Doubly Supervised Transfer Classifier for Computer-Aided Diagnosis with Imbalanced Modalities

Xiangmin Han, Xiaoyan Fei, Jun Wang, *Member, IEEE*, Tao Zhou, *Member, IEEE*, Shihui Ying, *Member, IEEE*, Jun Shi, *Member, IEEE*, and Dinggang Shen, *Fellow, IEEE*

***Abstract*—Transfer learning (TL) can effectively improve diagnosis accuracy of single-modal-imaging-based computer-aided diagnosis (CAD) by transferring knowledge from other related imaging modalities, which offers a way to alleviate the small-sample-size problem. However, medical imaging data generally have the following characteristics for the TL-based CAD: 1) The source domain generally has limited data, which increases the difficulty to explore transferable information for the target domain; 2) Samples in both domains often have been labeled for training the CAD model, but the existing TL methods cannot make full use of label information to improve knowledge transfer. In this work, we propose a novel doubly supervised transfer classifier (DSTC) algorithm. In particular, DSTC integrates the support vector machine plus (SVM+) classifier and the low-rank representation (LRR) into a unified framework. The former makes full use of the shared labels to guide the knowledge transfer between the paired data, while the latter adopts the block-diagonal low-rank (BLR) to perform supervised TL between the unpaired data. Furthermore, we introduce the Schatten-p norm for BLR to obtain a tighter approximation to the rank function. The proposed DSTC algorithm is evaluated on the Alzheimer's disease neuroimaging initiative (ADNI) dataset and the bimodal breast ultrasound image (BBUI) dataset. The experimental results verify the effectiveness of the proposed DSTC algorithm.**

***Index Terms*—Transfer learning, doubly supervised transfer classifier, modality imbalance, support vector machine plus, block-diagonal low-rank.**

I. INTRODUCTION

WITH the fast development of artificial intelligence, computer-aided diagnosis (CAD) has shown its effectiveness and efficiency to help improve diagnostic accuracy with consistency and repeatability [1][2]. Although the multi-modal-imaging-based CAD models generally achieve superior performance to the single-modal-imaging-based approaches [3][4], the latter ones have more popular and flexible applications, because not all hospitals are equipped with advanced multi-modal imaging devices [5]. However, compared with multi-modal medical images, single-modal medical images only provide partial information, such as structural or functional information, which generally limits CAD performance to a certain extent [1][6][7][8].

Transfer learning (TL) is an effective method to improve the model performance in the target domain by transferring knowledge from the related source domain [9]. It has been successfully applied to various medical image processing tasks [10]. From the clinical viewpoint, there are two types of knowledge transfer for developing a CAD model, *i.e.*, 1) transfer between different diseases and 2) transfer between different imaging modalities [10]. The former cases mainly consider the relevance between two diseases in the source and target domains with the same imaging modality. For example, Cheng *et al.* improved the magnetic resonance imaging (MRI) based diagnosis of Alzheimer's disease (AD) (target domain) by transferring knowledge from mild cognitive impairment (MCI) (source domain), since AD and MCI are considered to have inherent relevance [11]. In the latter cases, the source and target domains generally use two different modalities of the same disease. For example, Fei *et al.* proposed a parameter transfer deep neural network for the B-mode ultrasound (BUS) based CAD of breast cancers by taking the BUS and elastography ultrasound (EUS) as target and source domains, respectively, because EUS provides additional information pertaining to the biomechanical and functional properties of breast lesions [12]. Since the transfer between different modalities is easy to access and has more applications, it has attracted more attention in recent years.

In order to train a CAD model, the acquired imaging data are often labeled. Therefore, the supervised TL methods are valid, since they can transfer more effective knowledge from the source domain to the target domain under the guidance of label information. Moreover, it is a fact that the multi-modal imaging data scanned from a patient naturally share the same label. Therefore, learning using privileged information (LUPI) is suitable for this transfer task, because LUPI is a special

This work is supported by National Natural Science Foundation of China (81830058, 11971296, 62172228) and the 111 Project (D20031). (Corresponding authors: Jun Shi, Dinggang Shen)

X. Han, X. Fei, J. Wang and J. Shi are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, China. (Email: junshi@shu.edu.cn)

T. Zhou is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

S. Ying is with the Department of Mathematics, School of Science, Shanghai University, China.

D. Shen is with the School of Biomedical Engineering, ShanghaiTech University, China; Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China. (Email: Dinggang.Shen@gmail.com)

supervised TL paradigm, which is only conducted on the paired bi-modal data with shared labels [13]. Several pioneering works have successfully applied the LUPI-based classifiers, such as support vector machine plus (SVM+) and its variants [13][14][15], to improve the diagnostic performance of single-modal-imaging-based CAD with the help of additional modalities under the guidance of shared label [16][17][18][19].

It is worth noting that the single-modal imaging samples are more easily acquired than the multi-modal data in clinical practice, because only a few hospitals have full-modal imaging devices, or only partial patients could be scanned with multi-modalities. Therefore, we generally have *not only* the paired multi-modal images, *but also* some additional single-modal data for training a CAD model. This leads to the clinical modality imbalance issue for training a CAD model. However, the LUPI paradigm cannot address this TL problem between imbalanced modalities, because it can only handle the paired data with shared labels.

On the other hand, when developing a TL-based CAD model, it is a suitable way to select the commonly used modality as the target domain, which is equipped in more hospitals for clinical applications. For example, positron emission tomography (PET) device is very expensive and scarce, and thus MRI has more applications than PET to diagnose brain diseases, including AD; BUS device is widely equipped in almost all hospitals as a commonly used technique for diagnosis of breast cancers, while EUS is yet to be a routine diagnostic tool, especially in rural hospitals. Therefore, the MRI- or BUS-based CAD systems can benefit more patients, while its diagnosis performance could be improved by transferring knowledge from the corresponding PET or EUS data in the source domain.

Moreover, these two clinical characteristics result in the following issue: the samples in the source domain are generally less than those in the target domain for developing a CAD model. This increases the difficulty of the conventional TL algorithms to explore transferable information for the target domain, because these algorithms generally require sufficient training samples in the source domain to provide rich transferable knowledge [9]. While according to the domain adaption theory, the conventional TL algorithms automatically aligned domain divergence as the unsupervised domain adaption [20]. However, this unsupervised manner cannot fully utilize label information in the target domain to guide knowledge transfer. Currently, there are few works about the supervised TL to deal with this special clinical modality imbalance issue [21][22]. Therefore, it is necessary to develop a new TL method to effectively address this problem of modality imbalance in a supervised manner.

In this work, we propose a novel doubly supervised transfer classifier (DSTC) algorithm for CAD, which can effectively solve the abovementioned special clinical issue of modality imbalance. Specifically, the proposed DSTC integrates the SVM+ classifier and low-rank representation (LRR) for knowledge transfer between the paired and the unpaired data, respectively, into a unified framework. The experiments on two datasets indicate the effectiveness of the proposed DSTC.

The main contributions are two-fold:

1) A novel DSTC algorithm is proposed to improve the performance of a single-modal imaging-based CAD model, which integrates SVM+ and LRR into a unified framework. DSTC makes full use of both the shared and unshared label information to guide knowledge transfer between the paired and unpaired data, and thus effectively implements the transfer task for the abovementioned special issue of modality imbalance in clinical practice.
2) Different from the previous LRR-based TL methods that evaluate feature correlation between the source and target domains in feature space, the proposed DSTC also incorporates the label knowledge into LRR in a classifier for TL between the unpaired data with different labels. Furthermore, we implement the block-diagonal low-rank (BLR) with the Schatten-p norm to get a tighter approximation to the rank function.

II. RELATED WORK

A. Transfer Learning in CAD

As a classical LUPI algorithm, SVM+ replaces the slack variables in the standard SVM with a set of non-negative slack functions, in which additional data in the source domain is introduced to regularize the hinge loss [13]. Therefore, the hyperplane can be optimized with the guidance of the source domain during the training stage [13]. Various improved SVM+ algorithms, such as fast SVM+ [14], adaptive SVM+ [23], multi-view SVM+ [24], robust SVM+ [25], and random vector functional link network plus (RVFL+) [15], have been proposed for different classification tasks and obtained promising performance.

Currently, LUPI has been successfully applied in medical imaging-based CAD. For example, Duan *et al.* used the single nucleic polymorphisms as the source domain for the fundus image-based glaucoma detection with an SVM+ classifier [17]; Alahmadi *et al.* proposed a generalized matrix learning vector quantization classifier for diagnosis of MCI with the cognitive data as the target domain while functional MRI (fMRI) as the source domain [26]; Zheng *et al.* proposed an ensemble LUPI algorithm to improve the diagnostic performance of MRI-based CAD for brain diseases with another neuroimaging as the source domain [16][27]; Shi *et al.* developed a cascaded multi-column RVFL+ classifier for the MRI-based diagnosis of Parkinson's disease, in which the source domain was self-generated without another modality [18]; Li *et al.* compared different LUPI-based classifiers for AD diagnosis using MRI by regarding PET images as source domain [19]. The above works indicate that the LUPI paradigm can effectively promote the single-modal-imaging-based CAD with the help of additional modality as the source domain.

On the other hand, the conventional TL aims to improve the performance of the model in the target domain by leveraging the knowledge from the source domain [9], and it has been successfully applied to different CAD models [28]. For example, Cheng et al. proposed a multi-modality domain transfer support vector machine (SVM) algorithm for MCI

conversion prediction, where the data of AD and normal control (NC) subjects were used as the source domain to improve the predictive performance[29]; Wachinger *et al.* proposed a supervised domain adaptation method based on instance weighting for AD diagnosis, where the MRI images collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database are regarded as the source domain while MRI data collected from the Australian Imaging Biomarkers and Lifestyle Study of Aging database are regarded as the target domain [30]. All these studies indicate the effectiveness of the conventional TL for improving diagnostic performance.

