IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 October 2022, accepted 18 October 2022, date of publication 25 October 2022, date of current version 31 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3216884

RESEARCH ARTICLE

Domain Adaption Based on MSE Criterion and Progressive RKHS Subspace Learning (MSEpRKHS_DA)

YANZHEN QIU[1], SHUYU LIU[2], ZHENGMING MA[1], AND HUI HE[1]

[1]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
[2]Public Experimental Teaching Center, Sun Yat-sen University, Guangzhou 510006, China

Corresponding authors: Shuyu Liu (ljie@mail.sysu.edu.cn) and Zhengming Ma (issmzm@mail.sysu.edu.cn)

This work was supported in part by the Science and Technology Program of Guangzhou, China, under Grant 68000-42050001; and in part by the Natural Science Foundation of China under Grant 61773022.

ABSTRACT Reproducing Kernel Hilbert Space (RKHS) subspace learning is very popular among the domain adaption, which learns a latent RKHS subspace for the source domain and target domain, so that their distribution gap becomes smaller than in the original data space. There is a famous probability theory: two second-order moment random variables are equal if and only if their mean squared error (MSE) is zero. In this paper, firstly, we use second-order moment random variables to model the source domain and target domain. Then, we prove that a second-order moment random variable is still second-order moment after it is transformed into the RKHS subspace. Finally, we propose the MSE criterion to measure the distribution difference between source domain and target domain. To our best knowledge, we are the first to apply the MSE to RKHS subspace learning. And the experiments show the superiority of MSE criterion, which performs better than the common Maximum Mean Difference (MMD) and the Covariance Matrix (CovM) criteria. Furthermore, considering the robustness of the RKHS subspace learning framework to the data dimension, we propose the domain adaption framework of the progressive RKHS subspace learning (pRKHS-DA), which continuously updates the learned RKHS subspace. Each update takes the previous learned subspace as the starting point. The idea of pRKHS-DA is proposed for the first time in this paper. Finally, this paper proposes MSEpRKHS_DA model based on MSE criterion and pRKHS-DA framework. And experiments show that our model achieves higher classification accuracy than some state-of-the-art methods.

INDEX TERMS Domain adaption, MSE criterion, RKHS subspace learning.

I. INTRODUCTION

In the era of information explosion, the traditional statistical machine learning has been difficult to meet the need of the emerging applications, because it has two too big faults to ignore [1]: 1) it needs a large number of labeled samples to train models, which costs a lot of manpower to label samples; 2) it is based on the premise that source domain and target domain obey the same distribution, which varies from the reality. It's more common and practical that the source domain and the target domain have different distributions in real-world applications. For example, in the application of object recognition, the distribution of object images will change due to the changing light and different camera angles. Due to this undeniable fact, the trained-well model on source domain data often cannot achieve the expected results when it is directly applied to the target domain data, which obviously limits the generalization ability and knowledge reuse ability of the trained-well model [2]. In order to improve the performance of model in cross-domain tasks, domain adaption [1], [2] is proposed. Domain adaption aims to reduce the distribution difference between the source domain and the target

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

domain that different from but related with source domain, so that the knowledge obtained in the source domain can be well generalized in the target domain to realize the cross-domain migration [3], [4]. Domain adaption could improve the classification performance of the transferred trained models on target domain data, which avoids the need to train a new model on target domain data. In addition, if target domain has few or no labeled samples and is unable to train models with good performance, we can consider adopting the domain adaption method to pre-train the model in different but related source domain with a large number of labeled data, and then apply the trained model to the target domain with fine tuning. So, domain adaption overcomes the dilemma that the labeled data of the target domain is rare and difficult to obtain in practical applications.

The domain adaption based on Reproducing Kernel Hilbert Space (RKHS) subspace learning is very common [5], [6], [7]. In this method, first, the source domain data and the target domain data are transformed into RKHS; second, a RKHS subspace is learned by using the domain adaption criterion, that is, the criterion to measure the distribution difference; and then the source domain data and the target domain data on RKHS are projected into the learned subspace of RKHS, where their distributions can be same as much as possible. At present, the Maximum Mean Difference criterion (MMD) [8] based on the first-order moment and the Covariance Matrix criterion (CovM) [9] based on the second-order moment are the most common domain adaption criteria used for measuring the distribution gap. Since Gretton [8] in 2006 proposed MMD to measure distribution divergence in RKHS, many papers [12], [14], [15], [16], [5] used MMD criterion to discover a latent RKHS subspace. Besides, the proposed CovM criterion [9], [10] measures the distribution divergence between source domain and target domain. Since the first-order moment and second-order moment cannot represent non-Gaussian distribution totally and the domains usually obey non-Gaussian distributions in real applications, MMD criterion and CovM criterion have poor performance on measuring the non-Gaussian distribution gap. In order to solve the limitation of the MMD criterion and CovM criterion, we propose a new domain adaption criterion based on Mean Squared Error (MSE), which is fit for Gaussian and non-Gaussian distribution. The main aim of RKHS subspace learning is learn a RKHS subspace to decrease the distributions gap between source domain data and target domain data. However, no matter how to learn a subspace, the difference between source domain data and target domain data always exists. Then, in light of the robustness of data dimension in RKHS subspace learning framework, we propose a domain adaption framework based on the progressive RKHS subspace learning (pRKHS-DA). Unlike most existing RKHS subspace learning based domain adaption methods, which only learn the RKHS subspace once, pRKHS-DA framework repeatedly learns and optimizes the RKHS subspace, so that the distribution difference between the source domain data and the target domain data on RKHS subspace is gradually reduced. Our contributions in this paper are listed as follows:

1) We prove the RKHS subspace transformation validity, that is, the second-order moment random variables are still second-order moment after they are transformed into RKHS and then projected into RKHS subspace.
2) In light of MSE theorem and the RKHS subspace transformation validity, we apply the MSE theorem into RKHS subspace learning and then propose an effective domain adaption criterion MSE. The experiments provided show that MSE criterion achieves better results than MMD and CovM criteria.
3) We propose a new domain adaption framework based on the progressive RKHS subspace learning (pRKHS-DA), which gradually reduces the difference between source domain data and target domain data by continuously learning a new subspace.
4) We propose MSEpRKHS_DA model based on MSE criterion and pRKHS-DA framework. And the experiment results show that the our model outperforms some other state-of-the-art methods.

The rest of this paper is organized as follows: In Section II, we briefly review partial works related to domain adaption; In Section III, we introduce some necessary background of second-order moment random variable, RKHS subspace learning and domain adaption based on RKHS subspace learning; In Section IV, we give the proof of the transformation validity of RKHS subspace, propose the MSE criterion, pRKHS-DA framework, and MSEpRKHS-DA model which combines the MSE criterion and pRKHS-DA framework together. In Section V, the experiments show the validity of MSE criterion and MSEpRKHS_DA model respectively; And the conclusion is made in Section VI.

