + \frac{1}{N} \sum^{N} (\alpha \mathcal{L}^{class} + \beta \frac{\omega_k}{K_n} \sum^{K_n} \mathcal{L}_k^{sub}). \quad (3)$$

For the classification in testing, we utilize the centroids of training features as prototypical (Pan et al. 2019).

Experiments

We carried out experiments using CHD data to evaluate the effectiveness of our approach. We implemented our method and comparison methods using PyTorch and set $\alpha = 1$, $\lambda = 0.5$, and $\beta = 0.5$ consistently.

CHD is one of the most common types of birth defect, which usually results in the death of neonates. Therefore, early medical care and treatment can be helpful, which requires an efficient and accurate diagnosis. In clinical practice, clinicians rely on echocardiograms from five heart views. Collected raw data include echocardiogram videos, and a key-frame of each view is usually extracted for the assessment. Specifically, the collected views are from the parasternal long-axis (PSLAX), parasternal short-axis (PSSAX), apical four chambers (A4C), subxiphoid long-axis (SXLAX), and suprasternal long-axis (SSLAX).

To quantify the effect of subtype structure, four subtypes of CHD, including atrial septal defect (ASD), ventricular septal defect (VSD), patent ductus artery (PDA), and tetralogy of Fallot (TOF), are labeled by two clinicians or intraoperative records. We note that the fine-grained subtype label is not used in training, since large-scale labeling can be costly in clinical practice, while the normal/patient label can be relatively easy to acquire by primary clinicians. Of note, this work focuses on exploring the subtype-aware alignment for the conventional class-wise discriminative model.

Specifically, we evaluated our method on five-view echocardiogram datasets collected from two medical centers. We used 1,608 labeled source subjects (normal/patient) from Beijing Children's Hospital (BCH) and 800 unlabeled target subjects from Beth Israel Deaconess Medical Center (BIDMC). Each dataset consists of echocardiograms from five views that are sufficient for the diagnosis.

Abnormal regions are likely to be different from one subtype to another; as such, different subtypes can be easily detected from different views. Considering the large innerclass variation of the patient class, it is reasonable to enforce

the subtype-wise compactness. In Fig. 4 left, we can see that the source CHD samples of a subtype also tend to be distributed closely, demonstrating the underlying inner-subtype similarity. With our subtype-wise compactness objective as shown in Fig. 4 right, both the source and target samples are grouped into the high-density region w.r.t. subtypes.

Data Collection of Cardiac Ultrasound Images

• **BCH source domain**. A total of 1,608 echocardiogram datasets, including 823 healthy controls, 209 VSD, 276 ASD, 124 TOF, and 176 PDA, were collected using PHILIPS iE 33. The chest of each patient was exposed to the echocardiogram with the supine position. We set our transducer frequency between 3 to 8 MHz. We collected the five standard 2D views, i.e., PSLAX, PSSAX, A4C, SXLAX, and SSLAX.

• **BIDMC target domain**. A total of 800 echocardiogram datasets, including 300 healthy controls, 150 VSD, 150 ASD, 100 TOF, and 100 PDA, were collected. Similarly, the chest of each patient was exposed to the echocardiogram with the supine position. We used PHILIPS EPIQ 7C for echocardiogram imaging and set its transducer frequency from 3 to 8 MHz. We also collected the five standard 2D views, i.e., PSLAX, PSSAX, A4C, SXLAX, and SSLAX.

Notably, discrepancies between these two medical centers include imaging devices (PHILIPS iE 33 vs. EPIQ 7C), patient populations, and clinicians' echocardiogram imaging experience, which introduced domain shifts.

We also extracted the key frame from the video of each view. The heart is a dynamic organ and has different shape within a heart cycle. Thus, we chose a time frame that the clear defects were visible. Specifically, the key frame was corresponding to the isovolumic relaxation phase.

