Rethinking Joint Maximum Mean Discrepancy for Domain Adaptation

Wei Wang[1] **Haifeng Xia**[1] **Chao Huang**[1*] **Zhengming Ding**[2]
Cong Wang[3] **Haojie Li**[4] **Xiaochun Cao**[1]
[1]Shenzhen Campus of Sun Yat-sen University [2]Department of Computer Science, Tulane University
[3]University of California, San Francisco [4]Shandong University of Science and Technology
{wangwei29, xiahf5, huangch253, caoxiaochun}@mail.sysu.edu.cn zding1@tulane.edu
supercong94@gmail.com hjli@sdust.edu.cn

Abstract

In domain adaption (DA), joint maximum mean discrepancy (JMMD), as a famous distribution-distance metric, aims to measure joint probability distribution difference between the source domain and target domain, while it is still not fully explored and especially hard to be applied into a subspace-learning framework as its empirical estimation involves a tensor-product operator whose partial derivative is difficult to obtain. To solve this issue, we deduce a concise JMMD based on the Representer theorem that avoids the tensor-product operator and obtains two essential findings. First, we reveal the uniformity of JMMD by proving that previous marginal, class conditional, and weighted class conditional probability distribution distances are three special cases of JMMD with different label reproducing kernels. Second, inspired by graph embedding, we observe that the similarity weights, which strengthen the intra-class compactness in the graph of Hilbert Schmidt independence criterion (HSIC), take opposite signs in the graph of JMMD, revealing why JMMD degrades the feature discrimination. This motivates us to propose a novel loss JMMD-HSIC by jointly considering JMMD and HSIC to promote discrimination of JMMD. Extensive experiments on several cross-domain datasets could demonstrate the validity of our revealed theoretical results and the effectiveness of our proposed JMMD-HSIC.

1 Introduction

Domain adaptation (DA) has emerged as an effective technology to solve the well-known problem of domain discrepancy that frequently occurs in reality [1–5]. Many promising approaches have been suggested to mitigate this issue from different perspectives [6–10]. A major issue in DA is how to formulate a favorable probability distribution distance that can be applied to measure the proximity of two different probability distributions, thus numerous probability distribution-distance metrics have been proposed over the years. For example, the Quadratic and Kullback-Leibler distances derived from the Bregman divergence and generated by different convex functions are introduced to match two different probability distributions explicitly [11]. However, extending them into different models may be inflexible since they are parametric and require an intermediate density estimation [12]. The Wasserstein distance derived from the optimal transport problem exploits a transportation plan to align two different marginal [13], class conditional [14] or joint [15] probability distributions, but it might be inconvenient to be applied into a subspace-learning framework because it often comes down to a complex bi-level optimization problem [16].

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

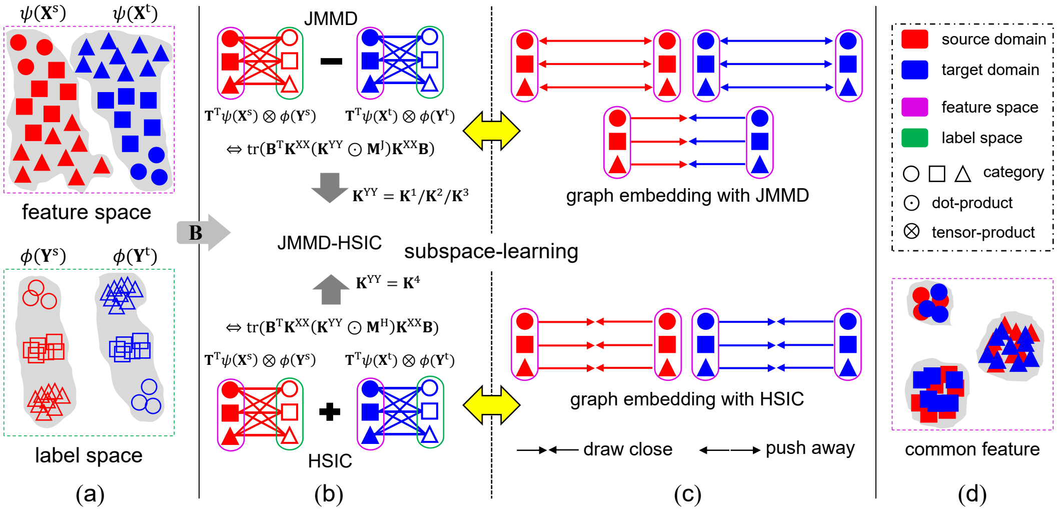

$\psi(\mathbf{X}^s)$
$\psi(\mathbf{X}^t)$
feature space
$\phi(\mathbf{Y}^s)$
$\phi(\mathbf{Y}^t)$
label space
(a)
B
JMMD
$\mathbf{T}^T\psi(\mathbf{X}^s) \otimes \phi(\mathbf{Y}^s)$
$\mathbf{T}^T\psi(\mathbf{X}^t) \otimes \phi(\mathbf{Y}^t)$
$\Leftrightarrow \mathrm{tr}(\mathbf{B}^T\mathbf{K}^{XX}(\mathbf{K}^{YY} \odot \mathbf{M}^J)\mathbf{K}^{XX}\mathbf{B})$
$\mathbf{K}^{YY} = \mathbf{K}^1/\mathbf{K}^2/\mathbf{K}^3$
JMMD-HSIC
subspace-learning
$\mathbf{K}^{YY} = \mathbf{K}^4$
$\Leftrightarrow \mathrm{tr}(\mathbf{B}^T\mathbf{K}^{XX}(\mathbf{K}^{YY} \odot \mathbf{M}^H)\mathbf{K}^{XX}\mathbf{B})$
$\mathbf{T}^T\psi(\mathbf{X}^s) \otimes \phi(\mathbf{Y}^s)$
$\mathbf{T}^T\psi(\mathbf{X}^t) \otimes \phi(\mathbf{Y}^t)$
HSIC
(b)
graph embedding with JMMD
graph embedding with HSIC
draw close
push away
(c)
source domain
target domain
feature space
label space
category
dot-product
tensor-product
common feature
(d)

Figure 1: The overview of our revealed theoretical results and proposed JMMD-HSIC. (a) We map the features (upper part) and labels (lower part) of the source and target domains into the RKHS, respectively; (b) The application of JMMD (upper part) and HSIC (lower part) in a subspace-learning framework; (c) The graph embedding interpretation of JMMD (upper part) and HSIC (lower part) in a subspace-learning framework; (d) Learned feature representations of the source and target domains after subspace learning. '$-$' in the module of JMMD means the feature-label dependence difference, and '$+$' in the module of HSIC means separately considering feature-label dependence in the two domains. ψ and ϕ are feature and label mappings for a RKHS. $\mathbf{T}$ and $\mathbf{B}$ are projection matrices for a projected RKHS. $\mathbf{T}\psi(\mathbf{X}) \otimes \phi(\mathbf{Y})$ is a tensor-product operator for the covariance. $\mathrm{tr}(\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J/H}})\mathbf{K}^{\mathrm{XX}}\mathbf{B})$ is the deduced concise JMMD/HSIC in a projected RKHS.

Gretton *et al.* propose a metric of maximum mean discrepancy (MMD), which empirically estimates the distance between two probability distributions in a reproducing kernel Hilbert space (RKHS) [17]. Due to its simplicity and solid theoretical foundation, it has been applied into a wide range of problems, such as deep generative models [18] and variational autoencoders [19], *etc*. Although MMD could establish marginal [12], class conditional [20], and weighted class conditional [21] probability distribution distances, it is still not fully explored about the joint probability distribution distance (*i.e.*, joint maximum mean discrepancy, JMMD) and is hard to be applied into a subspace-learning framework, as its empirical estimation involves a tensor-product operator whose partial derivative is difficult to obtain.

Specifically, when JMMD is considered in a subspace-learning framework, we project data features and their corresponding labels into a RKHS (Fig. 1(a)), and aim to exploit a projected RKHS by a feature projection matrix $\mathbf{T}$ with infinite dimensions (the upper part of Fig. 1(b)). In the projected RKHS or the common feature space, the feature-label dependence difference between the source domain and target domain is minimized, and their joint probability distributions [22] are aligned accordingly (Fig. 1(d)). Here, the covariance that involves a tensor-product operator describes the feature-label dependence. However, one issue remains to be overcome in this process: the partial derivative with respect to the infinite-dimensional $\mathbf{T}$ is hard to obtain. In this paper, based on the Representer theorem [23], we deduce a concise JMMD that avoids the tensor-product operator and optimizes a finite-dimensional $\mathbf{B}$ instead of $\mathbf{T}$, where the Representer theorem represents $\mathbf{T}^\top\psi(\mathbf{x})$ by finite values of a kernel function with corresponding coefficients from $\mathbf{B}$, and some matrix operation properties make the tensor-product disappear. Therefore, the derivative can be easily obtained.

With the concise JMMD, we obtain two essential findings. Firstly, we reveal the uniformity of JMMD by proving that previous popular marginal, class conditional, and weighted class conditional probability distribution distances are its three special cases with label kernels $\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$, and we also prove that they are reproducing kernels. This finding could provide theoretical guidance to refine JMMD by designing label kernels for different problems in DA. Moreover, recent work indicates that the procedure of distribution alignment degrades the feature discrimination unexpectedly [24, 25] but lacks theoretical support. To reveal this, similar to JMMD, we deduce a concise Hilbert Schmidt independence criterion (HSIC) [26], which models the feature-label dependence and maximizes the dependence to improve feature discrimination (*the features have better discrimination ability if the feature-label dependence is larger and vice versa*) (the lower part of Fig. 1(b)). We design a label reproducing kernel $\mathbf{K}^4$ for HSIC to strengthen the intra-class compactness or feature discrimination. Then, we explore the relationship between JMMD and HSIC inspired by the graph embedding viewpoint to better understand the reason for feature discrimination degradation in JMMD.

Specifically, we observe that the similarity weights, which strengthen intra-class compactness in the graph of HSIC, take opposite signs in the graph of JMMD. Thus, JMMD degrades the discrimination unexpectedly. As shown in Fig. 1(c), in the feature space, JMMD tries to *push data points from the same classes in the same domain further* and draw those from the same classes in different domains closer. In contrast, in the feature space, HSIC aims to *draw data points from the same classes in the same domain closer* (intra-class compactness). To this end, we propose a novel loss JMMD-HSIC by jointly considering JMMD and HSIC. Therefore, we can improve the discrimination of JMMD, leading to a better DA capacity (Fig. 1(d)). The whole pipeline of this paper is briefly depicted in Fig. 1, and our main contributions are summarized below,

- Based on the Representer theorem, we deduce a concise JMMD that avoids the tensor-product operator, and the derivative can be easily obtained so that it can be applied into a subspace-learning framework.
- With the concise JMMD, we reveal its uniformity by proving that previous distribution distances are its special cases with different label reproducing kernels. This finding yields theoretical guidance for refining JMMD by designing label kernels for different problems.
- To better understand the reason for feature discrimination degradation in JMMD, we reveal the relationship between JMMD and HSIC inspired by graph embedding. Then, a novel loss JMMD-HSIC is proposed to promote discrimination of JMMD.