II. RELATED WORKS

The transformation-based domain adaption methods are one of the important types of domain adaption. Generally speaking, the transformation-based aims to learn a representation space by using embedding or transformation, where the distribution of the target domain data could be more similar to that of source domain. In 2006, Gretton [8] put forward the MMD criterion to compare the distributions of samples in RKHS. Since then, the MMD criterion has been widely applied into transformation-based domain adaption methods to calculate the distribution divergence of domains. Pan et al proposed the MMDE [11] approach based on MMD criterion, which learns a low-dimensional space via the optimization of kernel matrix K to reduce the distribution gap of the projected source domain samples and target domain samples. However, MMDE needs the considerably expensive computation overhead to learn the kernel matrix K from data. In 2011, Pan and Yang proposed TCA [12]. TCA firstly maps the source and target domain into the RKHS through the kernel function. Secondly, it minimizes the MMD between the subspace representations of source domain samples and target domain samples to learn some transferable components across domains. Then these components are used to construct the RKHS subspace, where the distributions across domains are close to each other. Finally, the classification models trained on source domain samples in this subspace directly are applied on transformed target domain data. Compared with MMDE, TCA learns a RKHS subspace with low calculation, that is, low-rank matrix W, instead of the K. In addition, in order to further improve the transferring component performance of TCA, Pan proposed a semi-supervised TCA approach (SSTCA) [12]. The SSTCA contains two optimization objectives: one is the MMD minimization as in TCA, the other is the embedding and labels independence maximization and this independence is measured by Hilbert–Schmidt Independence Criterion (HSIC) [13]. Since then, domain adaption based on RKHS subspace learning has received considerable attention from the domain adaption community, which learns a suitable RKHS subspace and then maps the source domain data and the target domain data into this subspace to reduce the difference of the two domains' distributions. The key of the domain adaption based on RKHS subspace learning is to find an appropriate domain adaption criterion to measure the distributions gap between domains, so as to realize the distribution alignment of the source domain and the target domain [5]. Jiang et al proposed IGLDA [14] to uncover a latent RKHS subspace where the distributions of the source domain and target domain could be more similar and local geometries of labeled source domain data could be retained. Specially, IGLDA not only considers the global information of domains by minimizing the MMD to reduce the inter domain distribution difference, but also maintains the local geometry properties of the source domain by maximizing the intraclass distance of the labeled source domain data which improves the dependency between the labels and data, and also makes it easier to separate the samples in the latent subspace. In the same year, Yan et al proposed an unsupervised domain adaption method MIDA [15], which considers minimizing distribution difference between domains via MMD criterion and maximizing the distance between different classes of source domain data. Furthermore, Yan extended the MIDA to semi-supervised version (SMIDA) [15] via feature augmentation strategy to learn a better RKHS subspace. In 2019, Li et al proposed TIT approach [16]. On the one hand, TIT reduces the distribution gap by MMD criterion; on the other hand, it also proposes an effective and fast landmark selection method based on graph to reweigh the samples to enhance the ability of knowledge transfer. What's more, TIT extends RKHS subspace learning into unsupervised heterogeneous domain adaption (HDA) [17] by using multiple transformations to map different domains into a common and latent subspace. The LPJT [18] proposed by Li et al considers the knowledge transfer both feature level (aligning the distributions by MMD criterion) and sample level (preserving the neighborhood relationship of samples by landmark selection) in a unified framework. What's more, as in TIT, LPJT also utilizes the two different transformations for subspace learning, one for each domain, so LPJT can be applied in HDA too. For the SDRKHS-DA method [19], it also considers aligning the distributions by MMD criterion, in the meantime, introduces the dictionary learning into the RKHS subspace learning framework. That is, SDRKHS-DA uses source domain data as dictionary to code the target domain data and keeps the coding as sparse as possible at the same time, which makes the same kind of source domain data and target domain data close to each other in subspace to improve the performance of domain adaption. The CDSPP [20] algorithm extends the locality preserving projection [21] into HDA through the multiple transformation matrices and is able to learn a subspace of better separability. In addition to MMD criterion based on first-order statistics, the domain adaption criterion CovM based on second-order statistics to measure the distribution difference is proposed. For instance, the DACoM model [9] is designed to align the different domains' distributions by minimizing the distance between the covariance matrix of source domain data and that of target domain data. At the same time, the local geometric structure and discriminative information are preserved in DACoM model.

III. PRELIMINARY

In this section, some related background knowledge are introduced. First of all, we give some related contents of the second-order moment random variable. Next, we review the framework of the RKHS subspace learning. What's more, we introduce the concept of domain adaption based on RKHS subspace learning.

A. SECOND-ORDER MOMENT RANDOM VARIABLE

Given a random variable X which obeys the distribution $p\left(x\right)$, it becomes a second-order moment random variable if the condition $E\left[|X|^2\right] = \int_{\Omega} x^2 p\left(x\right) \mathrm{d}x < +\infty$ is satisfied. From the view of physics, a second-order moment random variable is the limited-energy random signal. Since all signals have limited energy in the real life, we can use second-order moment random variables to model them. So, the source domain data and target domain data in original data space can be treated as the samplings from two second-order moment random variables with different distributions respectively.

Assuming that a set H_1 contains all second-order moment variables:

$$H_1 = \left\{X \left| E\left[|X|^2\right] < +\infty\right.\right\} \tag{1}$$

According to [22], H_1 is a L^2 space that belongs to Hilbert space and its inner product is defined as

$$(X, Y)_{H_1} = E\left[XY^*\right],$$

where $\forall X, Y \in H_1$, the star denotes the complex conjugate, and the inner product specified by round brackets on L^2 space.

In light of the positive definiteness of inner product defined in Hilbert space L^2, any two elements from L^2 are equal if the norm of their difference is zero, which can be formulated as follows [23]:

$$\begin{aligned} X_1 = X_2 \Leftrightarrow \|X_1 - X_2\|_{H_1}^2 &= (X_1 - X_2, X_1 - X_2)_{H_1} \\ &= E\left[\|X_1 - X_2\|^2\right] = 0, \end{aligned} \tag{2}$$

where X_1 and X_2 are second-order moment random variables from the H_1 space. So, we can see that the necessary and sufficient condition for two second-order moment random variables to be equal is that the MSE between them is zero, which is famous in probability theory.

B. THE FRAMEWORK OF RKHS SUBSPACE LEARNING

The space H_2 includes all square integrable functions and its mathematical expression is given by

$$H_2 : \left\{ f \middle| f : \Omega \to \mathbb{R}, \int_\Omega |f(x)|^2 \mathrm{d}x < +\infty \right\}.$$

H_2 is a Hilbert space and the inner product of H_2 space is defined as

$$\langle f, g \rangle_{H_2} = \int_\Omega f(x)\, g^*(x) \mathrm{d}x,$$

where the star denotes the complex conjugate.

In particular, if there is a binary function $k(x', x) : \Omega \times \Omega \to \mathbb{R}$ that satisfies [24]:

1) For $\forall x \in \Omega$, $k(\cdot, x) \in H_2$;
2) For $\forall f \in H_2$ and $\forall x \in \Omega$,

$$f(x) = \langle f, k(\cdot, x) \rangle_{H_2}.$$

Then, we can call $k(x', x)$ as reproducing kernel.

And we can define a mapping function $\varphi(x) : \Omega \to \mathbb{R}$. For $\forall x \in \Omega$, we have $\varphi(x) = k(\cdot, x) \in H_2$. According to the property of the reproducing kernel, we have

$$\left\langle \varphi(x), \varphi(x') \right\rangle_{H_2} = k(x', x).$$

Given a set of samples $X = \{x_1, \ldots, x_N\} \subseteq \Omega$ and the reproducing kernel k, we can transform X into the Hilbert space H_2 to get $\varphi(X) = \{\varphi(x_1), \ldots, \varphi(x_N)\} \subseteq H_2$. And the new orthogonal basis ϑ_i of RKHS subspace H_s can be constructed through linear combination of these non-orthogonal feature vectors:

$$\vartheta_i = \sum_{j=1}^{N} w_{ji} \varphi(x_j), \quad i = 1, \ldots, d, \tag{3}$$

where d is the dimension of RKHS subspace.

Eq.(3) can be cast into the matrix form as follows

$$\Theta = \Phi W, \tag{4}$$

with

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{Nd} \end{bmatrix}, \quad \Theta = \begin{bmatrix} \vartheta_1 & \ldots & \vartheta_d \end{bmatrix},$$
$$\Phi = \begin{bmatrix} \varphi(x_1) & \ldots & \varphi(x_N) \end{bmatrix}.$$

The orthogonality of the new basis Θ satisfies the following condition:

$$\begin{bmatrix} \langle \vartheta_1, \theta_1 \rangle_{H_2} & \cdots & \langle \vartheta_1, \vartheta_d \rangle_{H_2} \\ \vdots & \ddots & \vdots \\ \langle \vartheta_d, \vartheta_1 \rangle_{H_2} & \cdots & \langle \vartheta_d, \vartheta_d \rangle_{H_2} \end{bmatrix} = \Theta^T \Theta = W^T K W = I_d. \tag{5}$$

where K is the kernel matrix given by

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix} \tag{6}$$

According to the subspace projection theorem in Hilbert space [24], the corresponding coordinates y_i of $\varphi(x_i)$ in the RKHS subspace basis H_s with Θ is given by

$$y_i = \begin{bmatrix} \langle \varphi(x_i), \vartheta_1 \rangle_{H_2} \\ \vdots \\ \langle \varphi(x_i), \vartheta_d \rangle_{H_2} \end{bmatrix} = W^T K_{iCol} \in \mathbb{R}^d. \tag{7}$$

where d is dimension of the RKHS subspace H_s, and K_{iCol} represents the ith column vector of the kernel matrix K.

C. THE DOMAIN ADAPTION BASED ON RKHS SUBSPACE LEARNING

In domain adaption, the labeled $X_s = \left\{x_1^s, \cdots, x_{n_s}^s\right\}$ and the unlabeled $X_t = \left\{x_1^t, \cdots, x_{n_t}^t\right\}$ are from source domain and target domain respectively, which obey different distributions.

Domain adaption based on RKHS subspace learning tries to find a latent RKHS subspace to minimize their distribution difference. First, the kernel transformation $\varphi(x) = k(\cdot, x)$ maps the data samples $X = X_s \cup X_t$ into the RKHS space H_2 to get $\varphi(X)$. Then, $\varphi(X)$ are projected into the RKHS subspace H_s. Specifically, the source and target domain samples on H_s are denoted as Y_s and Y_t respectively, where

$$\begin{aligned} Y_s &= \left[y_1^s, \ldots, y_{n_s}^s\right] = W^T K_s \in \mathbb{R}^{d \times n_s}, \\ Y_t &= \left[y_1^t, \ldots, y_{n_t}^t\right] = W^T K_t \in \mathbb{R}^{d \times n_t}, \quad K = [K_s, K_t] \end{aligned}$$

and

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix}, \quad N = n_s + n_t \tag{8}$$

Generally, the MMD criterion measures the distribution gap between the source domain data Y_s and the target domain data Y_t written as [12]

$$MMD(Y_s, Y_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} y_i^s - \frac{1}{n_t} \sum_{j=1}^{n_t} y_j^t \right\|_{H_2}^2,$$

where $\|\cdot\|_{H_2}$ is the RKHS norm.