We selected 80% and 20% subjects from the target domain for training and testing, respectively. Note that we only selected the subjects with all the key frames from the five views as our testing data. In addition, the key frames shown in training were not used for testing in a subject-independent manner.
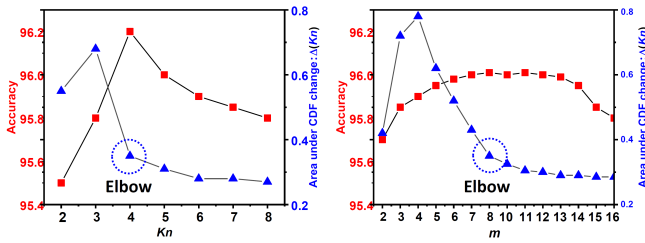
Figure 5: The sensitive analysis of K_n (right) and m of patient class in CHD dataset.

Figure 6: The backbone network via the multichannel convolutional neural networks for five-view echocardiograms analysis.

Pre-processing

The selected color key frames were first translated to the gray images, since the echocardiogram region has only single-channel 2 gray value information. Then, we cropped the region of interest (ROI) with a redefined sector mask. Since a typical input to convolutional neural networks (CNNs) has the size of 128×128, we resized the masked ROI to the 128×128 image. The five views were stacked following a specific sequence (i.e., SLAX, PSSAX, A4C, SXLAX, and SSLAX) to form our five channels training sample.

Backbone Network Structure for the CHD Diagnosis Task

Training a network with massive parameters using a limited number of training data samples usually results in overfitting, which causes a problem in medical image analysis. To alleviate the data constraints in our CHD diagnosis task, we propose to adopt the Depthwise Separable Convolution (DSC) (Howard et al. 2017) as our backbone to reduce the to-be-trained parameters. This strategy was introduced in the MobileNet (Howard et al. 2017) with a lightweight implementation. Notably, as a comparison, the DSC based CNNs with 1.32M weights performed similarly to AlexNet (Simonyan and Zisserman 2014) with 60M weights trained using the ImageNet dataset.

More specifically, the convolution operation in conventional CNNs is separated to the depth-wise and point-wise stages. The depth-wise stage processes each channel independently, while conventional CNNs process all channels together. Then, the 1×1 convolution is used in the subsequent point-wise stage. We adapt the two-stage convolution for our grayscale image analysis.

To fuse the information from the five views, we adopted the multi-channel DSC network. Following the MobileNet (Howard et al. 2017), we configured the first layer as the conventional convolution operation and set the stride as two. We also adopted two fully connected layers with the dimension of 1024 and 128, respectively. The structure of our backbone is shown in Fig. 6, and detailed in Table 1.

In addition to the multi-channel scheme, another feasible choice for multi-view information aggregation is the multi-branch network (Lee, Lee, and Kim 2016). However, our multi-channel framework has a few strengths over the multi-branch network. First, our multi-channel scheme is
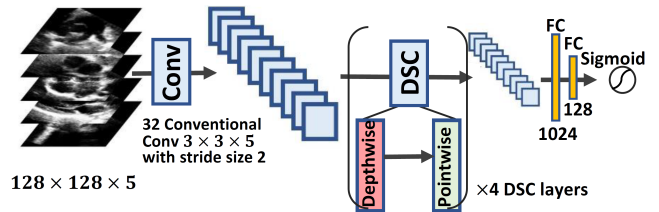
Input Size	Type / Stride	Filter Shape
$128 \times 128 \times 5$	Conv / s2	$32 \times$ of $3 \times 3 \times 5$
$64 \times 64 \times 32$	Conv dw / s1	$32 \times 3 \times 3$ dw
$64 \times 64 \times 32$	Conv pw / s1	$64 \times$ of $1 \times 1 \times 32$ pw
$64 \times 64 \times 64$	Conv dw / s2	$64 \times$ of 3×3 dw
$32 \times 32 \times 64$	Conv pw / s1	$128 \times$ of $1 \times 1 \times 64$ pw
$32 \times 32 \times 128$	Conv dw / s2	$128 \times$ of 3×3 dw
$16 \times 16 \times 128$	Conv pw / s1	$128 \times$ of $1 \times 1 \times 128$ pw
$16 \times 16 \times 128$	Conv dw / s2	$128 \times$ of 3×3 dw
$8 \times 8 \times 128$	Conv pw / s1	$128 \times$ of $1 \times 1 \times 128$ pw
$8 \times 8 \times 128$	Flatten	N/A
8192	FC1	1024
1024	FC2	128
128	Classifier	Softmax