However, both LUPI and conventional TL models cannot well handle the modality imbalance problem in clinical practice due to the limitation of their learning paradigm. To be specific, the LUPI can only be conducted on the paired modalities, whereas the remaining single-modal samples should be discarded. In this case, it brings the small-sample-size e issue, and the discarded samples could provide valuable information for improving the classification performance. In addition, the conventional TL methods assume that the source domain has much more data than the target domain to provide sufficient information for transfer. However, in clinical practice, the source domain usually has less data than the target domain, and the conventional TL cannot mine sufficient transferable information from the limited data in the source domain, thus degrading the learning performance. Moreover, the conventional TL methods do not make full use of label information to guide the transfer process. Generally, label information is expected to extract the inherent transferable information in both the target and source domains, and also reduce the discrepancy between the two domains.

To address the above modality imbalance issue in clinical application, we propose a new supervised TL method, which simultaneously transfers knowledge between both the paired data with shared labels and between the unpaired data with different labels.

B. Low-Rank Representation

Low-rank representation (LRR) aims to learn the lowest rank representation of all samples via a linear combination of bases in a given dictionary [31][32]. LRR has been successfully applied in different machine learning tasks, such as subspace clustering, classification, detection, semantic segmentation, and reconstruction [33].

LRR has also been introduced in TL in recent years. For example, Ding *et al.* developed a latent low-rank transfer subspace learning algorithm, which combined low-rank constraint and dictionary learning for knowledge transfer [34]; Xu *et al.* proposed a discriminative transfer subspace learning algorithm with low-rank and sparse constraints to reduce the disparity between the target and source domains [35]; Wang *et al.* proposed a class-specific reconstruction-based TL algorithm, in which the domain correlation was enhanced through a joint sparse and low-rank regularization with better block diagonal characteristic [36]; Wang *et al.* developed a multi-source domain adaption framework via LRR for multi-site ASD diagnosis based on fMRI [37]. These works demonstrate that LRR can effectively explore the intrinsic relationship in data to reduce the domain shift in TL tasks. However, these methods just evaluate the feature correlation between two domains in feature space, by ignoring the use of label information to guide knowledge transfer.

On the other hand, the above LRR-based TL methods generally adopt block-diagonal representation (BLR) to capture global semantic information between the target and source domains. BLR can effectively enlarge the intra-class discrimination and diminish the inter-class distance [38]. However, although the nuclear norm minimization for BLR is a convex problem with a global solution, it may over penalize large singular values, thus making the solution seriously deviate from the original solution [39].

To address these limitations, we propose a novel DSTC algorithm, in which LRR incorporates label knowledge into a classifier for TL between unpaired data. Furthermore, we implement BLR with the Schatten-*p* norm to achieve a more accurate recovery ability for the low-rank matrix [40].

III. METHODOLOGY

A. Notation

We define $\{\mathbf{X}_1, \mathbf{X}_2\}$ to be N_p paired multi-modal samples with shared labels, where $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times N_p}$ and $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times N_p}$ and D_1 and D_2 are the respective dimensions of features. Let $\mathbf{X}_1^* \in \mathbb{R}^{D_1 \times N_u}$ be additional N_u single-modal samples that have the same modality as $\mathbf{X}_1$. We further use $\{\mathbf{X}_1^*, \mathbf{X}_2\}$ to denote the unpaired data without shared labels. In our TL settings, we define $\mathbf{X}_2$ in the source domain while $\mathbf{X}_1$ and $\mathbf{X}_1^*$ in the target domain.

In binary classification, we further assume that the data are grouped according to categories, i.e., $N_p = [N_p^1, N_p^2]$ and $N_u = [N_u^1, N_u^2]$, where the superscripts 1 and 2 represent the positive and negative classes, respectively. The labels for the paired $\{\mathbf{X}_1, \mathbf{X}_2\}$ and $\mathbf{X}_1^*$ are denoted as $\mathbf{y} \in \mathbb{R}^{1 \times N_p}$ and $\mathbf{y}^* \in \mathbb{R}^{1 \times N_u}$, respectively. $\hat{\mathbf{y}}$ is the prediction of the paired data, and $\hat{\mathbf{y}}^*$ is the prediction of the unpaired data. Besides, we define augmented matrices of the feature as:

$$\begin{cases} \widetilde{\mathbf{X}}_1 \coloneqq \left[\mathbf{X}_1^{\boldsymbol{T}}, \mathbf{1}_{N_p}\right]^T \in \mathbb{R}^{(D_1+1) \times N_p} \\ \widetilde{\mathbf{X}}_2 \coloneqq \left[\mathbf{X}_2^{\boldsymbol{T}}, \mathbf{1}_{N_p}\right]^T \in \mathbb{R}^{(D_2+1) \times N_p} \\ \widetilde{\mathbf{X}}_1^* \coloneqq \left[\mathbf{X}_1^{*\boldsymbol{T}}, \mathbf{1}_{N_u}\right]^T \in \mathbb{R}^{(D_1+1) \times N_u} \end{cases}$$

where $\mathbf{1}_{N_p}$ denotes an all-one vector with N_p elements, $\mathbf{Q} \in \mathbb{R}^{N_p \times N_u}$ is the transformation matrix, and $\mathbf{w}_1 \in \mathbb{R}^{(D_1+1) \times 1}$ and $\mathbf{w}_2 \in \mathbb{R}^{(D_2+1) \times 1}$ are parameter vectors.

B. Framework of Doubly Supervised Transfer Classifier

Fig. 1 shows the framework of the proposed DSTC. For the paired data with shared labels, we transfer knowledge via the LUPI paradigm, while conducting additional TL for the unpaired data with different labels.

Our proposed DSTC incorporates the TL between both the paired and unpaired data into a unified framework, which can be formulated as follows:

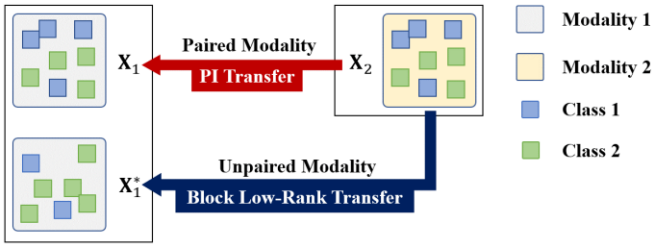

Fig. 1: Illustration of our proposed DSTC algorithm. $\mathbf{X}_1$ and $\mathbf{X}_1^*$ are the same modality worked in the target domain, while $\mathbf{X}_2$ is another modality in the source domain. $\mathbf{X}_1$ and $\mathbf{X}_2$ in the yellow ellipse form the paired modality with shared labels to transfer knowledge in the way of LUPI, while $\mathbf{X}_1^*$ and $\mathbf{X}_2$ in the blue ellipse build the unpaired modality with different labels to conduct additional TL.

$$L = L_{paried} + L_{unpaired} \tag{1}$$

where L_{paried} denotes the TL criterion for the paired data with shared labels, and $L_{unpaired}$ is the TL criterion for the unpaired data with different labels. The overall knowledge transfer mechanism of the proposed DSTC is shown in Fig. 1.

For L_{paried}, we aim to explore the inherent relation of shared labels between the paired data to guide the knowledge transfer. Several existing works indicate that SVM+ [13] can enhance the learning of the target classifier model through the guidance of the source domain by sharing label information. Thus, in this work, we introduce SVM+ to perform TL for the paired data with shared labels. The maximum margin criterion in SVM+ for paired data is given as:

$$L_{paried} = \min_{\mathbf{w}_1,\mathbf{w}_2} \frac{1}{2}(\|\mathbf{w}_1\|^2 + \gamma\|\mathbf{w}_2\|^2) + \lambda_1 {\mathbf{w}_2}^T \widetilde{\mathbf{X}}_2 \tag{2}$$
$$s.t.\ \mathbf{y}\odot\left({\mathbf{w}_1}^T\widetilde{\mathbf{X}}_1\right) \geq 1 - {\mathbf{w}_2}^T\widetilde{\mathbf{X}}_2$$

where $\mathbf{w}_1$ is the weight matrix for the classification hyperplane in the target domain, and $\mathbf{w}_2$ is the slack parameter matrix in the source domain; $\gamma > 0$ is the trade-off parameter, and λ_1 is the penalty parameter to balance the hinge loss term and regularizer term. Here, $\odot$ is the Hadamard product that performs element-wise multiplication.

Through the slack function ${\mathbf{w}_2}^T\widetilde{\mathbf{X}}_2$, the additional privileged information is introduced to regularize the hinge loss. Therefore, the classification hyperplane in the target domain can be tuned with the additional privileged information during the training stage, and then learning efficiency is further improved. [13].

For $L_{unparied}$, the distribution discrepancy between the source domain and target domain is minimized by finding an optimal transformation matrix. We assume that the knowledge can be propagated from both domains to the output space, and then the output of each target sample can be linearly reconstructed by those of the samples in the source domain, which is formulated as:

$$\hat{\mathbf{y}}^* = \hat{\mathbf{y}}\mathbf{Q} \tag{3}$$

where $\hat{\mathbf{y}}^* \coloneqq {\mathbf{w}_1}^T\widetilde{\mathbf{X}}_1^* \in \mathbb{R}^{\mathbf{1}\times N_u}$ and $\hat{\mathbf{y}} \coloneqq {\mathbf{w}_2}^T\widetilde{\mathbf{X}}_2 \in \mathbb{R}^{\mathbf{1}\times N_p}$.