2 Related Work

2.1 Maximum Mean Discrepancy

Gretton *et al.* introduce a probability distribution-distance metric MMD [17]. Along this direction, Pan *et al.* incorporate MMD into a subspace-learning framework to learn some transfer components [12] across the source domain and target domain. Duan *et al.* embed MMD into a multiple kernel learning framework to jointly learn a kernel function and a robust classifier [27] by minimizing the structural risk functional and the distribution mismatch. Long *et al.* down-weighted source instances irrelevant to target ones to realize more positive knowledge transfer based on MMD [28]. Ghifary *et al.* employ MMD to deal with distribution bias in a simple neural network [29]. Tzeng *et al.* propose a domain confusion loss based on MMD in a deep neural network [30]. Long *et al.* present a deep adaptation network where the multi-kernel MMD is applied into all task-specific layers [31].

To model more complicated probability distribution distances, Long *et al.* propose a class-wise MMD to align class conditional probability distributions [20], which can be conveniently optimized in many subspace-learning frameworks [32, 33]. To deal with an imbalanced dataset, Wang *et al.* establish a weighted class-wise MMD where class prior probabilities are introduced [21]. Concerning the label distribution shift problem, Yan *et al.* construct a weighted MMD where source instances are multiplied by special weights [34]. Moreover, Wang *et al.* raise a dynamic balanced MMD to quantitatively account for relative importance between the marginal and class conditional probability distribution distances [35]. Deng *et al.* propose an extended MMD to simultaneously minimize the intra-class dispersion and maximize the inter-class compactness [36].

Zhang *et al.* estimate the joint probability distribution discrepancy based on the Bayesian law and propose a discriminative joint probability distribution discrepancy [37]. In contrast, Long *et al.* estimate the uncentered feature-label covariance in the source and target domains, and consider the difference in covariance between the two domains as the joint probability distribution discrepancy. Then, they propose JMMD and construct a transfer network accordingly [22]. However, JMMD is still not fully explored and is very hard to be applied to a subspace-learning framework as it involves a tensor-product operator whose partial derivative is difficult to obtain. This paper theoretically deduces a concise JMMD and obtains two essential findings.

2.2 Hilbert Schmidt Independence Criterion

HSIC measures dependence between two random variables [26]. To preserve important data features, Dorri *et al.* regard original and transformed data features as two random variables and maximize their dependence during distribution alignment [38, 39]. Yan *et al.* minimize dependence between projected features and domain features (*e.g.*, background information) to learn a robust domain-invariant subspace [40]. Inspired by classifier adaptation, Wang *et al.* propose a projected HSIC to maximize feature-label dependence in a reconstruction DA framework [41]. Similarly, we aim to maximize the feature-label dependence while we further explore the relationship between JMMD and HSIC, and design a label reproducing kernel $\mathbf{K}^4$ for HSIC in this work. Besides, HSIC is computed on the whole domains [41] while we separately compute HSIC on the source and target domains.

2.3 Graph Embedding

Graph embedding as a prevalent technology has been applied to many problems due to its advantage in discovering complicated data structures and relationships, such as subspace-learning methods [42, 43] and graph convolutional networks [44, 45], *etc*. A lot of subspace-learning-based DA approaches aim to establish a similarity graph and exploit a feature mapping to embed the graph vertices into a desirable low-dimensional subspace so that some important data structures (local manifold structure [35], local discrimination structure [24], *etc.*) could be respected elegantly during the distribution alignment process. This paper does not focus on leveraging the graph embedding technique in the DA problem. Instead, we analyze the relationship between JMMD and HSIC inspired by graph embedding to illustrate the reason for feature discrimination degradation in JMMD.

3 Rethinking JMMD

In this paper, the bold uppercase letter $\mathbf{X}$ denotes a matrix, while the bold lowercase letter $\mathbf{x}$ is a column vector of $\mathbf{X}$. Moreover, $\mathbf{x}_{\rm i}$ is the i-th column vector of $\mathbf{X}$ and $\rm x_{ij}$ is the value of i-th row and j-th column of $\mathbf{X}$. $\otimes$ and $\odot$ are the tensor-product and dot-product operators.

Domain Adaptation. Given a labeled source domain $\mathcal{D}^{\rm s} := (\mathbf{X}^{\rm s}, \mathbf{Y}^{\rm s})$ and an unlabeled target domain $\mathcal{D}^{\rm t} := (\mathbf{X}^{\rm t}, \mathbf{Y}^{\rm t})$, where $\mathbf{X}^{\rm s/t} \in \mathbb{R}^{\rm m \times n^{s/t}}$, $\mathbf{Y}^{\rm s/t} \in \mathbb{R}^{\rm C \times n^{s/t}}$, 'm' is the feature dimension, '$\rm n^{s/t}$' is the number of source/target instances ($\rm n^s + n^t = n$) and 'C' is the number of shared categories. DA assumes that the two domains follow different joint probability distributions, *i.e.*, $\mathcal{P}^{\rm s}(\mathbf{X}^{\rm s}, \mathbf{Y}^{\rm s}) \neq \mathcal{P}^{\rm t}(\mathbf{X}^{\rm t}, \mathbf{Y}^{\rm t})$, but share the same feature and label spaces, and $\mathbf{Y}^{\rm t}$ is not available during the training process. $\mathbf{Y}^{\rm s/t}$ is the probability soft label, *i.e.*, the probability of a data sample $\mathbf{x}_{\rm i}$ that belongs to a class 'c' is $\rm y_{ci}$. Notably, the hard label (one-hot label) is a special case of $\mathbf{Y}^{\rm s/t}$ where $\rm y_{ci}$ =1 and $\rm y_{\bar{c}i}$ =0 ($\rm \bar{c} \neq$c) if $\mathbf{x}_{\rm i}$ belongs to the c-th class. DA aims to design a distribution-distance metric to minimize the divergence between their joint probability distributions so that the classifier trained on a source domain can be generalized into another target domain.

Reproducing Kernel Hilbert Space. RKHS is a Hilbert space ($\mathcal{H}$) of function $f : \Omega \rightarrow \mathbb{R}$ on a domain Ω, and its inner-product and Hilbert-Schmidt norm are $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $|| \cdot ||_{\mathcal{H}}$. The evaluation functional $f(\mathbf{x})$ can be reproduced by a reproducing kernel function $k(\mathbf{x}, \mathbf{x}^\top)$, *i.e.*, $\langle f(\cdot), k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = f(\mathbf{x})$, and the RKHS takes its name from this so-called reproducing kernel function. $k(\mathbf{x}, \cdot)$ can be viewed as an implicit mapping $\psi(\mathbf{x})$ (infinite dimensions) where $k(\mathbf{x}, \mathbf{x}^\top) = \langle \psi(\mathbf{x}), \psi(\mathbf{x}^\top) \rangle_{\mathcal{H}}$.

3.1 A Concise JMMD

In DA, JMMD as a famous probability distribution-distance metric is still not fully explored and hard to be applied into a subspace-learning framework as its empirical estimation involves a tensor-product operator whose partial derivative is nontrivial to obtain [22]. In this section, we first deduce a concise JMMD and reveal its uniformity. Then, we deduce a concise HSIC to explore the relationship between JMMD and HSIC inspired by graph embedding, revealing why JMMD degrades the feature discrimination. Finally, we propose a novel loss JMMD-HSIC to improve the discrimination of JMMD and apply it into a general subspace-learning framework.

Given a domain $\mathcal{D}$ sampled from a joint probability distribution $\mathcal{P}(\mathbf{X}, \mathbf{Y})$ [46], we project the data feature $\mathbf{x}$ and its corresponding label $\mathbf{y}$ into a RKHS, *i.e.*, $\psi(\mathbf{x})$ and $\phi(\mathbf{y})$, where ψ and ϕ are the feature and label mappings. Then, we utilize the uncentered covariance between the feature and label to represent a joint probability distribution, *i.e.*, $\mathcal{C}_{\mathrm{XY}} := \mathbb{E}_{\mathrm{XY}}(\psi(\mathbf{x}) \otimes \phi(\mathbf{y}))$. Here, we estimate the covariance $\mathcal{C}_{\mathrm{XY}}$ of domain $\mathcal{D}$ with the expectation $\mathbb{E}_{\mathrm{XY}}$ or mean μ_{XY} of all samples' covariance $\psi(\mathbf{x}) \otimes \phi(\mathbf{y})$ in a RKHS. The covariance essentially describes the dependence between feature and label. Due to the impossibility of obtaining all possible samples in domain $\mathcal{D}$ (an infinite number), JMMD [22] adopts the maximum likelihood estimate principle and utilizes finite samples to empirically estimate $\mu_{\mathrm{X^sY^s}}$ and $\mu_{\mathrm{X^tY^t}}$ for source domain and target domain, and then minimizes the loss of $||\mu_{\mathrm{X^sY^s}} - \mu_{\mathrm{X^tY^t}}||^2_{\mathcal{H}}$ to draw joint probability distributions of the two domains closer. Formally, JMMD and its concise form in a RKHS is defined as below,

$$\mathbb{D}_{\mathcal{H}}(\mathcal{P}^{\mathrm{s}}(\mathbf{X}^{\mathrm{s}}, \mathbf{Y}^{\mathrm{s}}), \mathcal{P}^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}, \mathbf{Y}^{\mathrm{t}})) = \left|\left|\mu_{\mathrm{X^sY^s}} - \mu_{\mathrm{X^tY^t}}\right|\right|^2_{\mathcal{H}} = \mathrm{tr}(\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J}})), \tag{1}$$

where $\mathbf{K}^{\mathrm{XX}}, \mathbf{K}^{\mathrm{YY}} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$ are the kernel matrices and they are computed by $\mathrm{k}^{\mathrm{XX}}_{\mathrm{ij}} = k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}^{\top})$, $\mathrm{k}^{\mathrm{YY}}_{\mathrm{ij}} = k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{i}}, \mathbf{y}_{\mathrm{j}}^{\top})$. k^{X} and k^{Y} are feature and label kernels. $\mathbf{M}^{\mathrm{J}} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$ is the MMD matrix for JMMD. Remarkably, the nonlinear functions ψ and ϕ do not need to be explicit, and the tensor-product operator disappears, more details could be found in ***Section A of the supplementary material***. To incorporate JMMD into a subspace-learning framework, we deduce the concise form of JMMD in a projected RKHS as below,

$$||\tfrac{1}{\mathrm{n^s}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n^s}}(\mathbf{T}^\top \psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}}) \otimes \phi(\mathbf{y}_{\mathrm{i}}^{\mathrm{s}})) - \tfrac{1}{\mathrm{n^t}} \textstyle\sum_{\mathrm{j}=1}^{\mathrm{n^t}}(\mathbf{T}^\top \psi(\mathbf{x}_{\mathrm{j}}^{\mathrm{t}}) \otimes \phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{t}}))||^2_{\mathcal{H}} = \mathrm{tr}(\mathbf{B}^\top \mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J}})\mathbf{K}^{\mathrm{XX}}\mathbf{B}). \tag{2}$$

As shown in the left side of Eq. (2), the dimension of feature projection matrix $\mathbf{T} \in \mathbb{R}^{\infty \times \mathrm{d}}$ is infinite since ψ is an infinite mapping, and a tensor-product operator is involved so that it is nontrivial to obtain the partial derivative with respect to **T**. To overcome this issue, we utilize the Representer theorem and some matrix operation properties to obtain the right side of Eq. (2). Remarkably, the infinite-dimensional **T** does not need to be optimized because we resort to optimize a finite-dimensional **B**, and the tensor-product operator disappears. Therefore, it is easy to obtain the partial derivative with respect to **B**, and JMMD could be applied into a subspace-learning framework. More details about the proof of Eq. (2) can be found in ***Section B of the supplementary material***.