In addition, the formulas of CovM criterion for the Y_s and Y_t is [9]

$$CovM\left(Y_s, Y_t\right) = \left\| \sum\nolimits_s - \sum\nolimits_t \right\|_F^2,$$

where $\|\cdot\|_F$ is Frobenius norm, $\sum_s$ and $\sum_t$ represent the covariance matrices of Y_s and Y_t.

IV. OUR WORKS

In this section, we give the proof of the validity of RKHS subspace transformation for second-order moment variables. Then, based on this proof, we propose an effective MSE criterion to match distributions. In addition, we propose a novel domain adaption model named MSEpRKHS-DA shown in Fig.1: First, the framework of domain adaption based on the progressive RKHS subspace learning (pRKHS-DA) is introduced. So far as we know, we are the first to propose the idea of pRKHS-DA framework. Then, we combine the MSE criterion with pRKHS-DA framework to put forward the MSEpRKHS-DA model.

A. RKHS SUBSPACE TRANSFORMATION VALIDITY

Supposing that X is a random variable, its samplings $\{x_1, \cdots, x_N\}$ come from the original data space Ω, and the new representation of X in H_2 subspace is a random vector Y according to section III-B:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_d \end{bmatrix} = \begin{bmatrix} \langle \varphi(X), \vartheta_1 \rangle_{H_2} \\ \vdots \\ \langle \varphi(X), \vartheta_d \rangle_{H_2} \end{bmatrix}.$$

Now, we prove that if X is a second-order moment random variable, Y is a second-order moment random vector by proving that each component Y_i of Y is second-order moment.

Theorem 1: If X is a second-order moment variable, then $\varphi(X)$ is the second-order moment random process, that is, $\forall x \in X, \varphi(X)(x) = k(x, X)$ is a second-order moment variable.

The proof of Y_i is second-order moment is showed in Appendix A.

B. MSE DOMAIN ADAPTION CRITERION

Given source domain data $X_s = \left\{x_1^s, \cdots, x_{n_s}^s\right\} \sim p(x)$ and target domain data $X_t = \left\{x_1^t, \cdots, x_{n_t}^t\right\} \sim q(x)$ where $p(x) \neq q(x)$. In line with Section III-A and Section IV-A, we can get the corresponding second-order moment Y_s and Y_t respectively. According to Eq.(2), we put forward the MSE criterion to measure the distribution discrepancy between Y_s and Y_t, which could find a shared latent RKHS subspace where the distributions of source and target domain can align better. And the MSE criterion for Y_s and Y_t is:

$$MSE\left(Y_s, Y_t\right) = E\left[\left\|Y_s - Y_t\right\|^2\right]. \tag{9}$$

Due to $Y_s = \left\{y_1^s, \cdots, y_{n_s}^s\right\}$ and $Y_t = \left\{y_1^t, \cdots, y_{n_t}^t\right\}$. So, the Eq. (9) can be rewritten as:

$$\begin{aligned} MSE\left(Y_s, Y_t\right) &= \frac{1}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \left\| y_i^s - y_j^t \right\|^2 \\ &= tr\left(W^T \Psi W\right) \end{aligned} \tag{10}$$

For the specific derivation of the Eq.(10) is shown in Appendix B.

C. pRKHS-DA FRAMEWORK

Given the source domain X_s and target domain X_t, the goal of domain adaption based on RKHS subspace learning is to learn a latent RKHS subspace W, where the distribution discrepancy between the corresponding transformed Y_s and Y_t can be smaller than in original data space, that is:

$$\left|prob\left(Y_s\right) - prob\left(Y_t\right)\right| \leq \left|prob\left(X_s\right) - prob\left(X_t\right)\right|,$$

where the $prob(\cdot)$ represents the probability distribution.

Obviously, the closer the distribution of Y_t to that of Y_s, the higher the classification accuracy can be achieved by directly transferring the classifier trained on labeled Y_s to Y_t. In this paper, we propose the framework of domain adaption based on progressive RKHS subspace learning (pRKHS-DA), which reduces $\left|prob\left(Y_s\right) - prob\left(Y_t\right)\right|$ gradually by treating the learned subspace as a new data space for the next subspace learning and repeating this process. Specifically, pRKHS-DA framework has the following characteristics:

1) Consider the original data space Ω_0, the i*th* data space Ω_i, and RKHS subspace $W^{(i)}$ learned for the i time. First, we transform the source domain data X_s and target domain data X_t in Ω_0 into the learned subspace $W^{(1)}$, and we get $Y_s^{(1)}$ and $Y_t^{(1)}$ respectively. Second, we regard the $W^{(1)}$ as the new data space, and $Y_s^{(1)}$ and $Y_t^{(1)}$ as the new source domain data $X_s^{(2)}$ and target domain data $X_t^{(2)}$. Mathematically, $\Omega_1 = W^{(1)}$, $X_s^{(2)} = Y_s^{(1)}$, and $X_t^{(2)} = Y_t^{(1)}$. Then, repeat this subspace learning many times. And in the repeated process, the distributions of the transformed source domain and target domain will be closer. In addition, the process of pRKHS-DA framework can be expressed as:

$$\begin{aligned} \Omega_0 \overset{W^{(1)}}{\rightarrow} \Omega_1 = W^{(1)} \rightarrow \cdots \rightarrow \Omega_k = W^{(k)} \overset{W^{(k+1)}}{\rightarrow} \Omega_{k+1} \\ = W^{(k+1)} \end{aligned}$$

At present, this paper cannot theoretically prove the convergence of the pRKHS-DA framework. However, the experiments provided later in this paper show that pRKHS-DA framework can effectively improve the classification accuracy of target domain data.

2) pRKHS-DA framework benefits from the framework of RKHS subspace learning. In the framework of RKHS subspace learning, the dimension of the mapped data on subspace can be set artificially. Therefore, as long as the fixed subspace dimension remains unchanged,

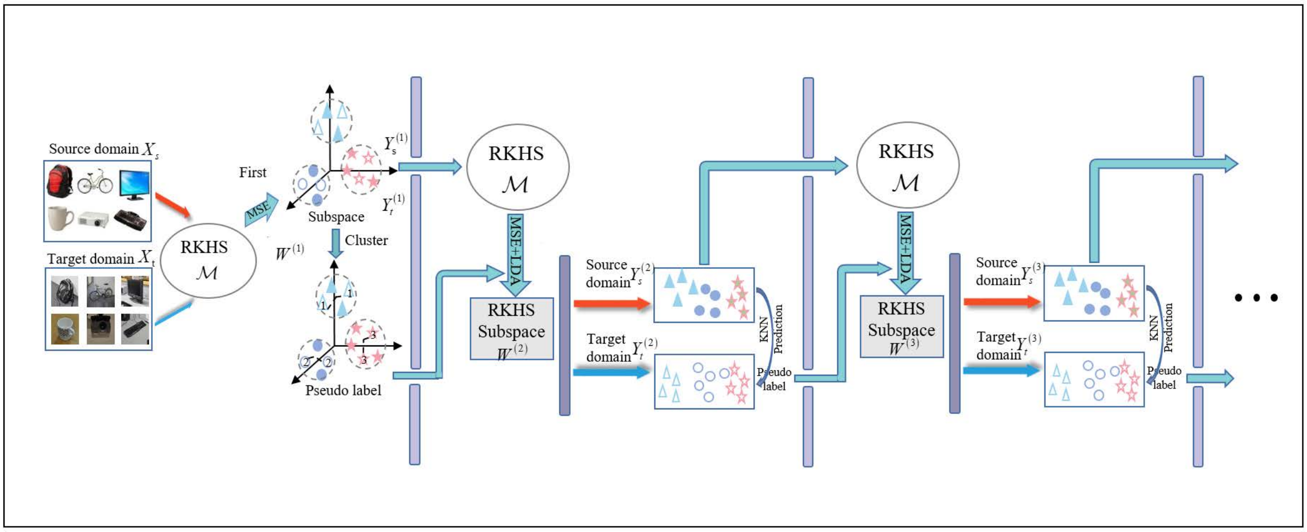


FIGURE 1. The process of MSEpRKHS-DA model which reduces the distribution gap between the projected source domain and target domain gradually. Firstly, we project all samples in the original data space into the learned subspace $W^{(1)}$ by minimizing the MSE criterion; Secondly, we learn a new subspace $W^{(2)}$ based on the previous subspace learning$W^{(1)}$; Then, we constantly learn a better subspace $W^{(k+1)}$ based on the $W^{(k)}$.

the problem of smaller and smaller data dimension will not occur in pRKHS-DA framework.