C. DSTC via BLR with Schatten-p Norm

To make the relevant samples in both domains more interlaced than irrelevant samples, we propose to apply low-rank regularization to assume that the output of each sample in the target domain can be reconstructed by those of its neighbors in the source domain. Furthermore, to make the output space in the target domain sufficiently discriminative, it is natural that the transformation matrix should transfer one class in the source domain to that of the target domain. Thus, the BLR regularization is further utilized to enhance the discriminant of the target classifier.

Low-Rank Regularization. In our algorithm, the output of each sample in the target domain is reconstructed by those of its neighbors in the source domain. To capture the underlying correlations among the neighbors, the transformation matrix $\mathbf{Q}$ should be low-rank [41], which can be formulated as:

$$\min_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{Q}} \|\mathbf{Q}\|_* \tag{4}$$
$$s.t.\ {\mathbf{w}_1}^T\widetilde{\mathbf{X}}_1^* = {\mathbf{w}_2}^T\widetilde{\mathbf{X}}_2\mathbf{Q}$$

where $\|\cdot\|_*$ is the nuclear norm of a matrix.

Block-Diagonal Low-Rank Regularization. Although low-rank regularization is used to impose on $\mathbf{Q}$ reveals structure information, it cannot lead to a discriminative feature representation. We denote $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{N_u}] \in \mathbb{R}^{N_p\times N_u}$ as an ideal transformation matrix, if it can transform information considering the class information [42]. $\mathbf{q}_i \in \mathbb{R}^{N_p}$, $i = 1, \cdots, N_u$, is the reconstruction code for the output of the i-th target sample. If the i-th target sample belongs to the k-th class ($k = 1, \cdots, K$), the entries in $\mathbf{q}_i$ for this class should take nonzero values, while the others are all zeros. We further assume that the transformation matrix $\mathbf{Q}$ should be block-diagonal [38]. However, the absolute block-diagonal structure is not easy to learn. Therefore, it is expected that the off-block-diagonal entries in $\mathbf{Q}$ are as small as possible. To this end, we define:

$$\mathbf{A} = \mathbf{1}_{N_p\times N_u} - \begin{bmatrix} \mathbf{1}_{N_p^1}\mathbf{1}_{N_u^1}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{N_p^2}\mathbf{1}_{N_u^2}^T \end{bmatrix} \tag{5}$$

where $\mathbf{1}_N$ denotes an all-one vector with N elements and $\mathbf{0}$ denotes an all-zeros vector.

To achieve a block-wise structure of $\mathbf{Q}$ with minimal off-block-diagonal entries, we minimize the off-block-diagonal entries and preserve the block-diagonal entries in $\mathbf{Q}$ by minimizing:

$$\min_{\mathbf{Q}} \| \mathbf{A}\odot\mathbf{Q} \|_F^2 \tag{6}$$

By integrating both Eq. (4) and Eq. (6), the objective function of BLR is formulated as:

$$L_{BLR} = \min_{\mathbf{Q}} \lambda_2\|\mathbf{Q}\|_* + \frac{\lambda_3}{2}\| \mathbf{A}\odot\mathbf{Q} \|_F^2$$
$$+\frac{1}{2}\| {\mathbf{w}_1}^T\widetilde{\mathbf{X}}_1^* - \mathbf{y}^* \|_F^2 \tag{7}$$
$$\text{s.t. } {\mathbf{w}_1}^T\widetilde{\mathbf{X}}_1^* = {\mathbf{w}_2}^T\widetilde{\mathbf{X}}_2\mathbf{Q}$$

where $\| {\mathbf{w}_1}^T\widetilde{\mathbf{X}}_1^* - \mathbf{y}^* \|_F^2$ is the supervised term to minimize the training error in the target domain.

Block-Diagonal Low-Rank Regularization with Schatten-p norm. Compared with the nuclear norm, the Schatten-p norm could recover signals more accurately while keeping a weaker restricted isometric property [43]. The Schatten-p norm of a matrix $\mathbf{Q} \in \mathbb{R}^{N_p\times N_u}$ is defined as the l_p-norm of its singular values as follows:

$$\|\mathbf{Q}\|_{S_p} \triangleq \left(\sum_{i=1}^{\min(N_p,N_u)} \sigma_i^p(\mathbf{Q})\right)^{\frac{1}{p}} \tag{8}$$

where $\mathbf{Q} = \mathbf{U} \cdot Diag(\sigma(\mathbf{Q})) \cdot \mathbf{V}^T$ is the SVD with two orthogonal matrices $\mathbf{U} \in \mathbb{R}^{N_p \times N_u}$ and $\mathbf{U} \in \mathbb{R}^{N_u \times N_u}$ where N_u =min(N_p, N_u), $\sigma_i(\mathbf{Q})$ is the i-th entry of singular values vector, where $i = 1,2,\dots, min(N_p, N_u)$. It follows from Eq. (5) that the gap between rank function (i.e., $p = 0$) and nuclear norm (i.e., $p = 1$) can be bridged by setting $0 < p < 1$ [44][45]. More especially, when the rank number is relatively larger, the nonconvex Schatten-p norm can show its superiority over the nuclear norm for relaxing a matrix rank function.

By replacing the nuclear norm in the original low-rank model, the model of Schatten-p norm-based LRR is obtained:

$$\min_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{Q}} \|\mathbf{Q}\|_{S_p}^p \tag{9}$$

By replacing the nuclear norm in Eq. (7), the objective function of TL for the unpaired data $L_{unpaired}$ is formulated as:

$$\begin{aligned} L_{unpaired} = \min_{\mathbf{Q}} \lambda_2 \|\mathbf{Q}\|_{S_p}^p + \frac{\lambda_3}{2} \| \mathbf{A} \odot \mathbf{Q} \|_F^2 \\ + \frac{1}{2} \| \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* - \mathbf{y}^* \|_F^2 \\ \text{s.t. } \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* = \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 \mathbf{Q} \end{aligned} \tag{10}$$

The detailed TL strategy for unpaired data in DSTC is shown in Fig. 2. It can be found that DSTC also incorporates the label knowledge into the classifier for TL between the unpaired data with different labels.

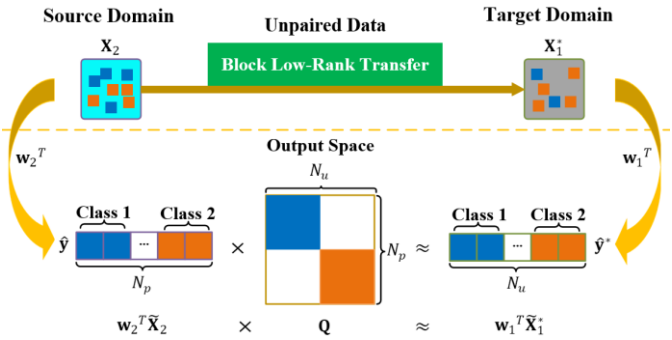

Fig. 2: Illustration of the unpaired TL strategy in DSTC. $\mathbf{X}_2$ and $\mathbf{X}_1^*$ are the source and target domains, respectively. They are separately transformed to the output space. Then, the BLR regularized TL is adopted to transfer knowledge and enhance the discriminative ability of the classifier in the target domain.

Combining L_{paried} and $L_{unpaired}$, we obtain the final formulation of the proposed DSTC as follows:

$$\begin{aligned} \boldsymbol{L} = \min_{\mathbf{w_1},\mathbf{w_2},\mathbf{Q}} \frac{1}{2}(\|\mathbf{w}_1\|^2 + \gamma\|\mathbf{w}_2\|^2) + \lambda_1 \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 \\ + \lambda_2 \|\mathbf{Q}\|_{S_p}^p + \frac{\lambda_3}{2} \| \mathbf{A} \odot \mathbf{Q} \|_F^2 + \frac{1}{2} \| \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* - \mathbf{y}^* \|_F^2 \\ s.t. \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* = \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 \mathbf{Q} \text{ and } \mathbf{y} \odot (\mathbf{w}_1{}^T \tilde{\mathbf{X}}_1) \geq \mathbf{1} - \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 \end{aligned} \tag{11}$$

where λ_1, λ_2, and λ_3 are trade-off parameters.

Based on Eq. (11), the LUPI mechanism in DSTC utilizes the shared labels of the paired data to maximize the margin between two classes. Meanwhile, DSTC also transfers knowledge between the two modalities of the unpaired data to further optimize the model. Minimizing $\|\mathbf{Q}\|_{S_p}^p$ guides to reveal the underlying structure of the data in both domains. Besides, minimization of $\| \mathbf{A} \odot \mathbf{Q} \|_F^2$ guides the off-block-diagonal entries in $\mathbf{Q}$ to be as small as possible, which makes the margin between classes in the target domain to be enlarged.