3.2 The Uniformity of JMMD

In this section, we introduce a theorem to illustrate that JMMD is a unified form of existing popular marginal, class conditional, and weighted class conditional probability distribution distances.

Theorem 1 *The marginal, class conditional, and weighted class conditional probability distribution distances are three special cases of JMMD with label reproducing kernels* $\boldsymbol{K}^1$, $\boldsymbol{K}^2$ *and* $\boldsymbol{K}^3$*, and more details about these three distances could be found in* ***Section C of the supplementary material****.* $\boldsymbol{K}^1 = \boldsymbol{1}_{n \times n}$ *is a matrix whose elements are all 1 with size of* $n \times n$*, and* $\boldsymbol{K}^2$, $\boldsymbol{K}^3$ *are defined as below,*

$$\mathrm{k}^2_{\mathrm{ij}} = \begin{cases} (\mathrm{n^sn^s})/(\mathrm{n^{s,c}n^{s,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ (\mathrm{n^tn^t})/(\mathrm{n^{t,c}n^{t,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ (\mathrm{n^sn^t})/(\mathrm{n^{s,c}n^{t,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ (\mathrm{n^tn^s})/(\mathrm{n^{t,c}n^{s,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ 0, \quad \text{otherwise}, \end{cases} \qquad \mathrm{k}^3_{\mathrm{ij}} = \begin{cases} 1, & \mathbf{x}_{\mathrm{i}} \in D^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ 1, & \mathbf{x}_{\mathrm{i}} \in D^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ 1, & \mathbf{x}_{\mathrm{i}} \in D^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ 1, & \mathbf{x}_{\mathrm{i}} \in D^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ 0, \quad \text{otherwise}, \end{cases} \tag{3}$$

where the superscript 's/t,c' denotes data points from the c-th class in the source/target domain. The proof of this theorem could be found in ***Section D of the supplementary material***, and we also prove that $\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$ are the reproducing kernels in ***Section E of the supplementary material***. Notably, Theorem 1 yields theoretical guidance to refine JMMD by designing more delicate label kernels for different problems in DA, and we will leave this open direction in our future work.

3.3 A Concise HSIC

HSIC [26] also utilizes the covariance to establish the feature-label dependence, and aims to maximize the dependence for a given domain to improve its feature discrimination. A problem is that the domain-specific dependence may be decreased or the feature discrimination is degraded unexpectedly when minimizing JMMD. In the next section, we will illustrate the reason for discrimination degradation in JMMD from the graph embedding viewpoint. Following JMMD, we utilize finite samples to empirically estimate $\mathcal{C}_{\mathrm{X}^{\mathrm{s}}\mathrm{Y}^{\mathrm{s}}}$ and $\mathcal{C}_{\mathrm{X}^{\mathrm{t}}\mathrm{Y}^{\mathrm{t}}}$, and maximize these two terms separately. Similarly, HSIC involves a tensor-product operator whose derivative is hard to obtain, thus we deduct a concise HSIC in an RKHS and a concise HSIC in a projected RKHS as follows,

$$\mathrm{tr}(\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{H}})), \quad \mathrm{tr}(\mathbf{B}^{\top}\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{H}})\mathbf{K}^{\mathrm{XX}}\mathbf{B}). \tag{4}$$

Notably, the concise HSIC is consistent with JMMD. Thus, it is easy to jointly consider them in a concise form, which will be introduced in 3.5. Besides, we design a label kernel for HSIC as below,

$$\mathrm{k}^{4}_{\mathrm{ij}} = \begin{cases} (-\mathrm{n}^{\mathrm{s}}\mathrm{n}^{\mathrm{s}})/(\mathrm{n}^{\mathrm{s,c}}\mathrm{n}^{\mathrm{s,c}}), & \mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}}, \mathrm{i} \neq \mathrm{j} \\ \big(\mathrm{n}^{\mathrm{s}}\mathrm{n}^{\mathrm{s}}(\mathrm{n}^{\mathrm{s,c}} - 1)\big)/\big(\mathrm{n}^{\mathrm{s,c}}\mathrm{n}^{\mathrm{s,c}}\big), & \mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}}, \mathrm{i=j} \\ (-\mathrm{n}^{\mathrm{t}}\mathrm{n}^{\mathrm{t}})/(\mathrm{n}^{\mathrm{t,c}}\mathrm{n}^{\mathrm{t,c}}), & \mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}}, \mathrm{i} \neq \mathrm{j} \\ \big(\mathrm{n}^{\mathrm{t}}\mathrm{n}^{\mathrm{t}}(\mathrm{n}^{\mathrm{t,c}} - 1)\big)/\big(\mathrm{n}^{\mathrm{t,c}}\mathrm{n}^{\mathrm{t,c}}\big), & \mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}}, \mathrm{i=j} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

There are some advantages with $\mathbf{K}^{4}$: 1) (4) will be a minimization problem which can be easily analyzed from the graph embedding viewpoint; 2) In 3.4, we will illustrate that the intra-class compactness (discrimination) of source domain and target domain can be improved with label kernel $\mathbf{K}^{4}$; 3) $\mathbf{K}^{4}$ is a reproducing kernel which is proved in ***Section E of the supplementary material***.

3.4 A Graph Embedding Viewpoint

In this section, we reveal that JMMD degrades the feature discrimination from the graph embedding viewpoint. Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_{\mathrm{n}}]$, we establish a nearest neighbor graph G with n vertices, where each vertex denotes a data point. Let $\mathbf{W}$ be the weight matrix of G, and w_{ij} measures the similarity weight between $\mathbf{x}_{\mathrm{i}}$ and $\mathbf{x}_{\mathrm{j}}$ in original feature space (the larger w_{ij} is and the closer they are and vice versa). Graph embedding technique [42] tries to find a desirable feature representation of $\mathbf{X}$ so that it could respect the relationship between each two data points in the original feature space. Formally, it aims to minimize the following objective function,

$$\textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}}(\mathrm{w}_{\mathrm{ij}}||\mathbf{z}_{\mathrm{i}} - \mathbf{z}_{\mathrm{j}}||_2^2) = \mathrm{tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\top}), \tag{6}$$

where $\mathbf{Z}$ is the embedding representation of $\mathbf{X}$. $\mathbf{L} = \mathbf{Q} - \mathbf{W}$ is the graph Laplacian matrix where $\mathbf{Q} = \mathbf{diag}(\mathbf{q}_1, \cdots, \mathbf{q}_{\mathrm{n}})$ is a diagonal matrix and $\mathbf{q}_{\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{ij}}$. Inspired by graph embedding, we regard $\mathbf{L}^{\mathrm{J}} = \mathbf{K}^{2} \odot \mathbf{M}^{\mathrm{J}}$ and $\mathbf{L}^{\mathrm{H}} = \mathbf{K}^{4} \odot \mathbf{M}^{\mathrm{H}}$ as two graph Laplacian matrices. Then, we reveal the similarity weight matrices $\mathbf{W}^{\mathrm{J}}$ and $\mathbf{W}^{\mathrm{H}}$ of JMMD and HSIC as follows,

$$\mathrm{w}^{\mathrm{J}}_{\mathrm{ij}} = \begin{cases} -1/(\mathrm{n}^{\mathrm{s,c}}\mathrm{n}^{\mathrm{s,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ -1/(\mathrm{n}^{\mathrm{t,c}}\mathrm{n}^{\mathrm{t,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ 1/(\mathrm{n}^{\mathrm{s,c}}\mathrm{n}^{\mathrm{t,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ 1/(\mathrm{n}^{\mathrm{t,c}}\mathrm{n}^{\mathrm{s,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ 0, & \text{otherwise,} \end{cases} \qquad \mathrm{w}^{\mathrm{H}}_{\mathrm{ij}} = \begin{cases} 1/(\mathrm{n}^{\mathrm{s,c}}\mathrm{n}^{\mathrm{s,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s,c}} \\ 1/(\mathrm{n}^{\mathrm{t,c}}\mathrm{n}^{\mathrm{t,c}}), & \mathbf{x}_{\mathrm{i}} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t,c}} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Then, the concise JMMD (2) and HSIC (4) in a projected RKHS can be rewritten as follows,

$$\textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}}(\mathrm{w}_{\mathrm{ij}}^{\mathrm{J}}||\mathbf{B}^\top\mathbf{k}_{\mathrm{i}}^{\mathrm{XX}}-\mathbf{B}^\top\mathbf{k}_{\mathrm{j}}^{\mathrm{XX}}||_{\mathrm{F}}^2),\quad \sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\Big(\mathrm{w}_{\mathrm{ij}}^{\mathrm{H}}||\mathbf{B}^\top\mathbf{k}_{\mathrm{i}}^{\mathrm{XX}}-\mathbf{B}^\top\mathbf{k}_{\mathrm{j}}^{\mathrm{XX}}||_{\mathrm{F}}^2), \tag{8}$$

where $\mathbf{B}^\top\mathbf{k}^{\mathrm{XX}}$ is the embedded feature representation and (8) is similar to (6). Specifically, the distance between $\mathbf{B}^\top\mathbf{k}_{\mathrm{i}}^{\mathrm{XX}}$ and $\mathbf{B}^\top\mathbf{k}_{\mathrm{j}}^{\mathrm{XX}}$ should be closer if $\mathrm{w}_{\mathrm{ij}} > 0$ but further if $\mathrm{w}_{\mathrm{ij}} < 0$ since the goal is to minimize (8). From (7), we observe that JMMD will *push two data points from the same classes in the same domain further*, and draw those from the same classes in different domains closer. HSIC will *draw two data points from the same classes in the same domain closer* (intra-class compactness). These observations illustrate that JMMD degrades feature discrimination as the similarity weights which strengthen intra-class compactness in the graph of HSIC, take opposite signs in the graph of JMMD. Besides, the designed $\mathbf{K}^4$ makes (4) a minimization problem.

3.5 The Proposed JMMD-HSIC

From the above analysis, we consider JMMD and HSIC by jointly minimizing (1)/(2) and (4) to propose JMMD-HSIC. With a little abuse of notations, we denote $\mathbf{K}^{\mathrm{J}}$ and $\mathbf{K}^{\mathrm{H}}$ for label kernels of JMMD and HSIC, then JMMD-HSIC in an RKHS and a projected RKHS are finalized as follows,

$$\mathrm{tr}(\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{J}}\odot\mathbf{M}^{\mathrm{J}}+\delta\mathbf{K}^{\mathrm{H}}\odot\mathbf{M}^{\mathrm{H}})),\quad \mathrm{tr}(\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{J}}\odot\mathbf{M}^{\mathrm{J}}+\delta\mathbf{K}^{\mathrm{H}}\odot\mathbf{M}^{\mathrm{H}})\mathbf{K}^{\mathrm{XX}}\mathbf{B}), \tag{9}$$

where δ aims to balance the relative importance between JMMD and HSIC, and we can adopt $\mathbf{K}^1$, $\mathbf{K}^2$ or $\mathbf{K}^3$ for $\mathbf{K}^{\mathrm{J}}$ and $\mathbf{K}^4$ for $\mathbf{K}^{\mathrm{H}}$. To validate our revealed theoretical results and the effectiveness of JMMD-HSIC, we incorporate it into a general subspace-learning framework, *i.e.*, $\arg\min\ \mathcal{L}^\star(\mathbf{B})+\mathcal{L}(\mathbf{B})$, where $\mathcal{L}^\star(\mathbf{B})$ is the JMMD-HSIC loss in a projected RKHS, and its partial derivative with respect to $\mathbf{B}$ is $2\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{J}}\odot\mathbf{M}^{\mathrm{J}}+\delta\mathbf{K}^{\mathrm{H}}\odot\mathbf{M}^{\mathrm{H}})\mathbf{K}^{\mathrm{XX}}\mathbf{B}$. $\mathcal{L}(\mathbf{B})$ is a general form for other losses.