3) In fact, pRKHS-DA is a framework of domain adaption based on RKHS subspace learning, and the rules used for each subspace learning is optional.

D. MSEpRKHS-DA MODEL

Based on all the above works, we propose a MSEpRKHS-DA model, which combines MSE criterion and pRKHS-DA framework.

Given the RKHS (H_2, k) generated from the original data space Ω_0, where k represents the reproducing kernel, the labeled-well source domain data $X_s = \left\{x_1^s, \cdots, x_{n_s}^s\right\} \subseteq \Omega_0$, and unlabeled target domain data $X_t = \left\{x_1^t, \cdots, x_{n_t}^t\right\} \subseteq \Omega_0$. Now, we elaborate the procedure of the MSEpRKHS-DA model in detail.

1) We first construct the subspace $W^{(1)}$ of H_2 via MSE criterion, and the optimization problem is formulated as follows:

$$\underset{W^{(1)}}{argmin} \quad \mu_1 MSE\left(Y_s^{(1)}, Y_t^{(1)}\right) + \mu_2 tr\left({W^{(1)}}^T W^{(1)}\right)$$
$$\text{s.t. } {W^{(1)}}^T K^{(1)} W^{(1)} = I_d \tag{11}$$

where

$$K^{(1)} = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix}, \; N = n_s + n_t$$

According to section III-B, we get the $Y_s^{(1)}$ and $Y_t^{(1)}$ on the optimized subspace $W^{(1)}$. Specifically, the first term $MSE\left(Y_s^{(1)}, Y_t^{(1)}\right)$ is the MSE criterion between the $Y_s^{(1)}$ and $Y_t^{(1)}$, the second term $tr\left({W^{(1)}}^T W^{(1)}\right)$ is used to limit the complexity of $W^{(1)}$.

2) Let $\Omega_1 = W^{(1)}$, $X_s^{(2)} = Y_s^{(1)}$, and $X_t^{(2)} = Y_t^{(1)}$. Then we use the clustering method ADPC-KNN [30] to label the $Y_t^{(1)}$ data, so we get the labeled data $Y^{(1)} = Y_s^{(1)} \cup Y_t^{(1)}$. In order to improve the performance of the domain adaption classification, we make further efforts to learn a new subspace $W^{(2)}$ based on $W^{(1)}$. We not only consider reducing the distribution difference via MSE criterion, but also consider maximizing $LDA\left(Y^{(2)}\right)$ [31], so that the projected samples with same labels are as close as possible and the projected samples with different labels are as far away as possible. And the optimization problem can be written as:

$$\underset{W^{(2)}}{argmin} \quad \mu_1 MSE\left(Y_s^{(2)}, Y_t^{(2)}\right) + \mu_2 tr\left({W^{(2)}}^T W^{(2)}\right)$$
$$-\lambda LDA\left(Y_s^{(2)} \cup Y_t^{(2)}\right)$$
$$\text{s.t. } {W^{(2)}}^T K^{(2)} W^{(2)} = I_d \tag{12}$$

3) Based on $W^{(2)}$, we repeat the above subspace learning Eq. (12) that gradually reduces the distribution gap, and then we get the ideal subspace $W^{(i)}$ and the corresponding $Y_s^{(i)}$ and $Y_t^{(i)}$.

4) Train the classifier on the projected samples $Y_s^{(i)}$, then use the trained classifier to label the projected samples $Y_t^{(i)}$.

V. EXPERIMENT

In this section, we have three types of experiments in total: first, we compare our MSE criterion with the common MMD and CovM criteria on four dataset; second, we compare our MSEpRKHS-DA model with seven domain adaption

FIGURE 2. The headphone samples selected from the four domains in Office-Caltech10 dataset.

methods on other four dataset; finally, we conduct the ablation studies of kernel function and RBF kernel size for the MSE criterion.

A. THE COMPARISON WITH MMD AND CovM CRITERION

We evaluate the proposed MSE criterion on four popular dataset: Office-Caltech10 dataset, handwritten digits dataset, text dataset, and VLSIC dataset. And we compare our criterion with the popular criteria: MMD criterion and CovM criterion. The four dataset used all downloaded from this website address.[1] The related instructions of this experiment are as follows:

1) We use the RBF kernel [25], $k(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|}{2\delta^2}}$ and $\delta = 10$, to map samples into the RKHS.
2) The dimension of the RKHS subspace H_s is set to $d = 30$ for the handwritten digits dataset and $d = 100$ for the other three dataset.
3) k-Nearest Neighbor method (knn) is used as the classifier. The principle of knn classifier is that the label of x is determined by the labels of the nearest k samples, that is, the label with the most occurrences in the k labels is the prediction label of x. And the experiments are carried out on $k = 1, 3, 5, 7$.

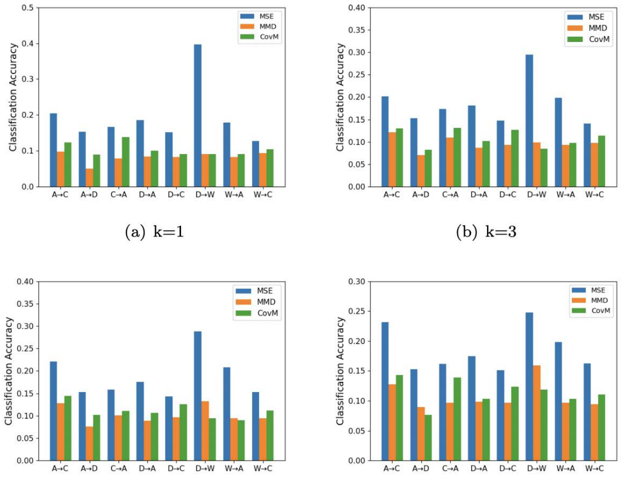
1) OFFICE-Caltech10 DATASET

Office-Caltech10 dataset [29] consists of four domains: Amazon (A), DSLR (D), Webcam (W), and Caltech (C). Each domain contains 10 same classes, such as backpack, monitor, headphone and so on. Examples of headphones from A, D, W, and C domains are shown in Fig. 2.

According to IGLDA [14], the Speed Up Robust Features (SURF) [26] of this dataset are first extracted; then the SURF of each domain are normalized. For each domain adaption classification task, we randomly selected two domains as the source domain and target domain respectively. In total, we carried out eight tasks: A→C, A→D, C→A, D→A, D→C, D→W, W→A, W→C. For instance, A→C means that Amazon domain is the source domain and Caltech domain is the target domain. And the results shown in Table 1 and Fig.3 indicate that MSE criterion outperforms MMD and CovM criteria on Office-Caltech10 dataset.

TABLE 1. The knn (k = 1,3,5,7) classification accuracy comparison of MSE, MMD, CovM respectively on Office-Caltech10 dataset.

$k = 1$	MSE	MMD	CovM	$k = 3$	MSE	MMD	CovM
A → C	**0.2048**	0.0971	0.1238	A → C	**0.2012**	0.1211	0.1300
A → D	**0.1529**	0.0510	0.0892	A → D	**0.1529**	0.0701	0.0828
C → A	**0.1670**	0.0793	0.1378	C → A	**0.1733**	0.1096	0.1315
D → A	**0.1858**	0.0835	0.1002	D → A	**0.1806**	0.0866	0.1023
D → C	**0.1523**	0.0825	0.0908	D → C	**0.1478**	0.0935	0.1264
D → W	**0.3966**	0.0915	0.0915	D → W	**0.2949**	0.0983	0.0847
W → A	**0.1795**	0.0825	0.0908	W → A	**0.1983**	0.0929	0.0971
W → C	**0.1273**	0.0935	0.1051	W → C	**0.1407**	0.0971	0.1140
$k = 5$	MSE	MMD	CovM	$k = 7$	MSE	MMD	CovM
A → C	**0.2208**	0.1282	0.1443	A → C	**0.2315**	0.1273	0.1434
A → D	**0.1529**	0.0764	0.1019	A → D	**0.1529**	0.0892	0.0764
C → A	**0.1587**	0.1013	0.1106	C → A	**0.1618**	0.0971	0.1388
D → A	**0.1754**	0.0887	0.1065	D → A	**0.1743**	0.0981	0.1033
D → C	**0.1434**	0.0962	0.1256	D → C	**0.1514**	0.0971	0.1238
D → W	**0.2881**	0.1322	0.0949	D → W	**0.2475**	0.1593	0.1186
W → A	**0.2077**	0.0939	0.0905	W → A	**0.1983**	0.0971	0.1033
W → C	**0.1532**	0.0944	0.1113	W → C	**0.1621**	0.0944	0.1104



(a) k=1

(b) k=3

(c) k=5

(d) k=7

FIGURE 3. The knn (k = 1,3,5,7) classification performance of MSE, MMD, CovM respectively on Office-Caltech10 dataset.