Table 1: The detailed network construction of the DSC-based five-view framework. We denote the depth-wise convolution with the size $H \times W$ as dw, which is the same for all channels. The point-wise convolution is denoted by pw, which has the size of $1 \times 1 \times N$. The conventional convolution operation is used in the first convolutional layer.

able to learn multi-view fusing in all of the layers adaptively, rather than simply concatenating each view in the late layer (Lee, Lee, and Kim 2016). Second, our multi-channel framework only uses a single forward model, which has much fewer to-be-learned weights. Considering that the echocardiograms from five views can share some similarity, the convolutional filters trained in each view may potentially be useful for one another. Thus, our framework with less to-be-learned weights can efficiently deal with the problem of small dataset size and the requirement of large memory in the implementation.

The class-level classification in our task aims to differentiate healthy controls from patients, which can be formulated as a binary classification problem. We thereby apply the sigmoid unit as our output layer, and adopt the binary CE loss as the supervision signal. We report the accuracy with the threshold of 0.5.

Evaluations

For comparison, we re-implemented the current state-of-the-art methods with the same backbone and experiment setting, where we chose the batch size to 64. The results are shown in Table. 2. Considering the imbalance of normal and patient proportion in the testing set, we also provide the area under

Method	Accuracy (%) ↑	AUC ↑
Source only	76.4±0.12	0.721±0.005
MCD (Saito et al. 2018)	88.6±0.15	0.856±0.003
GTA	90.9±0.17	0.873±0.005
CRST (Zou et al. 2019)	93.2±0.09	0.882±0.006
TPN (Pan et al. 2019)	93.4±0.14	0.885±0.004
SubUDA ($K_n = 4$)	96.2±0.13	0.910±0.003
SubUDA ($K_n = 1$)	94.7±0.11	0.902±0.004
SubUDA-μ_k^{st} ($K_n = 4$)	95.4±0.10	0.903±0.005
SubUDA-ω_k ($K_n = 4$)	96.0±0.13	0.908±0.004
SubUDA-DR ($K_n = 4$)	96.2±0.11	0.911±0.002
SubUDA-SG ($m = 8$)	96.0±0.12	0.907±0.004
SubUDA-SG-τ ($m = 8$)	95.5±0.14	0.902±0.003

Table 2: Experimental results for CHD. ↑ larger is better.

Figure 7: Comparison with the state-of-the-art methods w.r.t. $\mathcal{A}$-distance (left), and the sensitive analysis of τ (right).

the receiver operating characteristic curve (AUC) metric in addition to the accuracy metric.

MCD (Saito et al. 2018) and GTA (Sankaranarayanan et al. 2018) are the typical adversarial training frameworks to align the marginal distribution $p(x)$ at feature level or image level, respectively. The self-training is used to alternatively update the pseudo label of target samples and the network parameters (Zou et al. 2019). We note that the compared methods (Saito et al. 2018; Sankaranarayanan et al. 2018; Zou et al. 2019; Wu, Inkpen, and El-Roby 2020) use the fully-connected classifier after the encoder. The TPN (Pan et al. 2019) uses class centroids as a classifier. It proposes to align class centroids of the source and target sample to achieve the conditional alignment w.r.t. $p(x|y)$. Our Sub-UDA outperformed the state-of-the-art methods w.r.t. both the accuracy and AUC by a large margin, by introducing the subtype-aware constraint. The results indicate that the online subtype compactness can effectively help the classification in the target domain.

The domain adaptation theory suggests proxy $\mathcal{A}$-distance (Ben-David et al. 2007) as a measure of cross-domain discrepancy (Saito, Ushiku, and Harada 2017). In Fig. 7, we compare our SubUDA with the other state-of-the-art methods, and the smaller discrepancy has been observed by using the explicit compactness objective in our SubUDA.