4 Experiments

4.1 Datasets and Experimental Settings

To validate our revealed theoretical results and the effectiveness of JMMD-HSIC, we conduct extensive experiments on four benchmark datasets in cross-domain object recognition. D^1: **Office10-Caltech10** [47] consists of four domains, *i.e.*, Amazon, Dslr, Webcam, Caltech; D^2: **ImageCLEF-DA** includes three domains, *i.e.*, Caltech-256, ImageNet ILSVRC 2012, Pascal VOC 2012; D^3: **Office-31** [48] contains three domains, *i.e.*, Amazon, Dslr, Webcam; D^4: **Office-Home** [49] has four domains, *i.e.*, Art, Clipart, Product, Real-world.

As this paper mainly focuses on the problem that JMMD is hard to be applied into a subspace-learning framework, we incorporate the proposed JMMD-HSIC into three subspace-learning-based DA approaches, *i.e.*, joint distribution adaptation (JDA) [20], selective pseudo-labeling (SPL) [50], and optimal graph learning-based label propagation (OGL^{2}P) [51]. We abbreviate these three variants as JDA+JMMD-HSIC, SPL+JMMD-HSIC, and OGL^{2}P+JMMD-HSIC, respectively. For a fair comparison, all hyper-parameters remain consistent with the three approaches. Regarding δ, we uniformly set it to 0.5 for JDA+JMMD-HSIC, while assigning different values on the corresponding datasets for the other two variants after trials. On Office10-Caltech10, we use the SURF features with 800 dimensions [47] and the DECAF-6 features with 4096 dimensions [52]. On the other three datasets, we utilize the ResNet-50 features with 2048 dimensions [53]. Moreover, we adopt $\mathbf{K}^2$ for JMMD due to its superiority and $\mathbf{K}^4$ for HSIC.

4.2 Results

We compare our proposed approach with existing state-of-the-art shallow (SPL [50], PGCD [54], RMMD [55]) and deep DA approaches (BSP+MetaReg [56], DRDA [57], RSDA-MSTN [58], Jeffreys-DD [59], OGL^{2}P [51]) on D^3 and D^4. As can be seen from Tabs. 1 and 2, our proposed approach is better than the baseline methods SPL and OGL^{2}P on average, and has achieved 1.4%/0.8% and 0.9/0.8% improvements on the two datasets, respectively. Besides, OGL^{2}P+JMMD-HSIC could

Table 1: Comparison results on Office-31 with ResNet-50 features. A, D, W in the second row denotes domains of Amazon, Dslr, and Webcam, respectively.

Source	Venue	Amazon		Dslr		Webcam		Avg.
Target		D	W	A	W	A	D	
PGCD [54]	TIP’23	95.2	94.0	76.4	99.0	76.5	100.0	90.2
RMMD-I [55]	TNNLS’23	90.4	88.4	74.1	98.7	74.8	99.8	87.7
BSP+MetaReg [56]	TKDE’23	96.2	95.2	76.8	99.2	74.6	100.0	90.3
DRDA [57]	TIP’23	94.5	95.8	75.6	98.8	76.6	100.0	90.2
RSDA-MSTN [58]	TPAMI’24	96.1	95.9	77.8	99.3	78.2	100.0	91.2
Jeffreys-DD [59]	NeurIPS’24	95.9	94.9	76.0	99.1	74.6	100.0	90.1
SPL [50]	AAAI’20	93.0	92.7	76.4	98.7	76.8	99.8	89.6
SPL+JMMD-HSIC	-	95.8	95.5	78.5	99.1	77.0	100.0	91.0
OGL^{2}P [51]	TIP’25	96.2	95.5	77.5	98.7	76.8	99.4	90.7
OGL^{2}P+JMMD-HSIC	-	96.5	95.8	78.8	99.3	78.8	100.0	**91.5**

Table 2: Comparison results on Office-Home with ResNet-50 features. A, C, P, R in the second row denotes domains of Artistic, Clipart, Product, and Real-World, respectively.

Source	Venue	Artistic			Clipart			Product			Real-World			Avg.
Target		C	P	R	A	P	R	A	C	R	A	C	P	
PGCD [54]	TIP’23	57.7	77.2	79.1	59.1	74.3	72.7	61.2	54.2	79.3	70.0	58.4	82.7	68.8
RMMD-I [55]	TNNLS’23	58.4	77.8	79.3	61.6	72.8	73.0	62.7	55.3	78.9	70.4	60.1	83.2	69.5
BSP+MetaReg [56]	TKDE’23	58.0	75.5	78.9	65.0	74.7	75.0	67.9	57.2	81.8	74.7	63.5	83.8	71.3
DRDA [57]	TIP’23	58.2	74.2	81.2	65.6	75.1	73.3	65.8	57.1	80.4	75.6	63.2	85.1	71.2
RSDA-MSTN [58]	TPAMI’24	59.6	79.2	81.1	68.7	77.7	77.7	67.8	61.0	82.2	75.3	60.8	85.9	73.1
Jeffreys-DD [59]	NeurIPS’24	55.5	74.9	79.5	64.3	73.8	73.9	63.9	54.7	81.3	75.2	61.6	84.2	70.2
SPL [50]	AAAI’20	54.5	77.8	81.9	65.1	78.0	81.1	66.0	53.1	82.8	69.9	55.3	86.0	71.0
SPL+JMMD-HSIC	-	56.8	77.1	81.6	66.5	79.4	81.2	67.9	55.0	83.4	70.9	57.1	85.8	71.9
OGL^{2}P [51]	TIP’25	57.8	78.8	82.1	68.4	81.6	80.4	68.9	56.6	82.9	71.7	59.1	85.0	72.8
OGL^{2}P+JMMD-HSIC	-	58.3	79.6	82.5	69.3	81.9	80.9	69.5	57.9	83.3	73.4	61.2	85.3	**73.6**

achieve the best average results among all compared approaches, which has achieved 0.3% and 0.5% improvements compared with the second-best methods, *i.e.*, RSDA-MSTN. The comparison results on the other datasets could be found in ***Section F of the supplementary material***. Generally speaking, these results can show the effectiveness and competitiveness of our proposed JMMD-HSIC.

4.3 Feature Visualization

To further show the results of JMMD-HSIC, we visualize the feature distributions using the t-SNE algorithm as a common practice in this field [31, 22]. Fig. 2 shows the related results for JMMD, HSIC, and JMMD-HSIC in the SPL framework. The better the matching of points with the same color but different shapes, the smaller the distribution difference; disregarding shape, the tighter the clustering of points with the same color, the better the discriminability. As illustrated in Fig. 2(a), the original features perform badly on both the distribution alignment and discrimination. From Fig. 2(b), JMMD tries to align the feature distributions of the source domain and target domain, but damages the discrimination greatly. As depicted in Fig. 2(c), HSIC aims to enhance the discriminative structure in both the source and target domains. However, it exhibits poor distribution alignment, as highlighted by the dashed boxes where points of the same color but different shapes are distributed in different regions. As shown in Fig. 2(d), the feature distributions are matched better compared to HSIC (as highlighted by the dashed boxes where points of the same color but different shapes are distributed in the same region) and data points from the same classes tend to be closer (better discrimination) compared to JMMD. These observations could validate our revealed theoretical results and the effectiveness of JMMD-HSIC compared to JMMD and HSIC.

Table 3: Ablation study with different losses on Office-31 (average accuracy on 6 tasks) and Office-Home (average accuracy on 12 tasks) with ResNet-50 features.

Dataset	SPL	SPL			OGL^{2}P	OGL^{2}P		
		JMMD	HSIC	JMMD-HSIC		JMMD	HSIC	JMMD-HSIC
Office-31	89.6	87.3	89.9	**91.0**	90.7	88.5	90.6	**91.5**
Office-Home	71.0	68.7	69.0	**71.9**	72.8	71.3	71.5	**73.6**

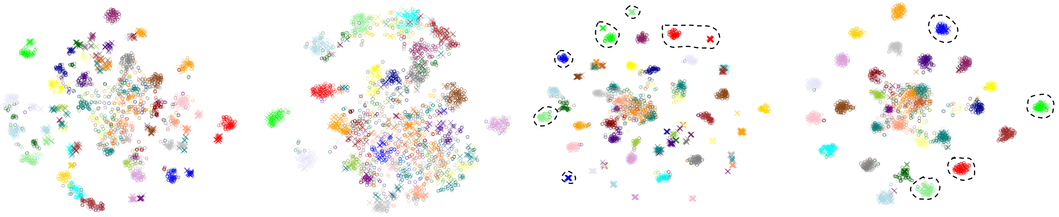
Table 4: Ablation study with different label kernels on Office-31 (average accuracy on 6 tasks) and Office-Home (average accuracy on 12 tasks) with ResNet-50 features.

Dataset	SPL	SPL+JMMD-HSIC			OGL^{2}P	OGL^{2}P+JMMD-HSIC		
		$\mathbf{K}^1$	$\mathbf{K}^2$	$\mathbf{K}^3$		$\mathbf{K}^1$	$\mathbf{K}^2$	$\mathbf{K}^3$
Office-31	89.6	85.6	**91.0**	90.9	90.7	86.7	**91.5**	**91.5**
Office-Home	71.0	64.9	**71.9**	71.8	72.8	69.8	**73.6**	**73.6**

4.4 Ablation Study

We further validate our revealed theoretical results by inspecting the feature distribution distance (JMMD metric) and feature-label independence. The smaller the JMMD metric, the better the distribution alignment; the smaller the feature-label independence, the better the discriminability. We run the methods of PCA, JDA+JMMD, JDA+HSIC and JDA+JMMD-HSIC on the dataset of D^1 with SURF features and utilize four different classifiers (1-nearest neighbor (1-NN), SVM, label propagation (LP) [60] and nearest class prototype (NCP) [50]) and two different label forms (hard and soft). We report average results of the two metrics on all DA tasks. Then, we compute these two metrics of each method on their embedded feature representations. Note that, in order to compute the true distance or metric, we have to use the ground-truth labels instead of the pseudo ones. However, the ground-truth target labels are only used for verification, not for training procedure [20]. As shown in Fig. 3, we could obtain the following observations. With JMMD, the JMMD metric is the smallest among the four methods but the feature-label independence is the largest among them, which indicates good distribution alignment but compromises discriminative structure. Conversely, with HSIC, the feature-label independence is the smallest among the four methods but the JMMD metric is the largest among them, which suggests consideration of discriminative structure but overlooking distribution alignment. In contrast, the proposed JMMD-HSIC strikes a good balance between distribution alignment and discriminability, which could lead to a better DA capacity. Moreover, PCA performs poorly in both aspects. In Tab. 3, we conduct ablation experiments using different loss functions on baselines of SPL and OGL^{2}P and have the following observations. Considering that the classifiers employed in SPL and OGL^{2}P frameworks require higher feature discriminability, we can observe that HSIC outperforms JMMD. Since the Office-31 dataset has smaller distribution discrepancies, the performance improvement of HSIC over JMMD is more significant on the Office-31 dataset (2.6%, 2.1%) compared to the Office-Home dataset (0.3%, 0.2%). Our proposed loss function achieves optimal results. Although both SPL and OGL^{2}P also incorporate loss functions for feature distribution alignment and discriminative feature learning, our approach provides more precise balancing between these two objectives by fundamentally analyzing how JMMD compromises feature discriminability. In other words, we more effectively mitigate the impact of distribution alignment on feature discriminability. Consequently, our method achieves superior performance compared to SPL and OGL^{2}P. These observations could validate our revealed theoretical results and the effectiveness of proposed JMMD-HSIC compared to JMMD and HSIC.