D. Optimization

The objective function in Eq.(11) is not jointly convex with respect to all variables. Thus, we utilize an alternating direction method to solve the problem efficiently. To achieve this, we first introduce two auxiliary variables $\mathbf{Z} = \mathbf{Q}$ and $\mathbf{E} = 1 - \mathbf{y} \odot (\mathbf{w}_1{}^T \tilde{\mathbf{X}}_1) - \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2$ to make the problem separable, and then construct the augmented Lagrangian function. The problem of Eq. (11) can be rewritten as:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_1, \mathbf{w}_2, \mathbf{Q}, \mathbf{Z}, \mathbf{E}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}(\|\mathbf{w}_1\|^2 + \gamma\|\mathbf{w}_2\|^2) \\ + \lambda_1 \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 + \lambda_2 \|\mathbf{Z}\|_{S_p}^p + \frac{\lambda_3}{2} \| \mathbf{A} \odot \mathbf{Q} \|_F^2 + \Phi(\boldsymbol{\beta}, \mathbf{Z} - \mathbf{Q}) \\ + (\mathbf{E})_+ + \frac{1}{2} \| \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* - \mathbf{y}^* \|_F^2 + \frac{1}{2} \| \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* - \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 \mathbf{Q} \|_F^2 \\ + \Phi(\boldsymbol{\alpha}, \mathbf{E} - 1 + \mathbf{y} \odot (\mathbf{w}_1{}^T \tilde{\mathbf{X}}_1) + \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2) \end{aligned} \tag{12}$$

where $(u)_+ := \max(u, 0)$ keeps the input scalar u unchanged if u is non-negative, and otherwise zero. The extension of vectors and matrices is simply applied element-wise. In addition, the $\Phi(\cdot)$ operation is defined as:

$$\Phi(\mathbf{M}, \mathbf{N}) = \frac{\mu}{2} + \frac{\lambda_2}{2} \|\mathbf{N}\|_F^2 + \langle \mathbf{M}, \mathbf{N} \rangle \tag{13}$$

where μ is a positive penalty scalar. $\boldsymbol{\alpha} \in \mathbb{R}^{1 \times N_p}$ and $\boldsymbol{\beta} \in \mathbb{R}^{N_p \times N_u}$ are Lagrangian multipliers.

1) Updating $\mathbf{w}_1$:

By fixing the irrelevant terms with respect to $\mathbf{w}_1$, and setting the derivative with respect to $\mathbf{w}_1$ to be zero, a closed-form solution for $\mathbf{w}_1$ can be given as:

$$\begin{aligned} \mathbf{w}_1 = \left(\mathbf{I} + 2\tilde{\mathbf{X}}_1^* \tilde{\mathbf{X}}_1^{*T} + \mu_1 \tilde{\mathbf{X}}_1 \tilde{\mathbf{X}}_1{}^T\right)^{-1} [\tilde{\mathbf{X}}_1^* \mathbf{Q}^T \tilde{\mathbf{X}}_2{}^T \mathbf{w}_2 + \tilde{\mathbf{X}}_1^* \mathbf{y}^{*T} \\ - \tilde{\mathbf{X}}_1 (\mathbf{y} \odot \boldsymbol{\alpha}^T) - \mu_1 \tilde{\mathbf{X}}_1 \left(\mathbf{y} \odot (\mathbf{E} - \mathbf{1} + \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2)^T\right)] \end{aligned} \tag{14}$$

where μ_1 is a positive penalty scalar.

2) Updating $\mathbf{w}_2$:

Similar to $\mathbf{w}_1$, with other variables fixed, the model is differentiable to $\mathbf{w}_2$, the solution for $\mathbf{w}_2$ can be derived as:

$$\begin{aligned} \mathbf{w}_2 = \left(\gamma \mathbf{I} + 2\tilde{\mathbf{X}}_2 \mathbf{Q} \mathbf{Q}^T \tilde{\mathbf{X}}_2{}^T\right)^{-1} [\tilde{\mathbf{X}}_2 (-\boldsymbol{\alpha} - \lambda_1)^T \\ - 2\tilde{\mathbf{X}}_2 \mathbf{Q} \tilde{\mathbf{X}}_1^{*T} \mathbf{w}_1 - \mu_1 \tilde{\mathbf{X}}_2 \left(\mathbf{E} - 1 + (\mathbf{y} \odot \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1)^T\right)] \end{aligned} \tag{15}$$

3) Updating $\mathbf{E}$:

Picking out the terms related to $\mathbf{E}$, we seek the minimum of each element in E and get the following updating formula:

$$\begin{aligned} \mathbf{E} = \boldsymbol{\Omega} \odot \left(1 - \mathbf{y} \odot (\mathbf{w}_1{}^T \tilde{\mathbf{X}}_1) - \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2\right) \left(1 + \frac{2\boldsymbol{\alpha}}{\mu_1}\right) \\ + \bar{\boldsymbol{\Omega}} \odot \left(\mathbf{1} - \mathbf{y} \odot \langle \mathbf{w}_1, \tilde{\mathbf{X}}_1 \rangle - \langle \mathbf{w}_2{}^T, \tilde{\mathbf{X}}_2 \rangle\right) \end{aligned} \tag{16}$$

where $\boldsymbol{\Omega}$ is an indicator matrix which is computed as

$$\boldsymbol{\Omega} = \left(\mathbf{1} - \mathbf{y} \odot (\mathbf{w}_1{}^T \tilde{\mathbf{X}}_1) - \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 - \frac{\boldsymbol{\alpha}}{\mu_1}\right) > 0 \tag{17}$$

4) Updating $\mathbf{Q}$:

By dropping those terms without $\mathbf{Q}$ and setting the derivative with $\mathbf{Q}$ as zero, the solution for $\mathbf{Q}$ can be computed as:

$$\begin{aligned} \mathbf{Q} = \left(\tilde{\mathbf{X}}_2{}^T \mathbf{w}_2 \mathbf{w}_2{}^T \tilde{\mathbf{X}}_2 + \mu_2 \mathbf{I} + \lambda_3 \mathbf{A} \mathbf{A}^T\right)^{-1} \\ (\tilde{\mathbf{X}}_2{}^T \mathbf{w}_2 \mathbf{w}_1{}^T \tilde{\mathbf{X}}_1^* + \mu_2 \mathbf{Z} - \boldsymbol{\beta}) \end{aligned} \tag{18}$$

where μ_2 is a positive number.

5) Updating $\mathbf{Z}$:

Solving the Schatten-p norm of a matrix usually involves the

SVD of heavy computation at each iteration, limiting its application in large-scale problems. Inspired by the matrix factorization strategy for nuclear norm [46], Schatten-*1/2* norm and Schatten-*2/3* norm [47][48] are bounded by the Bi-Frobenius norm, Bi-nuclear norm $\|\cdot\|_{BiN}$ and Forbenius/Nuclear hybrid norm as $\|\cdot\|_{F/N}$, respectively.

For matrix $\mathbf{Z} \in \mathbb{R}^{N_p \times N_u}$ with $rank(\mathbf{Z}) = r \leq d$ (i.e., the upper bounding of the $rank(\mathbf{Z})$). we can factorize Z into two smaller matrices $\mathbf{U} \in \mathbb{R}^{N_p \times d}$ and $\mathbf{V} \in \mathbb{R}^{N_u \times d}$ such that $\mathbf{Z} = \mathbf{U}^T\mathbf{V}$. The corresponding alternative formulation can be formulated as:

$$\begin{cases} p = 1: \min\limits_{\mathbf{Z},\mathbf{U},\mathbf{V}} \frac{\lambda_2}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \Phi(\boldsymbol{\beta}, \mathbf{Z} - \mathbf{Q}) \\ \qquad \text{s.t. } \mathbf{Z} = \mathbf{U}^T\mathbf{V} \\ p = \frac{1}{2}: \min\limits_{\mathbf{Z},\mathbf{U},\mathbf{V}} \frac{\lambda_2}{2}(\|\mathbf{U}\|_* + \|\mathbf{V}\|_*) + \Phi(\boldsymbol{\beta}, \mathbf{Z} - \mathbf{Q}) \\ \qquad \text{s.t. } \mathbf{Z} = \mathbf{U}^T\mathbf{V} \\ p = \frac{2}{3}: \min\limits_{\mathbf{Z},\mathbf{U},\mathbf{V}} \frac{\lambda_2}{3}(2\|\mathbf{U}\|_* + \|\mathbf{V}\|_F^2) + \Phi(\boldsymbol{\beta}, \mathbf{Z} - \mathbf{Q}) \\ \qquad \text{s.t. } \mathbf{Z} = \mathbf{U}^T\mathbf{V} \end{cases} \quad (19)$$

By solving the Eq. (18) with different choices of p, we can update $\mathbf{Z}$. As p=2/3 is the combination of $p = 1$ and p=1/2, we give the detailed solving procedure of p=2/3.

When p=2/3, let $\mathbf{M} = \mathbf{U}$. Updating $\mathbf{U}, \mathbf{V}, \mathbf{M}, \mathbf{Z}$ by solving:

$$\min_{\mathbf{U}} \Phi(\mathbf{S}_1, \mathbf{M} - \mathbf{U}) + \Phi(\mathbf{S}_2, \mathbf{U}^T\mathbf{V} - \mathbf{Z}) \quad (20)$$

$$\min_{\mathbf{V}} \frac{\lambda_2}{3}\|\mathbf{V}\|_F^2 + \Phi(\mathbf{S}_2, \mathbf{U}^T\mathbf{V} - \mathbf{Z}) \quad (21)$$

$$\min_{\mathbf{M}} \frac{2\lambda_2}{3}\|\mathbf{M}\|_* + \Phi(\mathbf{S}_1, \mathbf{M} - \mathbf{U}) \quad (22)$$

$$\min_{\mathbf{Z}} \Phi(\boldsymbol{\beta}, \mathbf{Z} - \mathbf{Q}) + \Phi(\mathbf{S}_2, \mathbf{U}^T\mathbf{V} - \mathbf{Z}) \quad (23)$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are Lagrangian multipliers.