2) TEXT DATASET

In fact, the used text dataset is a pre-processed subset of Reuters-21578 dataset,[2] which is divided into three domains: orgs, places, and people, and each domain contains two classes [14]. For this dataset, one classification task are set, that is, orgs domain is the source domain and places domain is the target domain. And this task is denoted by orgs → places.

As shown in the Table 2 and Fig.4, the MSE criterion has better classification performance than the other two criteria on $k = 1, 3, 5, 7$. Our criterion achieves the most significant improvement over MMD criterion on $k = 7$ and CovM criterion on $k = 3$, which is 5.66% higher than MMD criterion and 4.62% higher than CovM criterion respectively.

[1] https://github.com/jindongwang/transferlearning/tree/mas

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

TABLE 2. The orgs → places domain adaption classification results of MSE, MMD, CovM criterion respectively on the text dataset, where orgs domain is the source and places domain is the target.

orgs → places	MSE	MMD	CovM
k=1	**0.5465**	0.4372	0.5177
k=3	**0.5523**	0.5062	0.5091
k=5	**0.5446**	0.4976	0.5110
k=7	**0.5638**	0.5072	0.5283

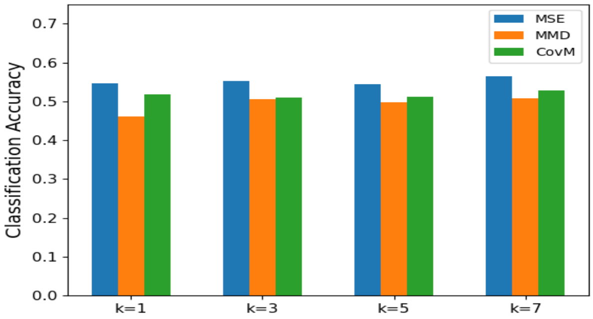
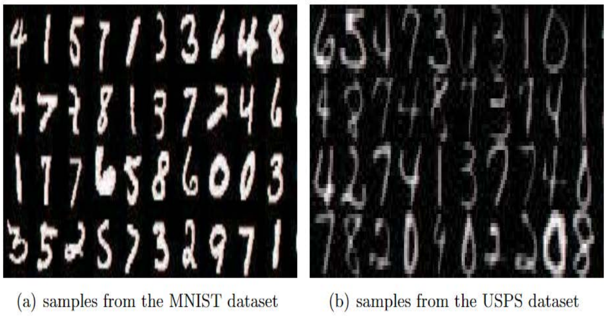


FIGURE 4. The classification accuracy of MSE, MMD, CovM criterion in different *k* on text dataset.

3) HANDWRITTEN DIGITS DATASET

The handwritten digits dataset consists of MNIST [27] and USPS [28] dataset with different distributions, which include handwritten 10 digits from 0 to 9. MNIST dataset contains 70000 sheets of 28 × 28 gray images, and the USPS dataset contains 11000 sheets of 16 × 16 gray images. Since the large amount of samples in this dataset and the limited processing power of our device, the subset of handwritten digits dataset is used in following experiments, which consists of 2000 images from MNIST and 1800 images from USPS that are all randomly selected. Then, some data preparation are done for this subset, which contain the uniformly scaling these gray images to 16 × 16 images, and then flattening each image into 256 dimensional vector. Some examples of the handwritten digits dataset are shown in Fig. 5. And MNIST and USPS dataset are taken as source and target domain by turns. Compared with MMD and CovM criteria, MSE criterion on MNIST → USPS task achieves 47.14% and 24.19% improvement in average classification accuracy. And for USPS → MNIST task, the average accuracy of MSE criterion is 13.64% higher than MMD and 24.10% higher than CovM criterion.



(a) samples from the MNIST dataset (b) samples from the USPS dataset

FIGURE 5. The samples from the handwritten digits dataset.

TABLE 3. The classification accuracy of MSE, MMD, CovM in handwritten digits dataset.

MNIST → USPS	MSE	MMD	CovM
k=1	**0.6489**	0.1739	0.4267
k=3	**0.6517**	0.1844	0.4150
k=5	**0.6683**	0.1878	0.4156
k=7	**0.6728**	0.2100	0.4167
USPS → MNIST	MSE	MMD	CovM
k=1	**0.3775**	0.2120	0.1000
k=3	**0.3865**	0.2450	0.1435
k=5	**0.3795**	0.2555	0.1410
k=7	**0.3675**	0.2530	0.1625

4) VLSIC DATASET

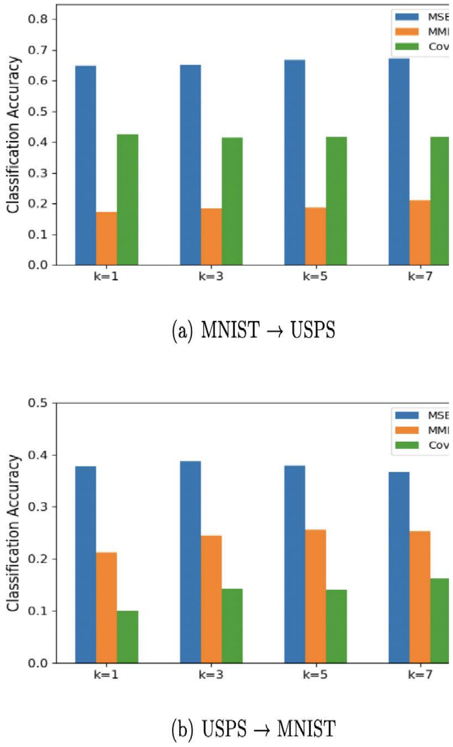
The VLSIC dataset consists of 5 domains from different distributions: VOC2007(V), LabelMe(L), SUN09(S), ImageNet(I), and Caltech101(C). Since the original data have very high dimension, we firstly applied principal component analysis (PCA) to reduce the dimension of original data from 4096 into 300. And we only selected the 5 classes shared by the five domains to conduct the experiments. We have set up six domain adaption tasks totally: C→L, C→S, C→V, I→C, I→V, V→L, and the results of these tasks are showed in Table 4 and Fig.7.

TABLE 4. The accuracy comparison of MSE, MMD, CovM on the VLSIC dataset.

$k = 1$	MSE	MMD	CovM	$k = 3$	MSE	MMD	CovM
C → L	**0.4620**	0.2771	0.2364	C → L	**0.4635**	0.2677	0.2003
C → S	**0.3827**	0.2367	0.1999	C → S	**0.3851**	0.1987	0.1496
C → V	**0.4437**	0.2707	0.2145	C → V	**0.4437**	0.2556	0.1807
I → C	**0.3781**	0.1929	0.2007	I → C	**0.3816**	0.1731	0.1816
I → V	**0.3353**	0.1842	0.1928	I → V	**0.3326**	0.1505	0.1431
V → L	**0.3823**	0.3008	0.3200	V → L	**0.3923**	0.3313	0.3343
$k = 5$	MSE	MMD	CovM	$k = 7$	MSE	MMD	CovM
C → L	**0.4646**	0.2944	0.2101	C → L	**0.4654**	0.3309	0.2161
C → S	**0.3851**	0.2188	0.1755	C → S	**0.3851**	0.2282	0.1755
C → V	**0.4437**	0.2823	0.1899	C → V	**0.4437**	0.3089	0.1931
I → C	**0.2678**	0.1767	0.1908	I → C	**0.2707**	0.1710	0.1830
I → V	**0.1525**	0.1517	0.1466	I → V	**0.1540**	0.1327	0.1437
V → L	**0.4040**	0.3566	0.3611	V → L	**0.3938**	0.3859	**0.3938**

B. COMPARE MSEpRKHS-DA MODEL WITH SOME STATE-OF-THE-ART METHODS

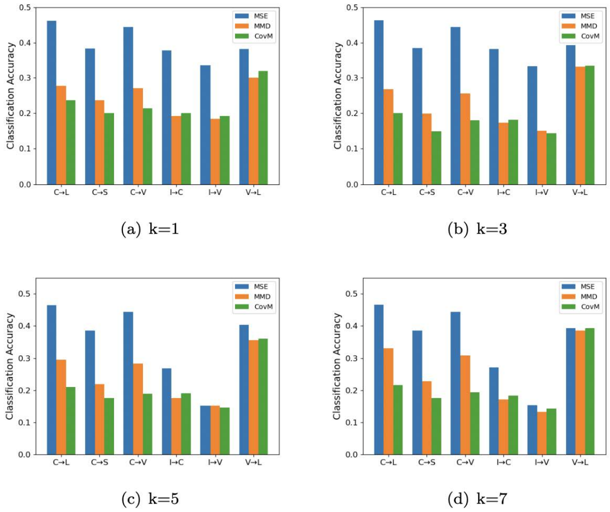
Here, we compare our MSEpRKHS-DA model with seven methods on four real dataset to evaluate its performance. And these methods are TCA [12], SSTCA [12], IGLDA [14], TIT [16], CDSPP [20], SDRKHS-DA [19], and LPJT [18].