In Tab. 4, we conduct ablation experiments to validate why we adopt the label kernel $\mathbf{K}^2$. It can be observed that the marginal distribution kernel $\mathbf{K}^1$, which neglects label information, leads to inaccurate distribution alignment and thus performs the worst. The results of the weighted class-conditional distribution kernel $\mathbf{K}^3$ and the conditional distribution kernel $\mathbf{K}^2$ are nearly identical because the class imbalance issue in the experimental dataset used in this paper is not severe—whereas the weighted class-conditional distribution kernel is specifically designed to address class imbalance.



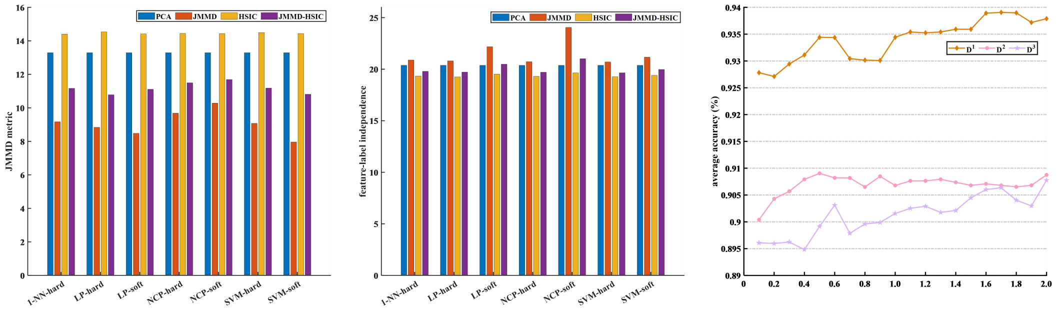
(a) original features

(b) SPL+JMMD

(c) SPL+HSIC

(d) SPL+JMMD-HSIC

Figure 2: Feature visualization of the DA task Amazon (source domain) → Webcam (target domain) from Office-31 dataset for SPL+JMMD, SPL+HSIC, and SPL+JMMD-HSIC. Different colors represent various classes, and '○' and '×' represent source domain and target domain, respectively.



(a) Different classifiers and labels

(b) Different classifiers and labels

(c) δ

Figure 3: Quantitative analysis for the JMMD (distribution distance, (a)) and HSIC (feature-label independence, (b)) metrics and sensitivity analysis for δ (c).

4.5 Sensitivity of Hyper-Parameter

We conduct the sensitivity analysis of δ in SPL+JMMD-HSIC to validate that the optimal results could be achieved under a stable range. We report average accuracy results of SPL+JMMD-HSIC on D^1, D^2, and D^3 with deep features. We plot average classification accuracy *w.r.t.*, different values of δ in Fig. 3(c), and choose $\delta \in [0.1, 2]$, which reflects the importance of HSIC for discrimination reinforcement in JMMD. It can be observed that the average accuracy often achieves its optimal value on a wide range for each dataset, which could display the stability of δ. Notably, the average results of D^4 with different values of δ fluctuate between 68.1%∼71.7%, smaller than those of $\mathrm{D}^1 \sim \mathrm{D}^3$. Therefore, we do not plot the change for D^4 to observe the fluctuation trends more distinctively. The rationale behind selecting δ within the range of [0.1, 2] is as follows: we are gradually mitigating the negative effects of JMMD on discriminability when $\delta \in [0.1, 1]$. We completely eliminate the negative effects of JMMD, and gradually promote discriminability when $\delta \in [1, 2]$. Therefore, the model performs better when δ takes larger values within this range.

5 Conclusions

JMMD is still not fully explored and is especially hard to be applied to a subspace-learning framework. To overcome this problem, we deduce a concise JMMD and obtain two essential findings, *i.e.*, the uniformity of JMMD and the reason for feature discrimination degradation in JMMD. To strengthen the discrimination of JMMD, we jointly consider JMMD and HSIC to propose a novel loss dubbed as JMMD-HSIC. Comprehensive tests carried out on some benchmark datasets could validate our revealed theoretical results, and show promising performance with our proposed JMMD-HSIC. For future work, we plan to explore: $\mathbf{1}$) designing an advanced label kernel to handle more diverse domain adaptation scenarios; $\mathbf{2}$) extending the framework to more complex visual tasks; $\mathbf{3}$) designing novel label kernels by systematically analyzing how different classifiers vary in their sensitivity to feature distribution alignment and feature discriminability.

Acknowledgments

This work was supported in part by the the National Key R&D Program of China (Grant No.2022ZD0119200), National Natural Science Foundation of China (Grant No.62306343, No.62301621, No.62025604), Guangdong Basic and Applied Basic Research Foundation (Grant No.2025A1515011322, No.2025A1515011398), China Postdoctoral Science Foundation (Grant No.2025T180435, No.2024M753741). Shenzhen Science and Technology Program (No.20231121172359002, 2023A008), Shenzhen General Research Project (No.JCYJ20241202125904007).

References

[1] Ye Wang, Junyang Chen, Mengzhu Wang, Hao Li, Wei Wang, Houcheng Su, Zhihui Lai, Wei Wang, and Zhenghan Chen. A closer look at classifier in adversarial domain generalization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 280–289, 2023.

[2] Wei Wang, Mengzhu Wang, Xiao Dong, Long Lan, Quannan Zu, Xiang Zhang, and Cong Wang. Class-specific and self-learning local manifold structure for domain adaptation. *Pattern Recognition*, 142:109654, 2023.

[3] Wei Wang, Hanyang Li, Ke Shi, Chao Huang, Yang Cao, Cong Wang, and Xiaochun Cao. Optimal graph learning and nuclear norm maximization for deep cross-domain robust label propagation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1407–1415, 2024.

[4] Wei Wang, Hanyang Li, Cong Wang, Chao Huang, Zhengming Ding, Feiping Nie, and Xiaochun Cao. Deep label propagation with nuclear norm maximization for visual domain adaptation. *IEEE Transactions on Image Processing*, 34:1246–1258, 2025.

[5] Wangzi Qi, Wei Wang, Chao Huang, Jie Wen, and Cong Wang. Batch singular value polarization and weighted semantic augmentation for universal domain adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[6] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the 15th European Conference on Computer Vision*, volume 11207, pages 297–313, 2018.

[7] Mehmet Pilanci and Elif Vural. Domain adaptation on graphs by learning aligned graph bases. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):587–600, 2022.

[8] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1180–1189, 2015.

[9] Léo Gautheron, Ievgen Redko, and Carole Lartizien. Feature selection for unsupervised domain adaptation using optimal transport. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 11052, pages 759–776, 2018.

[10] Lei Zhang, Shanshan Wang, Guang-Bin Huang, Wangmeng Zuo, Jian Yang, and David Zhang. Manifold criterion guided transfer learning via intermediate domain generation. *IEEE Transactions on Neural Networks and Learning Systems*, 30(12):3759–3773, 2019.

[11] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2010.

[12] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 22(2):199–210, 2011.

[13] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):1853–1865, 2017.

[14] Alain Rakotomamonjy, Rémi Flamary, Gilles Gasso, M. El Alaya, Maxime Berar, and Nicolas Courty. Optimal transport for conditional domain matching and label shift. *Machine Learning*, 111(5):1651–1670, 2022.

[15] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 3730–3739, 2017.

[16] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

[17] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 513–520, 2006.

[18] Yong Ren, Yucen Luo, and Jun Zhu. Improving generative moment matching networks with distribution partition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9403–9410, 2021.

[19] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5885–5892, 2019.

[20] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.

[21] Jindong Wang, Yiqiang Chen, Shuji Hao, Wenjie Feng, and Zhiqi Shen. Balanced distribution adaptation for transfer learning. In *Proceedings of the IEEE International Conference on Data Mining*, pages 1129–1134, 2017.

[22] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2208–2217, 2017.

[23] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[24] Wei Wang, Zhihui Wang, Haojie Li, Juan Zhou, and Zhengming Ding. Adaptive local neighbors for transfer discriminative feature learning. In *Proceedings of the 24th European Conference on Artificial Intelligence*, volume 325, pages 1595–1602, 2020.

[25] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: batch spectral penalization for adversarial domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1081–1090, 2019.

[26] Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, volume 3734, pages 63–77, 2005.

[27] Lixin Duan, Ivor W. Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012.

[28] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1410–1417, 2014.

[29] Muhammad Ghifary, W. Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence*, volume 8862, pages 898–904, 2014.

[30] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[31] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 97–105, 2015.

[32] Shuang Li, Chi Harold Liu, Limin Su, Binhui Xie, Zhengming Ding, C. L. Philip Chen, and Dapeng Wu. Discriminative transfer feature and label consistency for cross-domain image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4842–4856, 2020.

[33] Min Meng, Mengcheng Lan, Jun Yu, and Jigang Wu. Coupled knowledge transfer for visual data recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1776–1789, 2021.

[34] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 945–954, 2017.

[35] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S. Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the ACM Multimedia*, pages 402–410, 2018.

[36] Wanxia Deng, Qing Liao, Lingjun Zhao, Deke Guo, Gangyao Kuang, Dewen Hu, and Li Liu. Joint clustering and discriminative feature alignment for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 30:7842–7855, 2021.

[37] Wen Zhang and Dongrui Wu. Discriminative joint probability maximum mean discrepancy (DJP-MMD) for domain adaptation. In *IEEE International Joint Conference on Neural Networks*, pages 1–8, 2020.

[38] Fatemeh Dorri and Ali Ghodsi. Adapting component analysis. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 846–851, 2012.

[39] Fatemeh Dorri and Ali Ghodsi. Minimizing the discrepancy between source and target domains by learning adapting components. *Journal of Computer Science and Technology*, 29(1):105–115, 2014.

[40] Ke Yan, Lu Kou, and David Zhang. Learning domain-invariant subspace using domain features and independence maximization. *IEEE Transactions on Cybernetics*, 48(1):288–299, 2018.