For the ablation study, with $K_n = 1$, the subtype-wise alignment was reduced to the class-wise compactness. Besides, we used the suffix -DR, -ω_k, and -τ to denote the sub-UDA without dimension reduction head, subtype balance weight, and semi-hard target mining, respectively. Furthermore, the suffix -μ_k^{st} denotes using $\mu_k^{st} = \frac{\sum_{i=1}^{M_k^s + M_k^t} f(x_i^{st})}{M_k^s + M_k^t}$ as the subtype centroid, which is not robust to the subtype label shift. SubUDA-DR took $4\times$ clustering time, but the improvement was marginal. Therefore, we recommend using the dimension reduction head.

SubUDA-SG in Table 2 used the reliability-path based online sub-graph to replace k-means. With appropriate $m = 8$, the adaptively learned clustering achieved comparable performance to the K-means with $K_n = 4$. In Fig. 7 right, we can see that the semi-hard mining scheme in SubUDA-SG is not sensitive to τ for a large range, since the network
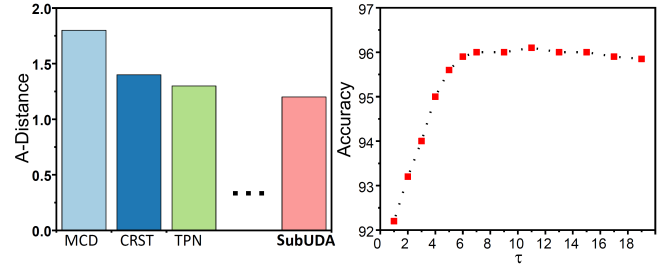
can flexibly learn to adjust the ratio of ϵ and τ in mapping space (Liu et al. 2017b). We note that too strict semi-hard mining (i.e., too small τ) can degenerate our SubUDA to conventional class centroids matching, since no target samples are selected to form a subtype cluster.

In Fig. 5, we provide a sensitive analysis of hyper-parameter K_n and m for two kinds of online clustering schemes. Consensus clustering can assess the clustering stability (Monti et al. 2003), and the optimal clustering is usually achieved in the elbow position of the area under CDF changes, where the CDF is for the consensus matrices[2]. We can see that the peak of accuracy usually coincides with the best clustering consensus metric, which indicates the good subtype clustering can boost the SubUDA performance. The choice of $K_n = 4$ also matches our prior knowledge of the CHD patient subtypes.

Since we have four clear subtypes in this task, using the concise k-means can be a straightforward solution. However, Fig. 5 right shows that the accuracy curve can be robust for a relatively large range of m, which is promising for the hyperparameter tuning of the case without the prior information of subtype numbers.

Conclusion

In this work, we presented a new UDA approach with a more realistic assumption that the subtype-wise conditional and label shifts widely exist, and can be adaptively aligned without the subtype label. We systematically investigate a flexible yet principled solution for the case with and without the prior knowledge of subtype numbers. Rather than the concise k-means, we further extend our framework with an online sub-graph scheme using the reliability-path, which can be scalable to many classes and subtypes with a few meta hyperparameters. The effectiveness of explicitly enforcing the subtype-aware compactness has been successfully demonstrated in the CHD transfer task.

Acknowledgments

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *NIPS*.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *CVPR*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*, 132–149.

Che, T.; Liu, X.; Li, S.; Ge, Y.; Zhang, R.; Xiong, C.; and Bengio, Y. 2019. Deep Verifier Networks: Verification of Deep Discriminative Models with Deep Generative Models. In *ArXiv*.

Chen, C.; Xie, W.; Huang, W.; Rong, Y.; Ding, X.; Huang, Y.; Xu, T.; and Huang, J. 2019a. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*.

Chen, Z.; Ding, R.; Chin, T.-W.; and Marculescu, D. 2019b. Understanding the Impact of Label Granularity on CNN-based Image Classification. *arXiv:1901.07012* .

De Fauw, J.; Ledsam, J. R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24(9): 1342–1350.

Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A. Y.; Foufou, S.; and Bouras, A. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing* 2(3): 267–279.