Thus, $\mathbf{U}, \mathbf{V}$ can be updated by:

$$\mathbf{U} = (\mathbf{M} + \mathbf{S}_1 + \mathbf{Z}\mathbf{V}^T - \mathbf{S}_2\mathbf{V}^T)^{-1}(\mathbf{I} + \mathbf{V}^T\mathbf{V}) \quad (24)$$

$$\mathbf{V} = \left(\frac{2\lambda_2}{3} + \mathbf{U}^T\mathbf{U}\right)^{-1}(\mathbf{U}\mathbf{Z} - \mathbf{S}_2\mathbf{U}^T) \quad (25)$$

Using the singular value thresholding (SVT) algorithm [49], the optimal $\mathbf{M}$ can be derived as:

$$\mathbf{M} = \mathcal{D}_{\frac{\lambda_1}{\mu_2}}(\mathbf{U} - \frac{\mathbf{S}_1}{\mu_2}) \quad (26)$$

where $\mathcal{D}_\tau(\cdot)$ is the singular value shrinkage operator. Given a matrix $\mathbf{P}$, the singular value decomposition (SVD) of matrix $\mathbf{P}$ is performed as $\mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma} = diag(\sigma_i)$, the operator can be computed by:

$$\mathcal{D}_\tau(\mathbf{P}) = \mathbf{U}\mathcal{D}_\tau(\boldsymbol{\Sigma})\mathbf{V}^T, \mathcal{D}_\tau(\boldsymbol{\Sigma}) = diag((\sigma_i - \tau)_+) \quad (27)$$

Finally, $\mathbf{Z}$ can be updated by

$$\mathbf{Z} = (\mathbf{Q} + \mathbf{S}_2 + \mathbf{U}^T\mathbf{V} - \boldsymbol{\beta}) \quad (28)$$

6) Updating α, β:

The multiplier $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are updated by:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha} + \theta(\mathbf{E} - 1 + \mathbf{y} \odot (\mathbf{w}_1{}^T\widetilde{\mathbf{X}}_1) + \mathbf{w}_2{}^T\widetilde{\mathbf{X}}_2, \quad (29)$$

$$\boldsymbol{\beta} = \boldsymbol{\beta} + \mu(\mathbf{Z} - \mathbf{Q}) \quad (30)$$

where $\theta = \min(\rho\theta, \theta_{max})$, $\mu = \min(\rho\mu, \mu_{max})$, and ρ is the learning rate.

The detailed procedure of the proposed DSTC is summarized in Algorithm 1.

Algorithm 1: Learning procedure of DSTC

Input: paired bimodal data $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{y}\}$, the single-modal data $\{\mathbf{X}_1^*, \mathbf{y}^*\}$;
Output: Parameters of target classifier: $\mathbf{w}_1$;
Construct the objective function $\boldsymbol{L}$ using (11);
for j =1, 2, …, T
  Update $\mathbf{w}_1$ according to Eq. (14);
  Update $\mathbf{w}_2$ according to Eq. (15);
  Compute $\mathbf{E}$ according to Eq. (16);
  Update $\mathbf{Q}$ according to Eq. (18);
  Compute $\mathbf{Z}$ according to Eq. (28);
  Update $\boldsymbol{\alpha}$ according to Eq. (29);
  Update $\boldsymbol{\beta}$ according to Eq. (30);
  if $\left|\boldsymbol{L}^{(j)} - \boldsymbol{L}^{(j-1)}\right| < \varepsilon$ **then**
    Go to **Output**;
  end if
end
return solution

IV. EXPERIMENTS AND RESULTS

A. Datasets and Data Preprocessing

The proposed DSTC algorithm was evaluated on two datasets, namely the ADNI dataset [50] and a bimodal breast ultrasound image (BBUI) dataset [51].

The used ADNI database includes 360 subjects (85 AD, 185 MCI, and 90 NC) with paired MRI and PET data, and 377 subjects (86 AD, 177 MCI, and 114 NC) with only MRI images. All the MRI images were scanned by the 1.5T devices. We extracted the region of interest (ROI) based features after the following image preprocessing on MRI data [52], including anterior commissure-posterior commissure (AC-PC) correction, intensity inhomogeneity correction by N3 algorithm [53], skull-stripping and removal of cerebellum with the algorithms by Wang et al. [54]. All images were then segmented into three tissues, i.e., grey matter, white matter and cerebrospinal fluid, by the FAST algorithm [55]. After an MRI was registered to a brain template with 93 manually labeled ROIs by HAMMER [56], the volumes of gray matter tissue were then calculated as a feature for each ROI. Thus, there were a total of 93 features corresponding to 93 ROIs. A PET image was then aligned to its corresponding MRI by a rigid registration, and then the average intensity value of each ROI was computed as a feature. Consequently, we finally extracted 93-dimensional features from MRI and PET images, respectively. Note that all the samples used in this work were baseline visits. Please refer to [57][58][59] for more details on the feature extraction of MRI and PET images, respectively.

The BBUI dataset was acquired from the Nanjing Drum Tower Hospital. It includes 106 pairs of bimodal ultrasound images from 52 benign tumor patients and 54 malignant tumor patients, and additional 158 single-modal BUS images from 77 benign tumor patients and 81 malignant tumor patients. The approval from the ethics committee of the hospital was obtained, and all patients had signed informed consent. All lesions were underwent biopsy and pathologically proven. Both the BUS and

EUS images were simultaneously scanned by the Mindary Resona7 ultrasound scanner with the L11-3 linear-array probe. A rectangle ROI, including the tumor region, was selected by an experienced sinologist from each ultrasound image. The statistical feature descriptors were calculated from the intensities of all pixels, including the mean, standard deviation, coefficient of variance, skewness, kurtosis, the entropy of histogram, area ratio, combined area ratio, and several percentiles. The texture features were extracted from the gray-level co-occurrence matrix (GLCM), including the energy, contrast, homogeneity, and entropy of GLCM. Moreover, the Hu moment invariants were also extracted as features. A total of 71-dimensional features were thus generated from each ROI in both BUS and EUS images. Please refer to reference [60] for more details about feature extraction.

In the experiment, we selected the widely used BUS as the target domain and EUS as the source domain in the BBUI dataset, and MRI as the target domain and PET as the source domain in the ADNI dataset.

B. Experimental Setup

To validate the effectiveness of the proposed DSTC, we compare it with the following related algorithms:

1) SVM: The widely used SVM algorithm was performed on MRI without TL.
2) LRR-SVM: It is a two-stage learning algorithm, which first trains the LRR model to generate a shared feature representation from both the source and target domains, and then feeds the new features to SVM for classification.
3) SVM+ [14]: The fast SVM+ algorithm, an improved version of the original SVM+, was compared as a baseline.
4) PMT-SVM [61]: It is a widely used TL SVM based on the projective model.
5) CT-SVM [62]: It is a TL-based SVM that incorporates correlation regularization for cross-domain recognition.
6) DTSL-LRSR [35]: It learns a discriminative transfer subspace via low-rank and sparse representation.
7) GSL [63]: It is a guide subspace learning-based TL algorithm that learns an invariant, discriminative, and domain agnostic subspace by subspace guidance, low-rank-based data guidance, and label guidance.
8) CRTL [36]: It combines low-rank and sparse constraints on the class-specific reconstruction coefficient matrix to preserve global and local data structures.
9) LSDT [64]: It is a reconstruction-based TL algorithm, which learns a sparse reconstruction coefficient matrix between the target and source domains in some latent space for domain adaptation.
10) MCTL [65]: It is a manifold criterion guided TL algorithm, which aims to learn a latent common subspace via a projection matrix for target and source domains.

All above algorithms perform different kinds of TL except SVM, among which SVM+, PMT-SVM, and CT-SVM conduct the classifier-level TL, while LRR-SVM, DSTL-LRSR, GSL, CRTL, LSDT, and MCTL are the low-rank-based feature-level TL algorithms.

We further conducted the ablation experiments to evaluate the effectiveness of DSTC:

1) DSTC-LR: It is performed low-rank regularization on the proposed DSTC for ablation study.
2) DSTC-BLR: It is performed BLR regularization on the proposed DSTC for ablation study.
3) DSTC-BLR$_{sp}$: We use DSTC-BLR$_{sp}$ to distinguish the proposed DSTC and the abovementioned ablation studies.

The five-fold nested cross-validation strategy was applied to all algorithms on both datasets [66]. The inner loop was repeated three times to find optimal hyperparameters, and the outer loop was repeated five times to evaluate the performance of the model. All the compared algorithms and DSTC used the same training/testing split on two datasets. The commonly used classification accuracy (ACC), sensitivity (SEN), specificity (SPE), Youden index (YI), positive predictive value (PPV), and negative predictive value (NPV) were selected as evaluation indices. The receiver operating characteristic (ROC), and the area under the curve (AUC) were also selected as evaluation indices. The results were reported in the format of mean ± SD (standard deviation).

We adopted a grid search strategy to search the optimal hyperparameters across a wider range. Table I shows the searching strategy ranges of all the parameters in all algorithms.