[41] Shanshan Wang, Lei Zhang, Wangmeng Zuo, and Bob Zhang. Class-specific reconstruction transfer learning for visual recognition across domains. *IEEE Transactions on Image Processing*, 29:2424–2438, 2020.

[42] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[43] Bin Zhang, Qianyao Qiang, Fei Wang, and Feiping Nie. Flexible multi-view unsupervised graph embedding. *IEEE Transactions on Image Processing*, 30:4143–4156, 2021.

[44] Xinhong Ma, Tianzhu Zhang, and Changsheng Xu. Gcan: Graph convolutional adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8266–8276, 2019.

[45] Ming Jin, Heng Chang, Wenwu Zhu, and Somayeh Sojoudi. Power up! robust graph convolutional network via graph powering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8004–8012, 2021.

[46] Charles R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186(186):273–273, 1973.

[47] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

[48] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the 11th European Conference on Computer Vision*, volume 6314, pages 213–226, 2010.

[49] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5385–5394, 2017.

[50] Qian Wang and Toby P. Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6243–6250, 2020.

[51] Wei Wang, Mengzhu Wang, Chao Huang, Cong Wang, Jie Mu, Feiping Nie, and Xiaochun Cao. Optimal graph learning-based label propagation for cross-domain image classification. *IEEE Transactions on Image Processing*, 34:1529–1544, 2025.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 1106–1114, 2012.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[54] Wenxu Wang, Zhencai Shen, Daoliang Li, Ping Zhong, and Yingyi Chen. Probability-based graph embedding cross-domain and class discriminative feature learning for domain adaptation. *IEEE Transactions on Image Processing*, 32:72–87, 2023.

[55] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):264–277, 2023.

[56] Shuang Li, Wenxuan Ma, Jinming Zhang, Chi Harold Liu, Jian Liang, and Guoren Wang. Meta-reweighted regularization for unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2781–2795, 2023.

[57] Zenan Huang, Jun Wen, Siheng Chen, Linchao Zhu, and Nenggan Zheng. Discriminative radial domain adaptation. *IEEE Transactions on Image Processing*, 32:1419–1431, 2023.

[58] Xiang Gu, Jian Sun, and Zongben Xu. Unsupervised and semi-supervised robust spherical space domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1757–1774, 2024.

[59] Ziqiao Wang and Yongyi Mao. On f-divergence principled domain adaptation: An improved framework. In *Advances in Neural Information Processing Systems*, 2024.

[60] Feiping Nie, Shiming Xiang, Yun Liu, and Changshui Zhang. A general graph-based semi-supervised learning with novel class discovery. *Neural Computing and Applications*, 19(4):549–555, 2010.

[61] Xianda Zhang. *Matrix Analysis and Applications*. Tsinghua University Press, 2013.

[62] Wei Wang, Baopu Li, Mengzhu Wang, Feiping Nie, Zhihui Wang, and Haojie Li. Confidence regularized label propagation based domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3319–3333, 2022.

[63] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3940–3949, 2020.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper aims to fully explore the distribution-distance metric, *i.e.*, JMMD, in the field of DA.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

Justification: We did not discuss the limitations of this work due to space constraints.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: We provided a complete proof for the theoretical results in the supplementary file due to space constraints.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided comprehensive information to facilitate the reproducibility of our work.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provided the references for the data and code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all the training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have conducted sensitivity analysis of hyper-parameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our experiments do not require significant computational resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured that the paper conforms, in every respect, with the Code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This paper focuses on a statistical distribution-distance metric JMMD in traditional machine learning and does not involve LLM.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A A Concise JMMD in an RKHS

Theorem 2 *In an RKHS, JMMD could be rewritten as the following concise form,*

$$\begin{aligned}&\mathbb{D}_{\mathcal{H}}\Big(\mathcal{P}^{\rm s}(\mathbf{X}^{\rm s},\mathbf{Y}^{\rm s}),\mathcal{P}^{\rm t}(\mathbf{X}^{\rm t},\mathbf{Y}^{\rm t})\Big)=\Big|\Big|\tfrac{1}{\rm n^s}\textstyle\sum_{\rm i=1}^{\rm n^s}\Big(\psi(\mathbf{x}_{\rm i}^{\rm s})\otimes\phi(\mathbf{y}_{\rm i}^{\rm s})\Big)-\tfrac{1}{\rm n^t}\sum_{\rm j=1}^{\rm n^t}\Big(\psi(\mathbf{x}_{\rm j}^{\rm t})\otimes\phi(\mathbf{y}_{\rm j}^{\rm t})\Big)\Big|\Big|_{\mathcal{H}}^2\\&=\mathrm{tr}\Big(\mathbf{K}^{\rm XX}(\mathbf{K}^{\rm YY}\odot\mathbf{M}^{\rm J})\Big),\end{aligned}\tag{10}$$

where $\mathbb{D}_{\mathcal{H}}$ denotes a distance metric between two joint probability distributions, *i.e.*, $\mathcal{P}^{\rm s}$ and $\mathcal{P}^{\rm t}$. We empirically estimate JMMD with the following steps: i) we utilize ψ and ϕ to map features and labels from the source domain and target domain to the RKHS, respectively; ii) we calculate the mean of the tensor product between feature and label for each domain; iii) we compute the difference between these two means. Notably, $\mathbf{x}_{\rm i}$ and $\mathbf{x}_{\rm j}$ are the i-th and j-th column vectors of $\mathbf{X}$.

Proof:

For convenience, we define $\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})$ and $\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})$ as shown in the following equations,

$$\begin{aligned}\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})&=\Big[\psi(\mathbf{x}_1^{\rm s})\otimes\phi(\mathbf{y}_1^{\rm s}),\cdots,\psi(\mathbf{x}_{\rm n^s}^{\rm s})\otimes\phi(\mathbf{y}_{\rm n^s}^{\rm s})\Big]\in\mathbb{R}^{\infty\times\rm n^s},\\\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})&=\Big[\psi(\mathbf{x}_1^{\rm t})\otimes\phi(\mathbf{y}_1^{\rm t}),\cdots,\phi(\mathbf{x}_{\rm n^t}^{\rm t})\otimes\phi(\mathbf{y}_{\rm n^t}^{\rm t})\Big]\in\mathbb{R}^{\infty\times\rm n^t}.\end{aligned}\tag{11}$$

Then, (10) could be rewritten as below,

$$\begin{aligned}&\mathbb{D}_{\mathcal{H}}\Big(\mathcal{P}^{\rm s}(\mathbf{X}^{\rm s},\mathbf{Y}^{\rm s}),\mathcal{P}^{\rm t}(\mathbf{X}^{\rm t},\mathbf{Y}^{\rm t})\Big)=\Big|\Big|\tfrac{1}{\rm n^s}\textstyle\sum_{\rm i=1}^{\rm n^s}\Big(\psi(\mathbf{x}_{\rm i}^{\rm s})\otimes\phi(\mathbf{y}_{\rm i}^{\rm s})\Big)-\tfrac{1}{\rm n^t}\sum_{\rm j=1}^{\rm n^t}\Big[\psi(\mathbf{x}_{\rm j}^{\rm t})\otimes\phi(\mathbf{y}_{\rm j}^{\rm t})\Big)\Big|\Big|_{\mathcal{H}}^2\\&=\mathrm{tr}\left(\begin{bmatrix}\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})^\top\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s}) & \Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})^\top\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})\\ \Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})^\top\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s}) & \Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})^\top\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})\end{bmatrix}\begin{bmatrix}\frac{\mathbf{1}_{\rm n^s\times 1}\mathbf{1}_{\rm n^s\times 1}^\top}{\rm n^sn^s} & \frac{-\mathbf{1}_{\rm n^s\times 1}\mathbf{1}_{\rm n^t\times 1}^\top}{\rm n^sn^t}\\ \frac{-\mathbf{1}_{\rm n^t\times 1}\mathbf{1}_{\rm n^s\times 1}^\top}{\rm n^tn^s} & \frac{\mathbf{1}_{\rm n^t\times 1}\mathbf{1}_{\rm n^t\times 1}^\top}{\rm n^tn^t}\end{bmatrix}\right).\end{aligned}\tag{12}$$

where $\mathbf{1}_{\rm n^s\times 1}$ and $\mathbf{1}_{\rm n^t\times 1}$ are two column vectors whose elements are all ones with sizes of $\rm n^s$ and $\rm n^t$.

Moreover, we have the following equations,

$$\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})^\top\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})=\mathbf{K}^{\rm X^sX^s}\odot\mathbf{K}^{\rm Y^sY^s},\tag{13}$$

$$\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})^\top\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})=\mathbf{K}^{\rm X^tX^t}\odot\mathbf{K}^{\rm Y^tY^t},\tag{14}$$

$$\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})^\top\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})=\mathbf{K}^{\rm X^sX^t}\odot\mathbf{K}^{\rm Y^sY^t},\tag{15}$$

$$\Gamma^{\rm t}(\mathbf{x}^{\rm t},\mathbf{y}^{\rm t})^\top\Gamma^{\rm s}(\mathbf{x}^{\rm s},\mathbf{y}^{\rm s})=\mathbf{K}^{\rm X^tX^s}\odot\mathbf{K}^{\rm Y^tY^s}.\tag{16}$$

where $\mathbf{K}^{\rm X^sX^s},\cdots,\mathbf{K}^{\rm Y^sY^s},\cdots\in\mathbb{R}^{\rm n^s\times n^s},\cdots\mathbb{R}^{\rm n^s\times n^s},\cdots$ are the kernel matrices and they are computed by $\mathrm{k}_{\rm ij}^{\rm X^sX^s}=k^{\rm X}(\mathbf{x}_{\rm i}^{\rm s},\mathbf{x}_{\rm j}^{\rm s\top}),\cdots,\mathrm{k}_{\rm ij}^{\rm Y^sY^s}=k^{\rm Y}(\mathbf{y}_{\rm i}^{\rm s},\mathbf{y}_{\rm j}^{\rm s\top}),\cdots$. Here $k^{\rm X}$ and $k^{\rm Y}$ are feature and label kernels.

Therefore, we can rewrite (12) using the feature kernel matrix $\mathbf{K}^{\rm XX}$ and the label kernel matrix $\mathbf{K}^{\rm YY}$ as below,

$$
\begin{aligned}
&\mathbb{D}_{\mathcal{H}}\Big(\mathcal{P}^{\mathrm{s}}(\mathbf{X}^{\mathrm{s}}, \mathbf{Y}^{\mathrm{s}}), \mathcal{P}^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}, \mathbf{Y}^{\mathrm{t}})\Big) \\
&= \mathrm{tr}\left(\begin{bmatrix} \mathbf{K}^{\mathrm{X^sX^s}} & \mathbf{K}^{\mathrm{X^sX^t}} \\ \mathbf{K}^{\mathrm{X^tX^s}} & \mathbf{K}^{\mathrm{X^tX^t}} \end{bmatrix} \odot \begin{bmatrix} \mathbf{K}^{\mathrm{Y^sY^s}} & \mathbf{K}^{\mathrm{Y^sY^t}} \\ \mathbf{K}^{\mathrm{Y^tY^s}} & \mathbf{K}^{\mathrm{Y^tY^t}} \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{\mathrm{n^s}\times 1}\mathbf{1}_{\mathrm{n^s}\times 1}^{\top}}{\mathrm{n^s n^s}} & \frac{-\mathbf{1}_{\mathrm{n^s}\times 1}\mathbf{1}_{\mathrm{n^t}\times 1}^{\top}}{\mathrm{n^s n^t}} \\ \frac{-\mathbf{1}_{\mathrm{n^t}\times 1}\mathbf{1}_{\mathrm{n^s}\times 1}^{\top}}{\mathrm{n^t n^s}} & \frac{\mathbf{1}_{\mathrm{n^t}\times 1}\mathbf{1}_{\mathrm{n^t}\times 1}^{\top}}{\mathrm{n^t n^t}} \end{bmatrix} \right) \\
&= \mathrm{tr}\big(\mathbf{K}^{\mathrm{XX}} \odot \mathbf{K}^{\mathrm{YY}}\mathbf{M}^{\mathrm{J}}\big),
\end{aligned} \tag{17}
$$

where $\mathbf{K}^{\mathrm{XX}} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ and $\mathbf{K}^{\mathrm{YY}} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ are feature and label kernel matrices for all source and target domains, and $\mathrm{n} = \mathrm{n^s} + \mathrm{n^t}$.