(a) MNIST → USPS

(b) USPS → MNIST

FIGURE 6. The classification accuracy of MSE, MMD, CovM on handwritten digits dataset; (a) MNIST dataset is the source domain, and USPS dataset is the target domain; (b) the domain setting is just the opposite of (a).



(a) k=1

(b) k=3

(c) k=5

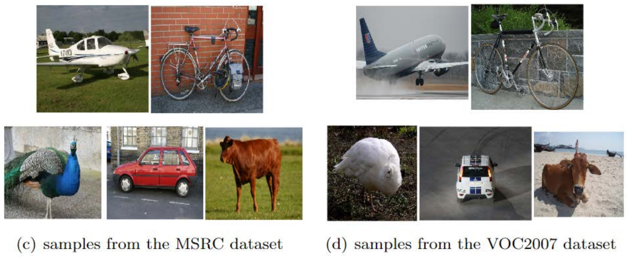
(d) k=7

FIGURE 7. The performance of MSE, MMD, CovM on VLSIC dataset.

And the knn (k = 1) classifier is used. In addition, the used reproducing kernel is Laplacian kernel $k(x, y) = exp\left(\frac{\|x-y\|}{\sigma}\right)$. For convenience, we bold the highest classification accuracy in each task.

1) MSRC-VOC2007 DATASET

MSRC-VOC2007 dataset [32] is a standard dataset for object classification, including MSRC dataset and VOC2007 dataset. The MSRC dataset has 4223 RGB images of 18 objects, while the VOC2007 dataset has 5011 RGB images of 20 objects. For this domain adaption classification experiment, we select five categories common to MSRC dataset and VOC2007 dataset: aircraft, bicycle, bird, car, cattle, sheep, and select 50 images from each category for experiment. Fig.8 shows the examples of the common categories of MSRC dataset and VOC2007 dataset.



(c) samples from the MSRC dataset

(d) samples from the VOC2007 dataset

FIGURE 8. The samples of the five common categories in MSRC dataset and VOC2007 dataset.

In the experiment, we record MSRC and VOC2007 as M and V respectively, and these two dataset are used as source domain and target domain in turn. Therefore, we totally have two tasks: M→V and V→M. And the parameter settings of MSEpRKHS-DA model are: $\mu_1 = 10.0$, $\mu_2 = 1.0$, $\lambda = 0.4$ The dimension of subspace is 40.

TABLE 5. The performance of MSEpRKHS-DA model and the state-of-the-art methods in MSRC-VOC2007 dataset.

Method	M→V	V→M	Average
TCA[12]	0.3720	0.4280	0.4000
SSTCA[12]	0.3240	0.3840	0.3540
IGLDA[14]	0.3760	0.4480	0.4120
TIT[16]	0.3520	0.3640	0.3580
SDRKHS-DA[19]	0.3640	0.4280	0.3960
CDSPP[20]	0.2490	0.3683	0.3086
LPJT[18]	0.3600	0.3800	0.3700
Ours	**0.4200**	**0.4880**	**0.4540**

According to the classification accuracy of Table 5, our model achieves higher classification accuracy than the other seven methods in M→V and V→M, and the accuracies of the two tasks are improved by $4.80\% - 17.10\%$ and $4.00\% - 12.40\%$ respectively compared with the other methods.

2) PIE DATASET

There are 40000 32 × 32 gray images in the PIE dataset.[3] These images record 68 volunteers' different postures and expressions under changing light conditions. In this experiment, we use five subsets of this dataset: PIE05, PIE07, PIE09, PIE27 and PIE29, where PIE05, PIE07, PIE09, PIE27

[3] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

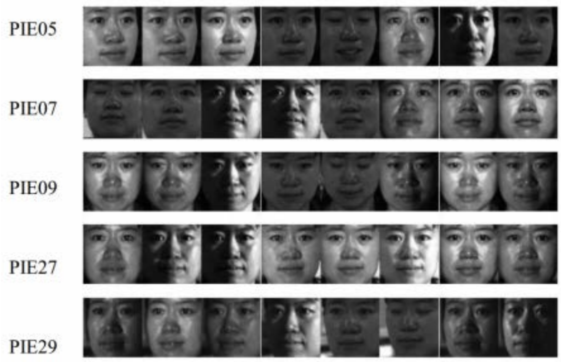


FIGURE 9. The samples of the used five subsets of PIE dataset.

and PIE29 contain images of the face is towards the left, the sky, the earth, the front, and the right respectively. And each subset is treated as a domain. The samples of these five subsets used in this experiment are shown in Fig.9.

TABLE 6. The classification accuracy of our model and the baseline methods on 20 domain adaption tasks of PIE dataset.

Task	Method							
	TCA[12]	SSTCA[12]	TIT[16]	IGLDA[14]	SDRKHS-DA[19]	CDSPP[20]	LPJT[18]	Ours
PIE05→PIE07	0.1357	0.1492	0.1848	0.1418	0.1351	0.1175	**0.2500**	0.2400
PIE05→PIE09	0.1373	0.1403	0.1716	0.1409	0.1360	0.1141	0.1900	**0.2365**
PIE05→PIE27	0.3253	0.3277	0.3214	0.3355	0.3265	0.2739	0.3900	**0.4551**
PIE05→PIE29	0.1054	0.1066	0.1501	0.1029	0.1054	0.1006	**0.2000**	0.1844
PIE07→PIE05	0.1501	0.1495	0.1441	0.1630	0.1495	**0.4412**	0.2300	0.2809
PIE07→PIE09	0.5282	0.5074	0.4547	0.5147	0.5276	0.1876	0.3700	**0.5729**
PIE07→PIE27	0.1901	0.2034	0.1727	0.1968	0.1901	**0.4319**	0.3300	0.3527
PIE07→PIE29	0.2898	0.2813	0.2935	0.2868	0.2898	0.1759	0.2400	**0.3456**
PIE09→PIE05	0.1441	0.1456	0.1303	0.1546	0.1435	**0.4301**	0.2300	0.2578
PIE09→PIE07	0.5089	0.5058	0.5052	0.5071	0.5077	0.1886	0.3400	**0.5482**
PIE09→PIE27	0.1947	0.2070	0.1835	0.2082	0.1944	**0.4304**	0.3700	0.3689
PIE09→PIE29	0.2911	0.2678	0.3186	0.2813	0.2898	0.1756	0.2500	**0.3719**
PIE27→PIE05	0.3340	0.3448	0.3259	0.3523	0.3334	0.2919	0.4000	**0.4688**
PIE27→PIE07	0.1971	0.2388	0.2468	0.2063	0.1971	0.1250	**0.4200**	0.4002
PIE27→PIE09	0.2071	0.2286	0.2555	0.2120	0.2071	0.1189	0.4100	**0.4216**
PIE27→PIE29	0.1636	0.1691	0.2353	0.1679	0.1636	0.1014	0.1700	**0.2702**
PIE29→PIE05	0.1113	0.1008	0.1074	0.1227	0.1104	**0.4380**	0.1900	0.1966
PIE29→PIE07	0.2879	0.2646	0.2971	0.2855	0.2861	0.1886	0.2100	**0.3198**
PIE29→PIE09	0.2953	0.2776	0.2586	0.2819	0.2941	0.1850	0.2300	**0.3431**
PIE29→PIE27	0.1514	0.1511	0.1397	0.1655	0.1517	**0.4337**	0.2100	0.2761
Average	0.2374	0.2383	0.2448	0.2414	0.2369	0.2475	0.2815	**0.3456**

In this experiment, we set up a total of 20 classification tasks, and the five subsets are as source domain and target domain in turn. Each task can be expressed as: source domain → target domain. And the parameters of MSEpRKHS-DA model are set as: $\mu_1 = 1.0, \mu_2 = 1.0, \lambda = 4.0$. The dimension of subspace is 150.

From Table 6, we can see that among the 20 classification tasks in the PIE dataset, the proposed model has achieved the best accuracies in the 75% tasks. Moreover, the average classification accuracy of MSEpRKHS-DA is 10.82%, 10.72%, 10.07%, 10.42%, 10.86%, 9.81%, and 6.41% higher than that of TCA, SSTCA, TIT, IGLDA, SDRKHS-DA, CDSPP, and LPJT respectively.

3) ORL DATASET

ORL dataset[4] is a face dataset, which contains 400 gray images with 92 ×112 size. In particular, this dataset is composed of 400 images taken by 40 volunteers at different times and each person took 10 pictures with different expressions.