TABLE I
THE SEARCHING STRATEGY RANGE OF PARAMETERS

Algorithms	The searching strategy range of parameters
SVM	C: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}
LRR-SVM	C: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}
SVM+	C: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}; γ: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}
PMT-SVM	τ: {10^{-5}, 10^{-4}, …, 10^{3}, 10^{4}, 10^{5}}
CT-SVM	ρ: {0, 0.2, 0.4, 0.6, 0.8, 1}; a: {1, 5, 10, 15, 20}; C: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}
DTSL-LRSR	α: {10^{-4}, 10^{-3}, …, 10^{1}, 10^{2}, 10^{3}}; β: {10^{-4}, 10^{-3}, …, 10^{1}, 10^{2}, 10^{3}};
GSL	α:{10^{-4}, 10^{-3}, …, 10^{1}, 10^{2}, 10^{3}}; β: {10^{-4}, 10^{-3}, …, 10^{1}, 10^{2}, 10^{3}}; λ: {10^{-4}, 10^{-3}, …, 10^{1}, 10^{2}, 10^{3}} C: {10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}}
CRTL	σ: {0.1, 0.2, 0.3, …, 1.8, 1.9, 2.0}
LSDT	λ_1: { 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}} λ_2: { 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}}
MCTL	τ: {0, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}}; λ_1: {10^{-4}, 10^{-3}, …, 10^{1}, 10^{2}, 10^{3}}
DSTC	γ: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}; λ_1: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}; λ_2: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}; λ_3: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5}

C. Comparisons with Different p

We conduct the following experiments to compare the classification performance of our DSTC under the different choices of parameter p, with p=1 (nuclear norm), p=1/2, and p=2/3. The results are shown in Fig. 3.

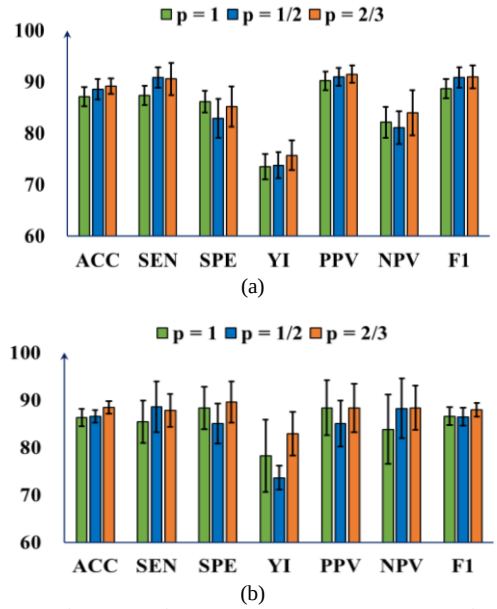

Fig. 3: The classification performance with varied parameter p of the proposed DSTC on the task of (a) AD vs. NC, (b) BUS-based breast cancer classification.

The best performance for both AD and breast cancer is achieved by p=2/3, since it obtains a better balance between enforcing low-rank and separating sparse outliers. Thus, we fix parameter p=2/3 throughout the experiments in this paper.

D. Results on MRI-based AD Classification

Table II shows results comparison of different algorithms for the AD classification task. From Table II, it can be observed that the proposed DSTC algorithm outperforms all the other compared algorithms, indicating its effectiveness. Moreover, it can be obtained that SVM+ achieves better classification performance than LRR-SVM, which indicates the superiority of the LUPI paradigm. Specifically, DSTC achieves the best performance with the best mean classification accuracy of 90.14±1.37%, sensitivity of 91.34±3.09%, specificity of 87.17±3.54%, YI of 78.51±3.23%, F1 score of 91.00±2.23%, PPV of 90.68±1.78%, and NPV of 88.16±4.97%, which improves at least 4.69%, 2.87%, 2.49%, 7.92%, 3.87%, 3.51%, and 3.8% on the corresponding indices over the other TL algorithms. This is mainly because that the proposed DSTC makes full use of the shared labels to effectively guide knowledge transfer and transfer auxiliary knowledge between unpaired data with different labels.

Table III shows the results of ablation experiments for AD classification. We can see that DSTC-BLR improves of 1.52%, 2.05%, 2.17%, 1.89%, 1.59%, and 1.04% on accuracy, sensitivity, YI, F1 score, PPV, and NPV, respectively, over DSTC-LR. It indicates that the BLR helps to promote the discriminative ability of different classes with improved performance of DSTC-BLR. Moreover, DSTC-BLR$_{SP}$ achieves the improvements of 1.54%, 0.8%, 2.07%, 2.87%, 1.75%, 2.57%, and 1.34% on accuracy, sensitivity, specificity, YI, F1 score, PPV, and NPV, respectively, over DSTC-BLR. It further indicates the effectiveness of the Schatten-p norm in recovering the low-rank matrix.

Fig. 4 shows ROC curves and the corresponding AUC values for different algorithms. The proposed DSTC-BLR$_{SP}$ algorithm achieves the best AUC value of 0.960, which further indicates its effectiveness.

TABLE II
CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS FOR AD CLASSIFICATION (UNIT: %)

	ACC	SEN	SPE	YI	F1	PPV	NPV
SVM	80.87±1.75	81.54±5.99	79.85±7.15	61.39±3.97	80.70±4.21	80.35±6.65	80.48±8.61
LRR-SVM	81.88±2.68	82.16±4.72	80.83±6.34	62.99±6.12	82.03±4.41	82.13±6.07	80.66±5.05
SVM+	82.88±2.72	82.94±5.55	83.27±7.45	66.21±7.57	82.51±4.46	82.53±7.21	81.58±9.90
PMT-SVM	83.30±3.54	83.24±5.80	82.46±6.14	65.70±7.53	83.79±5.04	84.61±6.37	80.39±9.05
CT SVM	83.68±3.39	82.97±5.74	83.63±5.02	66.60±6.87	83.93±5.18	85.10±6.23	81.05±6.36
DTSL-LRSR	84.54±4.19	83.61±5.35	84.97±5.08	68.58±8.58	84.81±5.53	86.19±6.84	81.05±8.23
GSL	83.68±2.97	81.94±4.73	84.68±4.02	66.62±5.70	83.74±4.89	85.78±6.16	79.75±6.36
CRTL	84.61±3.78	85.91±5.25	81.93±6.80	67.84±7.68	86.40±3.88	87.17±5.10	80.34±7.75
LSDT	85.27±1.83	86.65±3.42	83.94±4.55	70.59±3.23	85.64±3.19	85.04±6.44	84.36±5.72
MCTL	85.45±4.02	88.47±7.35	81.05±6.34	69.52±7.84	87.13±4.01	86.33±4.63	84.05±10.76
DSTC	**90.14±1.37**	**91.34±3.09**	**87.17±3.54**	**78.51±3.23**	**91.00±2.23**	**90.68±1.78**	**88.16±4.97**

TABLE III
ABLATION EXPERIMENT RESULTS FOR AD CLASSIFICATION (UNIT: %)

	ACC	SEN	SPE	YI	F1	PPV	NPV
DSTC-LR	87.08±2.84	88.49±4.05	84.98±3.84	73.47±5.20	87.36±4.23	86.52±6.35	85.78±7.75
DSTC-BLR	88.60±2.45	90.54±2.81	85.10±3.79	75.64±4.86	89.25±4.15	88.11±5.94	86.82±6.96
DSTC-BLR$_{SP}$	**90.14±1.37**	**91.34±3.09**	**87.17±3.54**	**78.51±3.23**	**91.00±2.23**	**90.68±1.78**	**88.16±4.97**

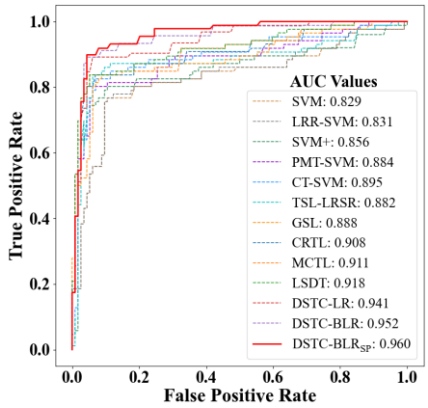

Fig. 4: ROC curves and AUC values of different algorithms for MRI-based AD classification.

E. Results on BUS-based Breast Cancer Classification

Table IV gives the classification results of different compared algorithms on the BUS-based CAD for breast cancer with EUS as the source domain. Similar to Table II, DSTC again outperforms all the compared algorithms, with the best mean classification accuracy of 88.41±1.33%, sensitivity of 87.90±4.03%, specificity of 89.22±4.09%, YI of 77.12±3.19%, F1 of 86.58±2.18%, PPV of 85.69±5.42%, and NPV of 90.54±3.89%. DSTC achieves significant improvements over the baseline SVM on all the indices. Moreover, DSTC improves at least 3.52%, 4.28%, 6.12%, 1.51%, and 2.39% on classification accuracy, specificity, YI, F1 score, and NPV over the other TL algorithms. It indicates that DSTC can improve classification performance by simultaneously transferring knowledge between both the paired data with shared labels and the unpaired data with different labels.

Table V shows the results of ablation experiments for breast cancer classification. It can be observed that DSTC-BLR improves 0.87%, 0.81%, 1.28%, 1.14%, 1.22%, and 0.79% on classification accuracy, sensitivity, YI, F1 score, PPV, and NPV, respectively, over DSTC-LR, which suggests that BLR can enhance the discriminative ability of classifier. Moreover, DSTC-BLR$_{SP}$ obtains the improvements of 1.85%, 0.84%, 2.74%, 3.58%, and 3.15%, respectively, on accuracy, sensitivity, specificity, YI, and NPV, respectively, over DSTC-BLR, which further indicates the of the effectiveness of the Schatten-p norm for low rank matrix.

Fig. 5 shows ROC curves and the corresponding AUC values for different and ablation algorithms. The proposed DSTC-BLR$_{SP}$ algorithm again achieves the best AUC value of 0.928.