According to $\mathrm{tr}\big(\mathbf{A} \odot \mathbf{BC}\big) = \mathrm{tr}\Big(\mathbf{A}(\mathbf{B} \odot \mathbf{C})\Big)$ where $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are symmetric matrices [61], thus $\mathrm{tr}\big(\mathbf{K}^{\mathrm{XX}} \odot \mathbf{K}^{\mathrm{YY}}\mathbf{M}^{\mathrm{J}}\big) = \mathrm{tr}\Big(\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J}})\Big)$. $\mathbf{M}^{\mathrm{J}}$ is calculated as below,

$$
\mathrm{m}_{\mathrm{ij}}^{\mathrm{J}} = \begin{cases} 1/(\mathrm{n^s n^s}), & \mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s}} \\ 1/(\mathrm{n^t n^t}), & \mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t}} \\ -1/(\mathrm{n^s n^t}), & \text{otherwise.} \end{cases} \tag{18}
$$

where $\mathcal{D}^{\mathrm{s}}$ and $\mathcal{D}^{\mathrm{t}}$ denote source and target domains.

□

B A Concise JMMD in a Projected RKHS

Theorem 3 *In a projected RKHS, JMMD could be rewritten as the following concise form,*

$$
\begin{aligned}
&\mathbb{D}_{\mathcal{H}}\Big(\mathcal{P}^{\mathrm{s}}(\mathbf{X}^{\mathrm{s}}, \mathbf{Y}^{\mathrm{s}}), \mathcal{P}^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}, \mathbf{Y}^{\mathrm{t}})\Big) \\
&= \Big|\Big| \tfrac{1}{\mathrm{n^s}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n^s}} \Big(\mathbf{T}^{\top}\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}}) \otimes \phi(\mathbf{y}_{\mathrm{i}}^{\mathrm{s}})\Big) - \tfrac{1}{\mathrm{n^t}} \textstyle\sum_{\mathrm{j}=1}^{\mathrm{n^t}} \Big(\mathbf{T}^{\top}\psi(\mathbf{x}_{\mathrm{j}}^{\mathrm{t}}) \otimes \phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{t}})\Big) \Big|\Big|_{\mathcal{H}}^{2} \\
&= \Big|\Big| \tfrac{1}{\mathrm{n^s}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n^s}} \Big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \big(\mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^{\top}\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}})\big) \otimes \phi(\mathbf{y}_{\mathrm{i}}^{\mathrm{s}})\Big) - \tfrac{1}{\mathrm{n^t}} \textstyle\sum_{\mathrm{j}=1}^{\mathrm{n^t}} \Big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \big(\mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^{\top}\psi(\mathbf{x}_{\mathrm{j}}^{\mathrm{t}})\big) \otimes \phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{t}})\Big) \Big|\Big|_{\mathcal{H}}^{2} \\
&= \mathrm{tr}\Big(\mathbf{B}^{\top}\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J}})\mathbf{K}^{\mathrm{XX}}\mathbf{B}\Big).
\end{aligned} \tag{19}
$$

where $\mathbf{T} \in \mathbb{R}^{\infty\times\mathrm{d}}$ is the feature projection matrix and 'd' is the dimension in the embedded subspace. Different from (10), we project $\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}})$ and $\psi(\mathbf{x}_{\mathrm{j}}^{\mathrm{t}})$ into an embedded subspace, and then empirically estimate JMMD.

Proof:

We begin by introducing the Representer theorem [23] as below,

Theorem 4 (Representer theorem) *It says that any function can be decomposed into finite values of a kernel function with corresponding coefficients [23].*

$$
\begin{aligned}
&\mathbf{T}^{\top}\psi(\mathbf{x}) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \Big(\mathbf{b}_{\mathrm{i}} k^{\mathrm{X}}(\mathbf{x}, \mathbf{x}_{\mathrm{i}})\Big) = \\
&\textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \Big(\mathbf{b}_{\mathrm{i}} \langle \psi(\mathbf{x}), \psi(\mathbf{x}_{\mathrm{i}}) \rangle\Big) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \Big(\mathbf{b}_{\mathrm{i}} \psi(\mathbf{x}_{\mathrm{i}})^{\top}\psi(\mathbf{x})\Big),
\end{aligned} \tag{20}
$$

where $\mathbf{b}_{\mathrm{i}} \in \mathbb{R}^{\mathrm{d}\times 1}$ and we definite a new projection matrix $\mathbf{B} = [\mathbf{b}_1^{\top}; \cdots; \mathbf{b}_{\mathrm{n}}^{\top}] \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$.

For convenience, we define $\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})$ and $\Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})$ as shown in the following equations according to the Representer theorem,

$$
\begin{aligned}
&\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s}) = \\
&\Big[\big(\textstyle\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{1}^{s})\otimes\phi(\mathbf{y}_{1}^{s}),\cdots,\big(\textstyle\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{n^{s}}^{s})\otimes\phi(\mathbf{y}_{n^{s}}^{s})\Big]\in\mathbb{R}^{\infty\times n^{s}}, \\
&\Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t}) = \\
&\Big[\big(\textstyle\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{1}^{t})\otimes\phi(\mathbf{y}_{1}^{t}),\cdots,\big(\textstyle\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{n^{t}}^{t})\otimes\phi(\mathbf{y}_{n^{t}}^{t})\Big]\in\mathbb{R}^{\infty\times n^{t}}.
\end{aligned}
\tag{21}
$$

where $\mathbf{b}_{i}\in\mathbb{R}^{d\times 1}$ and we definite a new projection matrix $\mathbf{B}=[\mathbf{b}_{1}^{\top};\cdots;\mathbf{b}_{n}^{\top}]\in\mathbb{R}^{n\times d}$ $(n=n^{s}+n^{t})$. Then, (19) could be rewritten as below,

$$
\begin{aligned}
&\mathbb{D}_{\mathcal{H}} \\
&=\Big|\Big|\tfrac{1}{n^{s}}\textstyle\sum_{i=1}^{n^{s}}\Big(\big(\textstyle\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{i}^{s})\otimes\phi(\mathbf{y}_{i}^{s})\Big)-\tfrac{1}{n^{t}}\textstyle\sum_{j=1}^{n^{t}}\Big(\big(\textstyle\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{j}^{t})\otimes\phi(\mathbf{y}_{j}^{t})\Big)\Big|\Big|_{\mathcal{H}}^{2} \\
&=\mathrm{tr}\left(\begin{bmatrix}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s}) & \Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})\end{bmatrix}\begin{bmatrix}\frac{\mathbf{1}_{n^{s}\times1}\mathbf{1}_{n^{s}\times1}^{\top}}{n^{s}n^{s}} & \frac{-\mathbf{1}_{n^{s}\times1}\mathbf{1}_{n^{t}\times1}^{\top}}{n^{s}n^{t}}\\ \frac{-\mathbf{1}_{n^{t}\times1}\mathbf{1}_{n^{s}\times1}^{\top}}{n^{t}n^{s}} & \frac{\mathbf{1}_{n^{t}\times1}\mathbf{1}_{n^{t}\times1}^{\top}}{n^{t}n^{t}}\end{bmatrix}\begin{bmatrix}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\\ \Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})^{\top}\end{bmatrix}\right) \\
&=\mathrm{tr}\left(\begin{bmatrix}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\\ \Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})^{\top}\end{bmatrix}\begin{bmatrix}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s}) & \Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})\end{bmatrix}\begin{bmatrix}\frac{\mathbf{1}_{n^{s}\times1}\mathbf{1}_{n^{s}\times1}^{\top}}{n^{s}n^{s}} & \frac{-\mathbf{1}_{n^{s}\times1}\mathbf{1}_{n^{t}\times1}^{\top}}{n^{s}n^{t}}\\ \frac{-\mathbf{1}_{n^{t}\times1}\mathbf{1}_{n^{s}\times1}^{\top}}{n^{t}n^{s}} & \frac{\mathbf{1}_{n^{t}\times1}\mathbf{1}_{n^{t}\times1}^{\top}}{n^{t}n^{t}}\end{bmatrix}\right) \\
&=\mathrm{tr}\left(\begin{bmatrix}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s}) & \Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})\\ \Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})^{\top}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s}) & \Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})^{\top}\Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})\end{bmatrix}\begin{bmatrix}\frac{\mathbf{1}_{n^{s}\times1}\mathbf{1}_{n^{s}\times1}^{\top}}{n^{s}n^{s}} & \frac{-\mathbf{1}_{n^{s}\times1}\mathbf{1}_{n^{t}\times1}^{\top}}{n^{s}n^{t}}\\ \frac{-\mathbf{1}_{n^{t}\times1}\mathbf{1}_{n^{s}\times1}^{\top}}{n^{t}n^{s}} & \frac{\mathbf{1}_{n^{t}\times1}\mathbf{1}_{n^{t}\times1}^{\top}}{n^{t}n^{t}}\end{bmatrix}\right).
\end{aligned}
\tag{22}
$$