Gu, X.; Sun, J.; and Xu, Z. 2020. Spherical Space Domain Adaptation With Robust Pseudo-Label Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9101–9110.

Han, Y.; Liu, X.; Sheng, Z.; Ren, Y.; Han, X.; You, J.; Liu, R.; and Luo, Z. 2020. Wasserstein Loss-Based Deep Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

He, G.; Liu, X.; Fan, F.; and You, J. 2020a. Classification-Aware Semi-Supervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

He, G.; Liu, X.; Fan, F.; and You, J. 2020b. Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 912–913.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* .

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*.

Kouw, W. M. 2018. An introduction to domain adaptation and transfer learning. *arXiv:1812.11806* .

Lee, D.; Lee, J.; and Kim, K.-E. 2016. Multi-view automatic lip-reading using neural network. In *Asian conference on computer vision*, 290–302. Springer.

Litjens, G.; Ciompi, F.; Wolterink, J. M.; de Vos, B. D.; Leiner, T.; Teuwen, J.; and Išgum, I. 2019. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular Imaging* 12(8): 1549–1565.

Liu, S.; Wang, Y.; Yang, X.; Lei, B.; Liu, L.; Li, S. X.; Ni, D.; and Wang, T. 2019a. Deep learning in medical ultrasound analysis: a review. *Engineering* .

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In *ICML*.

Liu, X. 2020. Disentanglement for discriminative visual recognition. *arXiv preprint arXiv:2006.07810* .

Liu, X.; B.V.K., K.; Yang, C.; Tang, Q.; and You, J. 2018a. Dependency-aware Attention Control for Unconstrained Face Recognition with Image Sets. In *European Conference on Computer Vision*.

Liu, X.; Che, T.; Lu, Y.; and Yang, C. 2020a. AUTO3D: Novel view synthesis through unsupervisely learned variational viewpoint and global 3D representation. *ECCV* .

Liu, X.; Fan, F.; Kong, L.; Xie, W.; Lu, J.; and You, J. 2020b. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing* .

Liu, X.; Ge, Y.; Yang, C.; and Jia, P. 2018b. Adaptive metric learning with deep neural networks for video-based facial expression recognition. *Journal of Electronic Imaging* 27(1): 013022.

Liu, X.; Guo, Z.; Jia, J.; and Kumar, B. 2019b. Dependency-aware Attention Control for ImageSet-based Face Recognition. In *IEEE Transactions on Information Forensics and Security*.

Liu, X.; Guo, Z.; Li, S.; You, J.; and B.V.K, K. 2019c. Dependency-aware Attention Control for Unconstrained Face Recognition with Image Sets. In *ICCV*.

Liu, X.; Han, X.; Qiao, Y.; Ge, Y.; Li, S.; and Lu, J. 2019d. Unimodal-uniform constrained wasserstein training for medical diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0.

Liu, X.; Hu, B.; Liu, X.; Lu, J.; You, J.; and Kong, L. 2020c. Energy-constrained Self-training for Unsupervised Domain Adaptation. *ICPR* .

Liu, X.; Hu, B.; Liu, X.; Lu, J.; You, J.; and Kong, L. 2021a. Energy-constrained self-training for unsupervised domain adaptation. *arXiv preprint arXiv:2101.00316* .

Liu, X.; Jin, L.; Han, X.; Lu, J.; You, J.; and Kong, L. 2020d. Identity-aware Facial Expression Recognition in Compressed Video. *ICPR* .

Liu, X.; Jin, L.; Han, X.; Lu, J.; You, J.; and Kong, L. 2021b. Identity-aware facial expression recognition in compressed video. *arXiv preprint arXiv:2101.00317* .

Liu, X.; Jin, L.; Han, X.; and You, J. 2020e. Mutual Information Regularized Identity-aware Facial ExpressionRecognition in Compressed Video. *arXiv preprint arXiv:2010.10637* .

Liu, X.; Kong, L.; Diao, Z.; and Jia, P. 2017a. Line-scan system for continuous hand authentication. *Optical Engineering* 56(3): 033106.