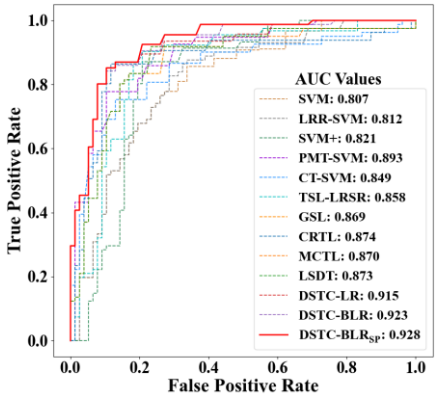

Fig. 5: ROC curves and AUC values of different algorithms for BUS-based breast cancer classification.

TABLE IV
CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS FOR BREAST TUMOR CLASSIFICATION (UNIT: %)

	ACC	SEN	SPE	YI	F1	PPV	NPV
SVM	76.79±2.77	77.40±6.15	77.32±7.87	54.72±6.94	76.71±3.56	77.03±8.66	77.03±6.69
LRR-SVM	78.02±3.50	77.39±5.88	79.50±7.79	56.89±7.95	77.99±3.99	79.45±8.29	76.92±6.46
SVM+	79.98±2.67	86.06±8.25	76.14±10.53	62.23±4.88	80.21±3.22	76.85±10.48	85.40±8.77
PMT-SVM	80.76±2.92	82.97±5.31	79.45±8.79	62.43v6.90	81.35±2.79	80.74±8.48	81.27±6.76
CT SVM	79.90±2.61	82.68±2.95	77.30±6.28	59.98±6.08	80.85±2.70	79.48±6.24	80.23±4.80
DTSL-LRSR	81.65±1.75	84.48±4.35	79.25±1.79	63.74±3.99	81.96±2.40	79.93±4.65	83.17±6.06
GSL	83.50±2.51	85.08±5.13	83.01±7.32	68.09±5.31	82.69±3.27	81.30±8.56	85.95±5.67
CRTL	83.52±2.35	85.66±5.06	82.68±7.12	68.34±5.06	83.14±3.12	81.67±8.65	85.50±6.75
LSDT	84.89±2.04	87.53±4.45	83.47±5.58	71.00±4.08	83.79±2.57	81.07±7.46	88.15±6.67
MCTL	84.64±1.40	85.14±4.58	84.94±4.85	70.08±3.30	85.07±1.79	85.55±5.77	83.71±6.22
DSTC	**88.41±1.33**	**87.90±4.03**	**89.22±4.09**	**77.12±3.19**	**86.58±2.18**	**85.69±5.42**	**90.54±3.89**

TABLE V
ABLATION EXPERIMENT RESULTS FOR BREAST TUMOR CLASSIFICATION (UNIT: %)

	ACC	SEN	SPE	YI	F1	PPV	NPV
DSTC-LR	85.69±2.20	86.25±4.99	86.01±6.20	72.26±4.12	84.97±2.46	84.42±6.98	86.60±7.37
DSTC-BLR	86.56±1.37	87.06±4.78	86.48±4.01	73.54±2.51	86.11±1.68	85.64±4.73	87.39±5.84
DSTC-BLR$_{SP}$	**88.41±1.33**	**87.90±4.03**	**89.22±4.09**	**77.12±3.19**	**86.58±2.18**	**85.69±5.42**	**90.54±3.89**

F. Convergence Analysis

Fig. 6 shows the convergence property of proposed DSTC-BLR$_{SP}$, DSTC-BLR, and DSTC-LR on both AD and breast classification. We can see that the objective function of the proposed DSTC decreases when the iteration number increases. It can also be seen that the proposed algorithms converge within 20 iterations on both classification tasks. Due to the inexact solution of DSTC-LR and DSTC-BLR, the objective function of DSTC-LR does not monotonically decrease. However, it still decreases sharply, indicating that the proposed DSTC has a good convergence property.

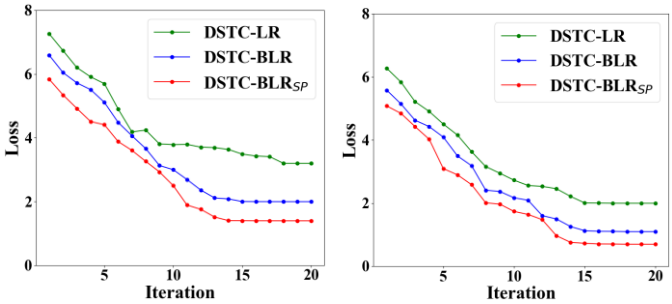

Fig. 6: Converge curve of the proposed DSTC-BLR and DSTC-LR on the task of (a) AD vs. NC. (b) BUS-based breast cancer classification.

V. Discussion

In this work, we propose a novel DSTC algorithm to solve the clinical problem of modality imbalance. We first integrate the SVM+ classifier and LRR into a unified framework, and then introduce the Schatten-p norm to achieve a tighter approximation of the rank function. The experimental results on the ADNI dataset and BBUI dataset indicate the effectiveness of the proposed DSTC.

In clinical practice, single-modal imaging-based CAD generally has wide and more flexible applications than multi-modal methods. Extensive studies demonstrate that TL can effectively improve the single-modal imaging-based CAD by transferring knowledge from other related imaging modalities or diseases [5]. However, existing TL methods, such as LUPI and the conventional TL, still have some limitations in solving the special clinical issue of modality imbalance. For example, the LUPI paradigm can only handle the paired data with shared labels, and the remaining unpaired data should be discarded [16][17][18][19]. In fact, the discarded samples may also provide valuable information for improving classification performance. On the other hand, the conventional TL algorithms generally require sufficient training samples in the source domain to provide enough transferable knowledge [9]. However, in clinical practice, the commonly used modalities are always adopted as the target domain, and the source domain generally cannot provide enough data when the source domain modalities are not completely popularized. This clinical phenomenon makes knowledge transfer between the source and target domains more difficult.

Moreover, the samples in both domains are commonly labeled, but the conventional TL algorithms automatically aligned domain divergence as the unsupervised domain adaption according to the domain adaption theory [20]. Thus, these algorithms rarely utilize this label information to reduce the discrepancy between the two domains so as to guide the transfer process [28][30]. The proposed DSTC integrates the SVM+ classifier and LRR into a unified framework, which can perform knowledge transfer between both paired data and unpaired data. Compared with existing TL methods, DSTC is more flexible for applications and can address the modality imbalance problem to a certain extent. It is worth noting that if there are also some single-modal imaging data in the source domain, the proposed can well handle this scenario. The additional single-model imaging data in the source domain only increases the numbers of the unpaired data, and does not change the training procedure of the DSTC algorithm.

Although LRR can shrink the unfavorable representation from off-block-diagonal elements, and thus obtain more discriminative representation, it may overpenalize large singular values for solving the nuclear norm minimization problem [39]. To this end, the label knowledge is incorporated into the LRR in classifier for TL between the unpaired data with different labels, which is different from the previous LRR-based TL methods that evaluate the feature correlation between the source and target domains in feature space. Besides, the Schatten-p norm minimization with small p-values requires significantly fewer measurements [67]. Furthermore, although there are no shared labels for the unpaired data, The Schatten-p norm ensures that the transformation matrix should transfer one class in the source domain to that of the target domain.

The proposed DSTC still has some room for improvement. It is known that deep learning has achieved great success in the field of medical image analysis. In our previous work [51], both the maximum mean discrepancy criterion-based feature-level TL in convolutional neural network (CNN) and the SVM+ classifier are integrated into the same framework. Thus, the proposed DSTC can be embedded into the deep learning models to conduct both the feature- and classifier-level knowledge transfer simultaneously, which is sure to further improve the transfer performance. Moreover, the softmax-based LUPI algorithm should be studied to replace SVM+ in DSTC in the future, which can handle the multi-class classification for more CAD tasks, and also can be easily integrated into the CNN models.

VI. Conclusion

This work proposes a novel DSTC algorithm that integrates the SVM+ classifier and LRR into a unified framework. The experimental results indicate that the proposed DSTC outperforms all the compared algorithms, including the conventional SVM+ and TL classifiers, and the state-of-the-art low-rank regularized TL algorithms. These comparison results also suggest that the proposed doubly supervised TL method has the potential in more applications for medical imaging-based CAD.

ACKNOWLEDGMENT

The authors would like to thank Mrs. Weijun Zhou with the First Affiliated Hospital of Nanjing University Medical School for providing the BBUI dataset in our experiments.

REFERENCES

[1] D. Shen, G. Wu, and H.I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221-248. 2017.

[2] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60-88., 2017.

[3] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 173-183, 2017.

[4] B. Jie, D. Zhang, B. Cheng, and D. Shen, "Manifold regularized multitask feature learning for multimodality disease classification," *Hum. Brain Mapp.*, vol. 36, no. 2, pp. 489-507, 2015.

[5] H. Hermessi, O. Mourali, and E. Zagrouba, "Multimodal medical image fusion review: Theoretical background and recent advances," *Signal Process.*, vol. 183, no. 108036, 2021.

[6] S. Liu *et al.*, "Multimodal neuroimaging feature learning for multi-class diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132-40, 2014.

[7] Y. Fan *et al.*, "Multivariate examination of brain abnormality using both structural and functional MRI," *NeuroImage*, vol. 36, no. 4, pp. 1189-1199, 2007.

[8] Y. Fan *et al.*, "Unaffected family members and schizophrenia patients share brain structure patterns: a high-dimensional pattern classification study," *Biol. Psychiatry*, vol. 63, no. 1, pp. 118-124, 2008.

[9] S. J. Pan, and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, 2010.

[10] V. Cheplygina, M. de Bruijne, J.P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal*. vol. 54, pp. 280-296, 2019.