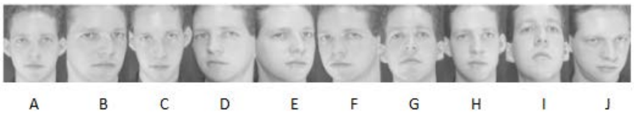


FIGURE 10. The 10 different expressions of a volunteer in ORL dataset.

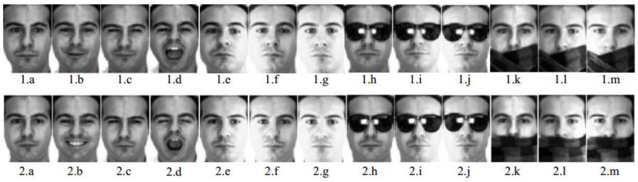


FIGURE 11. The 26 images of one volunteer in AR dataset. And 1.a-1.m represent the images taken in the first day, 2.a-2.m represent the images taken in the another day.

We number the 10 expressions of each volunteer as A-J. Therefore, each capital represents 40 images with the same expression from the 40 volunteers. Fig. 10 shows the 10 different expressions of a volunteer in ORL dataset.

In this experiment, we do some preprocessing for the ORL dataset. First, because the size of the original image is too large, we resize the images to 32 × 32, and then vectorize the resized images. Therefore, after the above processing, the experimental data are 1024 dimensional vectors. We collect images from A and B together as source domain data, denoted as AB. The remaining C-J are used as eight target domains, with 40 photos in each target domain. And the parameters of MSEpRKHS-DA model are set as: $\mu_1 = 10.0, \mu_2 = 1.0, \lambda = 0.4$. In addition, the dimension of subspace is 20.

TABLE 7. The performance of our model and the baseline methods on 8 domain adaption tasks of ORL dataset.

Task	Method							
	TCA[12]	SSTCA[12]	TIT[16]	IGLDA[14]	SDRKHS-DA[19]	CDSPP[20]	LPJT[18]	Ours
AB→C	0.7500	0.6250	0.7500	0.7500	0.7500	0.6250	0.6500	**0.7750**
AB→D	0.5250	0.2750	0.4750	0.5250	0.4750	0.4050	0.4700	**0.5500**
AB→E	**0.6500**	0.3250	0.5500	0.6250	0.5500	0.4250	0.5500	**0.6500**
AB→F	0.5500	0.4000	0.5250	0.5250	**0.6250**	0.5550	0.5000	0.5750
AB→G	0.5250	0.3000	**0.6000**	0.5500	**0.6000**	0.4300	0.4200	0.5250
AB→H	0.6750	0.3750	0.6500	**0.7000**	0.5750	0.5900	0.4000	**0.7000**
AB→I	**0.5750**	0.3000	**0.5750**	0.5250	0.5250	0.4200	0.5500	0.5250
AB→J	0.6500	0.4500	0.6000	**0.6750**	0.6000	0.4700	0.3500	**0.6750**
Average	0.6125	0.3813	0.5906	0.5893	0.5875	0.4900	0.4863	**0.6219**

As shown in Table 7, our model has achieved the highest domain adaption classification accuracy except AB→F, AB→G, and AB→I. But the proposed model has the highest average classification accuracy, which is increased by 0.94%, 24.06%, 3.13%, 3.26%, 3.44%, 13.19%, and 13.56% respectively compared with TCA, SSTCA, TIT IGLDA, SDRKHS-DA, CDSPP, and LPJT.

4) AR DATASET

The AR dataset[5] is an important standard dataset in face recognition. The original AR dataset contains more than 4000 RGB images of 126 persons. This experiment use the most common subset of AR dataset, which consists of

[4] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[5] http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html

2600 images taken by 100 persons on two days and 14 days apart. Each person takes 13 images each day under the changing light, face makeup, and expression. So each person has 26 images in total, and the image size is 165 × 120. In this experiment, we number 26 different photos of each person as 1.a-1.m and 2.a-2.m. The 26 images of one volunteer in the subset used are shown in Fig 11. 1.a-1.m are the 13 images taken by the volunteer on the first day, and 2 a-2.m are taken under the same conditions after 14 days.

We resize each gray image to 60 × 43. According to Fig. 11, due to the images of 1.a and 2.a are taken under normal circumstances and have more same facial features, so 1.a and 2.a are merged into source domain a. For the remaining images of 1.b-1.h and 2.b-2.h, we collect the images taken under the same conditions together as one target domain, so we can get 7 target domains b-h, and each domain has 200 images. For example, target domain b includes 1.b and 2.b images from 100 persons. And the parameters of MSEpRKHS-DA model are set as: $\mu_1 = 1.0$, $\mu_2 = 1.0$, $\lambda = 0.4$. The dimension of subspace is 30.

TABLE 8. The performance of MSEpRKHS-DA and the baseline methods on AR dataset.

Task	Method							
	TCA[12]	SSTCA[12]	TIT[16]	IGLDA[14]	SDRKHS-DA[19]	CDSPP[20]	LPJT[18]	Ours
a→b	0.9150	0.2150	0.5300	0.9200	0.9320	0.7938	0.6300	**0.9400**
a→c	0.8700	0.3100	0.5400	0.8850	**0.9290**	0.7506	0.6000	0.8500
a→d	0.5750	0.0750	0.4850	0.5850	0.5650	0.4938	0.4100	**0.6550**
a→e	0.8550	0.2150	0.5100	0.8700	0.9250	0.7369	0.7200	**0.8800**
a→f	0.8150	0.1650	0.5550	0.8150	0.7630	0.7038	0.6900	**0.8250**
a→g	0.6650	0.1450	0.4600	0.6700	0.6020	0.5669	0.5900	**0.6750**
a→h	0.5700	0.0750	0.4000	0.5700	**0.6450**	0.4731	0.5700	0.6200
Average	0.7521	0.1714	0.4971	0.7593	0.7659	0.6455	0.6014	**0.7779**

Table 8 shows the results of our proposed model and seven other methods on 7 different tasks in AR dataset. According to the results, we can find that in more than half of the tasks, our MSEpRKHS-DA model is the most outstanding. And the final average classification accuracy of the proposed model is also the best. Compared with TCA, SSTCA,TIT,IGLDA,SDRKHS-DA,CDSPP, and LPJT, our model improves the average classification accuracy by 2.57%, 60.64%, 20.07%, 1.86%, 1.2%, 13.23%, and 17.64% respectively.

C. ABLATION STUDIES FOR THE MSE CRITERION

Here, we conduct some ablation studies about kernel function and RBF kernel size for MSE criterion. First, we conduct the ablation study about kernel function on orgs→places classification task, where MSE, MMD, and CovM criteria use different kernels. From the Table 9, we can see that MSE criterion combined with different kernel functions all outperform MMD and CovM criteria, while the MSE criterion based on RBF kernel has highest classification accuracy among the all kernels.

In addition, the ablation experiment about RBF kernel size δ has been studied. We also conduct the classification task on orgs→places, where the δ of RBF kernel changes from 10^{-3} to 10^{3}. As shown in Table 10, the classification accuracy of MSE and CovM criteria varies from the RBF kernel size δ, but the classification accuracy of MMD criterion does not change. At the same time, the accuracy of MSE criterion is always higher than that of MMD and CovM in the process of δ change.

TABLE 9. The classification accuracy of the MSE, MMD, and CovM criteria on different kernels.

kernel function	MSE	MMD	CovM
RBF kernel	**0.5465**	0.4372	0.5177
Linear kernel	**0.5264**	0.4372	0.5043
Polynomial kernel	**0.5379**	0.4372	0.4938
Laplacian kernel	**0.5101**	0.4372	0.4890

TABLE 10. The classification accuracy of MSE, MMD, and CovM criteria on different RBF kernel size δ.

$log_{10}(\delta)$	MSE+RBF kernel	MMD+RBF kernel	CovM+RBF kernel
-3.0	**0.5139**	0.4372	0.4756
-2.5	**0.5139**	0.4372	0.4505
-2.0	**0.5139**	0.4372	0.4669
-1.5	**0.4746**	0.4372	0.4631
-1.0	**0.4775**	0.4372	0.4535
-0.5	**0.4861**	0.4372	0.4842
0.0	**0.5570**	0.4372	0.4593
0.5	**0.5254**	0.4372	0.4851
1.0	**0.5465**	0.4372	0.5177
1.5	**0.6347**	0.4372	0.5513
2.0	**0.5992**	0.4372	0.4775
2.5	**0.6309**	0.4372	0.5043
3.0	**0.6299**	0.4372	0.5053

So, the above ablation studies show that our criterion does have better performance of measuring the difference gap than MMD and CovM on different kernel functions and different kernel sizes.

VI. CONCLUSION

In this paper, inspired by the validity of RKHS subspace transformation, we propose a new and effective domain adaption criterion MSE, which is more valid than MMD and CovM criteria theoretically and experimentally. And then we are first to propose the domain adaption framework based on the progressive RKHS subspace learning pRKHS-DA, which gradually reduces the distribution difference between source and target domain by constantly learning new subspace. Finally, combining MSE criterion with pRKHS-DA framework, we propose the MSEpRKHS-DA model. And the experiment results on four dataset show that our model outperforms some state-of-the-art methods.