Liu, X.; Kumar, B. V.; Jia, P.; and You, J. 2019e. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recognition* 88: 1–12.

Liu, X.; Li, Z.; Kong, L.; Diao, Z.; Yan, J.; Zou, Y.; Yang, C.; Jia, P.; and You, J. 2018c. A joint optimization framework of low-dimensional projection and collaborative representation for discriminative classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 1493–1498. IEEE.

Liu, X.; Vijaya Kumar, B.; You, J.; and Jia, P. 2017b. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPR*.

Liu, X.; Xing, F.; Yang, C.; Kuo, C.-C. J.; Babu, S.; Fakhri, G. E.; Jenkins, T.; and Woo, J. 2021c. VoxelHop: Successive Subspace Learning for ALS Disease Classification Using Structural MRI. *arXiv preprint arXiv:2101.05131* .

Liu, X.; Xing, F.; Yang, C.; Kuo, C.-J.; El Fakhri, G.; and Woo, J. 2020f. Symmetric-Constrained Irregular Structure Inpainting for Brain MRI Registration with Tumor Pathology. In *MICCAI BrainLes*.

Liu, X.; Zou, Y.; Che, T.; Ding, P.; Jia, P.; You, J.; and Kumar, B. V. 2019f. Conservative Wasserstein Training for Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, X.; Zou, Y.; Kong, L.; Diao, Z.; Yan, J.; Wang, J.; Li, S.; Jia, P.; and You, J. 2018d. Data Augmentation via Latent Space Interpolation for Image Classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 728–733. IEEE.

Liu, X.; Zou, Y.; Song, Y.; Yang, C.; You, J.; and K Vijaya Kumar, B. 2018e. Ordinal Regression with Neuron Stick-breaking for Medical Diagnosis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. *ICML* .

Magliacane, S.; van Ommen, T.; Claassen, T.; Bongers, S.; Versteeg, P.; and Mooij, J. M. 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NIPS*.

Maraci, M. A.; Bridge, C. P.; Napolitano, R.; Papageorghiou, A.; and Noble, J. A. 2017. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. *Medical image analysis* 37: 22–36.

Monti, S.; Tamayo, P.; Mesirov, J.; and Golub, T. 2003. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 52(1-2): 91–118.

Moreno-Torres, J. G.; Raeder, T.; Alaiz-RodríGuez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45(1): 521–530.

Pan, Y.; Yao, T.; Li, Y.; Wang, Y.; Ngo, C.-W.; and Mei, T. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*.

Saito, K.; Ushiku, Y.; and Harada, T. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2017. Adversarial Dropout Regularization. *arXiv preprint arXiv:1711.01575* .

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3723–3732.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 499–515. Springer.

Wu, Y.; Inkpen, D.; and El-Roby, A. 2020. Dual Mixup Regularized Learning for Adversarial Domain Adaptation. *ECCV* .

Wu, Y.; Winston, E.; Kaushik, D.; and Lipton, Z. 2019. Domain adaptation with asymmetrically-relaxed distribution alignment. *arXiv:1903.01689* .

Xing, F.; Stone, M.; Prince, J. L.; Liu, X.; Fakhri, G. E.; and Woo, J. 2021. Floor-of-the-mouth muscle function analysis using dynamic magnetic resonance imaging. In *SPIE Medical Imaging 2021: Image Processing*.

Xu, R.; Liu, P.; Wang, L.; Chen, C.; and Wang, J. 2020. Reliable Weighted Optimal Transport for Unsupervised Domain Adaptation. In *CVPR*.

Yeoh, E.-J.; Ross, M. E.; Shurtleff, S. A.; Williams, W. K.; Patel, D.; Mahfouz, R.; Behm, F. G.; Raimondi, S. C.; Relling, M. V.; Patel, A.; et al. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell* 1(2): 133–143.

Zhao, H.; Combes, R. T. d.; Zhang, K.; and Gordon, G. J. 2019. On learning invariant representation for domain adaptation. *arXiv:1901.09453* .

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.; and Wang, J. 2019. Confidence Regularized Self-Training. *ICCV* .