[11] B. Cheng, M. Liu, D. Shen, Z. Li, and D. Zhang, "Multi-domain transfer learning for early diagnosis of Alzheimer's disease," Neuroinform., vol. 15, no. 2, pp.115-132, 2017.

[12] X. Fei *et al.*, "Parameter transfer deep neural network for single-modal B-mode ultrasound-based computer-aided diagnosis," *Cognit. Comput.*, vol. 12, no. 6, pp. 1252-1264, 2020.

[13] V. Vapnik, and A. Vasjist, "A new learning paradigm: Learning using privileged information," in *Proc. Int. Joint Conf. Neural Network (IJCNN)*, 2009, pp. 544-557.

[14] W. Li, D. Dai, M. Tan, D. Xu, and L. V. Gool, "Fast algorithms for linear and kernel SVM+," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2258-2266.

[15] P.B. Zhang, and Z.X. Yang, "A new learning paradigm for random vector functional-link network: RVFL+," *Neural Netw.*, vol. 122, pp. 94-105, 2020.

[16] X. Zheng, J. Shi, S. Ying, Q. Zhang, and Y. Li, "Improving single-modal neuroimaging based diagnosis of brain disorders via boosted privileged information learning framework," in *Int. Workshop Machine Learning Med. Imaging (MLMI)*, 2016, pp. 95-103.

[17] L. Duan *et al*., "Incorporating privileged genetic information for fundus image based glaucoma detection," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2014, pp. 204-211.

[18] J. Shi *et al*., "Cascaded multi-column RVFL+ classifier for single-modal neuroimaging-based diagnosis of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 66, pp. 2362-2371, 2019.

[19] Y. Li, F. Meng, and J. Shi, "Learning using privileged information improves neuroimaging-based CAD of Alzheimer's disease: A comparative study," *Med. Biol. Eng. Comput.*, vol. 57, no. 7, pp. 1605-1616, 2019.

[20] Y. Li, L. Cheng, Y. Peng, Z. Wen, and S. Ying, "Manifold alignment and distribution adaptation for unsupervised domain adaptation," in *2019 IEEE Int. Conf. on Multimedia and Expo. (ICME)*, 2019, pp. 688-693.

[21] X. Fei, J. Wang, S. Ying, Z. Hu, and J. Shi, "Projective parameter transfer based sparse multiple empirical kernel learning machine for diagnosis of brain disease," *Neurocomputing*, vol. 413, pp. 271-283, 2020.

[22] X. Fei *et al.*, "Doubly supervised parameter transfer classifier for diagnosis of breast cancer with imbalanced ultrasound imaging modalities," *Pattern Recognit.*, vol. 120, pp.108139, 2021.

[23] N. Sarafianos, M. Vrigkas, and I.A. Kakadiaris, "Adaptive SVM+: learning with privileged information for domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2637-2644.

[24] J. Tang, Y. Tian, X. Liu, D. Li, J. Lv, and G. Kou, "Improved multi-view privileged support vector machine," *Neural Netw.*, vol. 106, pp. 96-109, 2018.

[25] X. Li, B. Du, C. Xu, Y. Zhang, L. Zhang, and D. Tao, "R-SVM+: robust learning with privileged information," in *Proc. Int. Joint Conf. Artificial Intell. (IJCAI)*, 2018, pp. 2411-2417.

[26] H.H. Alahmadi *et al.*, "Classifying cognitive profiles using machine learning with privileged information in mild cognitive impairment," *Front. Comput. Neurosc.*, vol. 10, pp. 1-17, 2016.

[27] X. Zheng, J. Shi, S. Ying, Q. Zhang, and Y. Li, "Improving MRI-based diagnosis of Alzheimer's disease via an ensemble privileged information learning algorithm," in *Int. Symposium Biomed. Imaging (ISBI)*, 2017, pp. 456-459.

[28] V. Cheplygina, M.D. Bruijne, and J.P.W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280-296, 2019.

[29] B. Cheng, D. Zhang, and D. Shen, "Domain transfer learning for MCI conversion prediction," in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI), pp. 82-90, 2012.

[30] C. Wachinger, and M. Reuter, "Domain adaptation for Alzheimer's disease diagnostics," *NeuroImage*, vol. 139, pp. 470-479, 2016.

[31] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 663-670.

[32] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171-184, 2012.

[33] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 1-33, 2014.

[34] Z. Ding, S. Ming, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2014, pp. 1192-1198.

[35] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 850-863, 2016.

[36] S. Wang, L. Zhang, W. Zuo, and B. Zhang, "Class-specific reconstruction transfer learning for visual recognition across domains," *IEEE Trans. Image Process.*, vol. 29, pp. 2424-2438, 2019.

[37] J. Wang *et al.*, "Multi-class ASD classification based on functional connectivity and functional correlation tensor via multi-source domain adaptation and multi-view sparse representation," *IEEE Trans. Med. Imaging*, vol. 39, no. 10, pp. 3137-3147, 2020.

[38] Z. Zhang, Y. Xu, L. Shao, and J. Yang, "Discriminative block-diagonal representation learning for image recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3111-3125, 2018.

[39] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient Schatten p-norm minimization," in *26th AAAI Conf. Artif. Intell.*, 2012, pp. 655-661.

[40] M. Mohammadi, M. Zadeh, and M. Skoglund, "Performance guarantees for Schatten-p quasi-norm minimization in recovery of low-rank matrices," *Signal. Process.*, vol. 114, pp. 225-230, 2015.

[41] P. Indyk, A. Vakilian, and Y. Yuan, "Learning-based low-rank approximations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 7402-7412.

[42] Z. Luo, Y. Zou, J. Hoffman, and F. Li, "Label efficient learning of transferable representations across domains and tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 165-177.

[43] L. Liu, W. Huang, and D. Chen, "Exact minimum rank approximation via Schatten p-norm minimization," *J. Comput. Appl. Math.*, vol. 267, no. 1, pp. 218-227, 2014.

[44] L. Luo, J. Yang, J. Qian, Y. Tai, and G. Lu, "Robust image regression based on the extended matrix variate power exponential distribution of dependent noise," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 9, pp. 2168-2182, 2017.

[45] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding, "Robust matrix completion via joint Schatten$_p$-norm and l$_p$-norm minimization," in *Proc. IEEE Int. Conf. Data. Mining (ICDM)*, 2012, pp. 566-574.

[46] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2004,

pp. 1329-1336.
[47] F. Shang, Y. Liu, and J. Cheng, "Scalable algorithms for tractable Schatten quasi-norm minimization," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 2016-2022.
[48] F. Shang, J. Cheng, Y. Liu, Z. Luo, and Z. Lin, "Bilinear factor matrix norm minimization for robust PCA: Algorithms and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2066-2080, 2018.
[49] J. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956-1982, 2010.
[50] C.R. Jack *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reason. Imaging*, vol. 27, no. 4, pp. 685-691, 2008.
[51] X. Han, J. Wang, W. Zhou, C. Chang, S. Ying, and J. Shi, "Deep doubly supervised transfer network for diagnosis of breast cancer with imbalanced ultrasound imaging modalities," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 141-149, 2020.
[52] Z. Xue, D. Shen, and C. Davatzikos, "Classic: consistent longitudinal alignment and segmentation for serial image computing," *NeuroImage*, vol. 30, no. 2, pp. 388-399, 2006.
[53] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imaging*, vol. 17, no. 1, pp. 87-97, 1998.
[54] Y. Wang et al., "Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates," *PLOS ONE*, vol. 9, no. 1, pp. e77810, 2014.
[55] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imaging*, vol. 20, no. 1, pp. 45-57, 2001.
[56] D. Shen, and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imaging*, vol. 21, no. 11, pp. 1421-1439, 2002.
[57] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, pp. 856-867, 2011.
[58] T. Zhou, M. Liu, K. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE Trans. Med. Imag.* vol. 38, no. 10, pp. 2411-2422, 2019.
[59] X. Chen, H. Zhang, L. Zhang, C. Shen, S.W. Lee, and D. Shen, "Extraction of dynamic functional connectivity from brain grey matter and white matter for MCI classification," *Hum. Brain Mapp.*, vol. 38, no. 10, pp.5019-5034, 2017.
[60] Q. Zhang *et al.*, "Sonoelastomics for breast tumor classification: A radiomics approach with clustering-based feature selection on sonoelastography," *Ultrasound Med. Biol.*, vol. 43, pp. 1058-1069, 2017.
[61] A. Aytar, and A. Zisserman, "Tabula rasa: model transfer for object category detection," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2252-2259.
[62] Y. Yeh, C. Huang, and Y. F. Wang, "Heterogeneous domain adaptation and classification by exploiting the correlation subspace," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2009-2018, 2014.
[63] L. Zhang, J. Fu, S. Wang, D. Zhang, Y. D. Zhao, and C. Chen, "Guide subspace learning for unsupervised domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3374-3388, 2020.
[64] L. Zhang, W. Zuo, and D. Zhang, "LSDT: Latent sparse domain transfer learning for visual adaptation," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp.1177-1191, 2016.
[65] L. Zhang, S. Wang, G.B. Huang, W. Zuo, J. Yang, and D. Zhang, "Manifold criterion guided transfer learning via intermediate domain generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3759-3773, 2019.
[66] M.J. Abdulaal, A.J. Casson, and P. Gaydecki, "Performance of nested vs. Non-nested SVM cross-validation methods in visual BCI: Validation study," in *European Signal Processing Conference (EUSIPCO)*, pp. 1680-1684, 2018.
[67] M. Zhang, Z. Huang, and Y. Zhang, "Restricted-p isometry properties of nonconvex matrix recovery," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4316-4323, 2013.