APPENDIX A

$$\begin{aligned} E\left[|Y_i|^2\right] &= E\left[\left|\langle \varphi(X), \vartheta_i\rangle_{H_2}\right|^2\right] \\ &= \int_{\Omega} \left|\langle \varphi(x), \vartheta_i\rangle_{H_2}\right|^2 p(x)\,\mathrm{d}x \end{aligned}$$

$$\begin{aligned}
&\leq \int_{\Omega} \left|\langle \varphi(x), \vartheta_i \rangle_{H_2}\right|^2 \mathrm{d}x \\
&= \int_{\Omega} \left|\left\langle \varphi(x), \sum_{j=1}^{N} \omega_{ji}\varphi(x_j)\right\rangle_{H}\right|^2 \mathrm{d}x \\
&= \int_{\Omega} \left|\sum_{j=1}^{N} \omega_{ji}\left\langle \varphi(x), \varphi(x_j)\right\rangle_{H_2}\right|^2 \mathrm{d}x \\
&= \int_{\Omega} \left|\sum_{j=1}^{N} \omega_{ji} k(x, x_j)\right|^2 \mathrm{d}x \\
&\leq \sum_{p=1}^{N}\sum_{q=1}^{N} \left|\omega_{pi}\omega_{qi}\right| \left|\int_{\Omega} k(x, x_q)\, k(x, x_p)\, \mathrm{d}x\right| \\
&\leq \sum_{p=1}^{N}\sum_{q=1}^{N} \left|\omega_{pi}\omega_{qi}\right| \\
&\quad \times \sqrt{\int_{\Omega} k^2(x, x_q)\, \mathrm{d}x}\sqrt{\int_{\Omega} k^2(x, x_p)\, \mathrm{d}x} \\
&< +\infty,
\end{aligned}$$

where the X obeys $p(x)$ and $0 \leq p(x) \leq 1$, and $k(x_q, x) \in H_2$ and $k(x_p, x) \in H_2$ are square integrable.

APPENDIX B

$$\begin{aligned}
MSE(Y_s, Y_t) &= \frac{1}{n_s n_t} \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \left\| y_i^s - y_j^t \right\|^2 \\
&= \frac{1}{n_s n_t} \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \left\| W^T \left(\tilde{k}_i - \tilde{k}_{(n_s+j)}\right) \right\|^2 \\
&= \frac{1}{n_s n_t} \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \left\| W^T \varphi_{ij} \right\|^2 \\
&= \frac{1}{n_s n_t} \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} tr\left(W^T \varphi_{ij}\varphi_{ij}^T W\right) \\
&= tr\left(W^T \left(\frac{1}{n_s n_t} \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \varphi_{ij}\varphi_{ij}^T\right) W\right) \\
&= tr\left(W^T \Psi W\right) \qquad (13)
\end{aligned}$$

where $\tilde{k}_i = \varphi(x_i^s)$ and $\tilde{k}_{n_s+i} = \varphi(x_j^t)$, $\varphi_{ij} = \tilde{k}_i - \tilde{k}_{(n_s+j)}$, and $\Psi = \frac{1}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \varphi_{ij}\varphi_{ij}^T$.

REFERENCES

[1] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, pp. 1345–1359, Nov. 2010, doi: 10.1109/TKDE.2009.191.

[2] J. Ghosn and Y. Bengio, "Bias learning, knowledge sharing," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Networks, (IJCNN), Neural Comput., New Challenges Perspect. New Millennium*, Jul. 2000, pp. 9–14, doi: 10.1109/IJCNN.2000.857806.

[3] S. Ozawa, A. Roy, and D. Roussinov, "A multitask learning model for online pattern recognition," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 430–445, Mar. 2009, doi: 10.1109/TNN.2008.2007961.

[4] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015, doi: 10.1109/MSP.2014.2347059.

[5] B. Chen, W. Lam, I. W. Tsang, and T.-L. Wong, "Discovering low-rank shared concept space for adapting text mining models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1284–1297, Jun. 2013, doi: 10.1109/TPAMI.2012.243.

[6] F. Zhuang, P. Luo, C. Du, Q. He, Z. Shi, and H. Xiong, "Triplex transfer learning: Exploiting both shared and distinct concepts for text classification," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1191–1203, Jul. 2014, doi: 10.1109/TCYB.2013.2281451.

[7] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2288–2302, Nov. 2014, doi: 10.1109/TPAMI.2013.249.

[8] B. S. J. P. T. Hofmann, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, May 2007, pp. 513–520.

[9] L. Li and Z. Zhang, "Semi-supervised domain adaptation by covariance matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2724–2739, Nov. 2019, doi: 10.1109/TPAMI.2018.2866846.

[10] Z. Zhang, M. Wang, and A. Nehorai, "Optimal transport in reproducing kernel Hilbert spaces: Theory and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1741–1754, Jul. 2020, doi: 10.1109/TPAMI.2019.2903050.

[11] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. AAAI*, vol. 8, 2008, pp. 677–682.

[12] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011, doi: 10.1109/TNN.2010.2091281.

[13] A. Gretton, "Measuring statistical dependence with Hilbert–Schmidt norms," in *Proc. Int. Conf. Algorithmic Learn. Theory*. Berlin, Germany: Springer, 2005, pp. 1–15.

[14] M. Jiang, W. Huang, Z. Huang, and G. G. Yen, "Integration of global and local metrics for domain adaptation learning via dimensionality reduction," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 38–51, Dec. 2017, doi: 10.1109/TCYB.2015.2502483.

[15] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 288–299, Jan. 2018, doi: 10.1109/TCYB.2016.2633306.

[16] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2144–2155, Jun. 2019, doi: 10.1109/TCYB.2018.2820174.

[17] M. Xiao and Y. Guo, "Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation," in *Machine Learning and Knowledge Discovery in Databases*. Cham, Switzerland: Springer, 2015, pp. 525–540.

[18] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6103–6115, Dec. 2019, doi: 10.1109/TIP.2019.2924174.

[19] W. Lei, Z. Ma, Y. Lin, and W. Gao, "Domain adaption based on source dictionary regularized RKHS subspace learning," *Pattern Anal. Appl.*, vol. 24, no. 4, pp. 1513–1532, Nov. 2021, doi: 10.1007/s10044-021-01002-x.

[20] Q. Wang and T. P. Breckon, "Cross-domain structure preserving projection for heterogeneous domain adaptation," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108362, doi: 10.1016/j.patcog.2021.108362.

[21] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2003, pp. 1–10.

[22] S. Saitoh and Y. Sawano, *Theory of Reproducing Kernels and Applications*. Singapore: Springer, 2016.

[23] K. Yosida, *Functional Analysis*. Cham, Switzerland: Springer, 2012.

[24] I. V. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, vol. 152. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[25] I. Steinwart, D. Hush, and C. Scovel, "An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4635–4643, Oct. 2006.

[26] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proc. European Conf. Comput. Vis.* Berlin, Germany: Springer, 2006.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

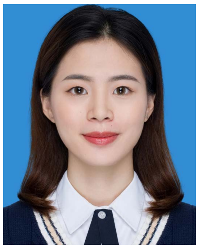
[28] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994, doi: 10.1109/34.291440.

[29] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073, doi: 10.1109/CVPR.2012.6247911.

[30] L. Yaohui, M. Zhengming, and Y. Fang, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy," *Knowl.-Based Syst.*, vol. 133, pp. 208–220, May 2017.

[31] H. Liu and W.-S. Chen, "A novel random projection model for linear discriminant analysis based face recognition," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, Jul. 2009, pp. 112–117, doi: 10.1109/ICWAPR.2009.5207431.

[32] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1410–1417, doi: 10.1109/CVPR.2014.183.



YANZHEN QIU received the B.S. degree in electronic information engineering from Shenzhen University, Shenzhen, China, in 2020. She is currently pursuing the master's degree with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China. Her current research interest includes domain adaption.

SHUYU LIU received the Ph.D. degree from Sun Yat-sen University, China, in 2010. He is currently the Director of the Experimental Teaching Center and the School of Pharmacy, Shenzhen, Sun Yat-sen University. His research interests include biological image processing, biomedical engineering, and intelligent medicine.

ZHENGMING MA received the B.Sc. degree in radio technology and the M.Sc. degree in electronic and communication system from the South China University of Technology, Guangzhou, China, in 1982 and 1985, respectively, and the Ph.D. degree in pattern recognition and intelligent control from Tsinghua University, Beijing, China, in 1989. He is currently a Professor with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou. His current research interest includes machine learning.

HUI HE received the B.S. degree in electronic information engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2020. She is currently pursuing the master's degree with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China. Her current research interest includes machine learning.