Similar to the proof of Theorem 2, we rewrite $\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})$, $\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})$, $\Theta^{t}(\mathbf{x}^{t},\mathbf{y}^{t})^{\top}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})$, $\cdots$ using feature and label kernels. First,

$$
\begin{aligned}
&\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s})^{\top}\Theta^{s}(\mathbf{x}^{s},\mathbf{y}^{s}) \\
&=\begin{bmatrix}
\Big\langle\big(\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{1}^{s})\otimes\phi(\mathbf{y}_{1}^{s}),\big(\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{1}^{s})\otimes\phi(\mathbf{y}_{1}^{s})\Big\rangle & \cdots & \cdots\\
\Big\langle\big(\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{2}^{s})\otimes\phi(\mathbf{y}_{2}^{s}),\big(\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{1}^{s})\otimes\phi(\mathbf{y}_{1}^{s})\Big\rangle & \cdots & \cdots\\
\cdots & \cdots & \cdots\\
\Big\langle\big(\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{n^{s}}^{s})\otimes\phi(\mathbf{y}_{n^{s}}^{s}),\big(\sum_{l=1}^{n}\mathbf{b}_{l}\psi(\mathbf{x}_{l})^{\top}\big)\psi(\mathbf{x}_{1}^{s})\otimes\phi(\mathbf{y}_{1}^{s})\Big\rangle & \cdots & \cdots
\end{bmatrix},
\end{aligned}
\tag{23}
$$

where $\langle\bullet,\bullet\rangle$ denotes the inner product between two vectors. Moreover, we have the following equation,

$$
\begin{aligned}
&\Big\langle \big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^\top\big)\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}}) \otimes \phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{s}}), \big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^\top\big)\psi(\mathbf{x}_{\mathrm{k}}) \otimes \phi(\mathbf{y}_{\mathrm{m}})\Big\rangle \\
&= \Big(\big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^\top\big)\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}}) \otimes \phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{s}})\Big)^\top \Big(\big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^\top\big)\psi(\mathbf{x}_{\mathrm{k}}) \otimes \phi(\mathbf{y}_{\mathrm{m}})\Big) \\
&= \Big(\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}})^\top\big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \psi(\mathbf{x}_{\mathrm{l}})\mathbf{b}_{\mathrm{l}}^\top\big) \otimes \phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{s}})^\top\Big)\Big(\big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^\top\big)\psi(\mathbf{x}_{\mathrm{k}}) \otimes \phi(\mathbf{y}_{\mathrm{m}})\Big) \\
&= \Big(\psi(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}})^\top\big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \psi(\mathbf{x}_{\mathrm{l}})\mathbf{b}_{\mathrm{l}}^\top\big)\big(\textstyle\sum_{\mathrm{l}=1}^{\mathrm{n}} \mathbf{b}_{\mathrm{l}}\psi(\mathbf{x}_{\mathrm{l}})^\top\big)\psi(\mathbf{x}_{\mathrm{k}})\Big) \otimes \Big(\phi(\mathbf{y}_{\mathrm{j}}^{\mathrm{s}})^\top\phi(\mathbf{y}_{\mathrm{m}})\Big) \\
&= \Big[k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}},\mathbf{x}_1), k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}},\mathbf{x}_2), \cdots, k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}},\mathbf{x}_{\mathrm{n}})\Big]\mathbf{B}\mathbf{B}^\top\Big[k^{\mathrm{X}}(\mathbf{x}_1,\mathbf{x}_{\mathrm{k}}), k^{\mathrm{X}}(\mathbf{x}_2,\mathbf{x}_{\mathrm{k}}), \cdots, k^{\mathrm{X}}(\mathbf{x}_{\mathrm{n}},\mathbf{x}_{\mathrm{k}})\Big]^\top \otimes k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{j}}^{\mathrm{s}},\mathbf{y}_{\mathrm{m}}) \\
&= \Big[k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}},\mathbf{x}_1), k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}},\mathbf{x}_2), \cdots, k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}}^{\mathrm{s}},\mathbf{x}_{\mathrm{n}})\Big]\mathbf{B}\mathbf{B}^\top\Big[k^{\mathrm{X}}(\mathbf{x}_1,\mathbf{x}_{\mathrm{k}}), k^{\mathrm{X}}(\mathbf{x}_2,\mathbf{x}_{\mathrm{k}}), \cdots, k^{\mathrm{X}}(\mathbf{x}_{\mathrm{n}},\mathbf{x}_{\mathrm{k}})\Big]^\top k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{j}}^{\mathrm{s}},\mathbf{y}_{\mathrm{m}}) \\
&= \mathbf{K}_{(\mathrm{i},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{k})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{j}},\mathbf{y}_{\mathrm{m}}),
\end{aligned}
\tag{24}
$$

where the subscripts $(\mathrm{i},\bullet)$ and $(\bullet,\mathrm{k})$ denote the i-th row vector and the k-th column vector of a given matrix, respectively. Then, we can obtain the following equation,

$$
\begin{aligned}
&\Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}})^\top\Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}}) = \\
&\begin{bmatrix}
\mathbf{K}_{(1,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_1,\mathbf{y}_1) & \cdots & \mathbf{K}_{(1,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_1,\mathbf{y}_{\mathrm{n^s}}) \\
\mathbf{K}_{(2,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_2,\mathbf{y}_1) & \cdots & \mathbf{K}_{(2,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_2,\mathbf{y}_{\mathrm{n^s}}) \\
\cdots & \cdots & \cdots \\
\mathbf{K}_{(\mathrm{n^s},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}},\mathbf{y}_1) & \cdots & \mathbf{K}_{(\mathrm{n^s},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}},\mathbf{y}_{\mathrm{n^s}})
\end{bmatrix}.
\end{aligned}
\tag{25}
$$

Similarly, $\Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}})^\top\Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}}) =$

$$
\begin{bmatrix}
\mathbf{K}_{(\mathrm{n^s}+1,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s}+1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+1},\mathbf{y}_{\mathrm{n^s}+1}) & \cdots & \mathbf{K}_{(\mathrm{n^s}+1,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+1},\mathbf{y}_{\mathrm{n}}) \\
\mathbf{K}_{(\mathrm{n^s}+2,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s}+1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+2},\mathbf{y}_{\mathrm{n^s}+1}) & \cdots & \mathbf{K}_{(\mathrm{n^s}+2,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+2},\mathbf{y}_{\mathrm{n}}) \\
\cdots & \cdots & \cdots \\
\mathbf{K}_{(\mathrm{n},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s}+1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n}},\mathbf{y}_{\mathrm{n^s}+1}) & \cdots & \mathbf{K}_{(\mathrm{n},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n}},\mathbf{y}_{\mathrm{n}})
\end{bmatrix}.
\tag{26}
$$

$\Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}})^\top\Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}}) =$

$$
\begin{bmatrix}
\mathbf{K}_{(1,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s}+1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_1,\mathbf{y}_{\mathrm{n^s}+1}) & \cdots & \mathbf{K}_{(1,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_1,\mathbf{y}_{\mathrm{n}}) \\
\mathbf{K}_{(2,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s}+1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_2,\mathbf{y}_{\mathrm{n^s}+1}) & \cdots & \mathbf{K}_{(2,\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_2,\mathbf{y}_{\mathrm{n}}) \\
\cdots & \cdots & \cdots \\
\mathbf{K}_{(\mathrm{n^s},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n^s}+1)}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}},\mathbf{y}_{\mathrm{n^s}+1}) & \cdots & \mathbf{K}_{(\mathrm{n^s},\bullet)}^{\mathrm{XX}}\mathbf{B}\mathbf{B}^\top\mathbf{K}_{(\bullet,\mathrm{n})}^{\mathrm{XX}}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}},\mathbf{y}_{\mathrm{n}})
\end{bmatrix}.
\tag{27}
$$

$\Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}})^\top\Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}}) =$

$$\begin{bmatrix} \mathbf{K}^{\mathrm{XX}}_{(\mathrm{n^s}+1,\bullet)}\mathbf{B}\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}_{(\bullet,1)}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+1},\mathbf{y}_1) & \cdots & \mathbf{K}^{\mathrm{XX}}_{(\mathrm{n^s}+1,\bullet)}\mathbf{B}\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}_{(\bullet,\mathrm{n^s})}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+1},\mathbf{y}_{\mathrm{n^s}}) \\ \mathbf{K}^{\mathrm{XX}}_{(\mathrm{n^s}+2,\bullet)}\mathbf{B}\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}_{(\bullet,1)}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+2},\mathbf{y}_1) & \cdots & \mathbf{K}^{\mathrm{XX}}_{(\mathrm{n^s}+2,\bullet)}\mathbf{B}\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}_{(\bullet,\mathrm{n^s})}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n^s}+2},\mathbf{y}_{\mathrm{n^s}}) \\ \cdots & \cdots & \cdots \\ \mathbf{K}^{\mathrm{XX}}_{(\mathrm{n},\bullet)}\mathbf{B}\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}_{(\bullet,1)}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n}},\mathbf{y}_1) & \cdots & \mathbf{K}^{\mathrm{XX}}_{(\mathrm{n},\bullet)}\mathbf{B}\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}_{(\bullet,\mathrm{n^s})}k^{\mathrm{Y}}(\mathbf{y}_{\mathrm{n}},\mathbf{y}_{\mathrm{n^s}}) \end{bmatrix}. \tag{28}$$

According to (25) $\sim$ (28), we could obtain the following equation,

$$\begin{aligned} \mathbb{D}_{\mathcal{H}} &= \mathrm{tr}\left(\begin{bmatrix} \Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}})^\top\Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}}) & \Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}})^\top\Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}}) \\ \Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}})^\top\Theta^{\mathrm{s}}(\mathbf{x}^{\mathrm{s}},\mathbf{y}^{\mathrm{s}}) & \Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}})^\top\Theta^{\mathrm{t}}(\mathbf{x}^{\mathrm{t}},\mathbf{y}^{\mathrm{t}}) \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{\mathrm{n^s}\times 1}\mathbf{1}^\top_{\mathrm{n^s}\times 1}}{\mathrm{n^s n^s}} & \frac{-\mathbf{1}_{\mathrm{n^s}\times 1}\mathbf{1}^\top_{\mathrm{n^t}\times 1}}{\mathrm{n^s n^t}} \\ \frac{-\mathbf{1}_{\mathrm{n^t}\times 1}\mathbf{1}^\top_{\mathrm{n^s}\times 1}}{\mathrm{n^t n^s}} & \frac{\mathbf{1}_{\mathrm{n^t}\times 1}\mathbf{1}^\top_{\mathrm{n^t}\times 1}}{\mathrm{n^t n^t}} \end{bmatrix} \right) \\ &= \mathrm{tr}\Big(\mathbf{B}^\top\mathbf{K}^{XX}(\mathbf{K}^{\mathrm{YY}}\odot\mathbf{M}^{\mathrm{J}})\mathbf{K}^{\mathrm{XX}}\mathbf{B}\Big). \end{aligned} \tag{29}$$

□

C Probability Distribution Distances

C.1 Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) [12] establishes the mean embedding of the marginal probability distribution in a RKHS endowed by the kernel k^{X} (feature mapping ψ), and using finite samples to empirically estimate the distance between $\mu_{\mathrm{X^s}}$ (mean embedding of source domain) and $\mu_{\mathrm{X^t}}$ (mean embedding of target domain) with the Hilbert-Schmidt norm as the following equation,

$$\begin{aligned} \mathbb{D}_{\mathcal{H}}\Big(\mathcal{P}^{\mathrm{s}}(\mathbf{X}^{\mathrm{s}}),\mathcal{P}^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}})\Big) &= \Big|\Big|\mathbb{E}\Big(\psi(\mathbf{X}^{\mathrm{s}})\Big) - \mathbb{E}\Big(\psi(\mathbf{X}^{\mathrm{t}})\Big)\Big|\Big|^2_{\mathcal{H}} = \Big|\Big|\mu_{\mathrm{X^s}} - \mu_{\mathrm{X^t}}\Big|\Big|^2_{\mathcal{H}} \\ &= ||\tfrac{1}{\mathrm{n^s}}\textstyle\sum_{\mathrm{i}=1}^{\mathrm{n^s}}\psi(\mathbf{x}_{\mathrm{i}}) - \tfrac{1}{\mathrm{n^t}}\sum_{\mathrm{j}=1}^{\mathrm{n^t}}\psi(\mathbf{x}_{\mathrm{j}})\Big|\Big|^2_{\mathcal{H}} = \mathrm{tr}(\mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{M}}), \end{aligned} \tag{30}$$

where $\mathbf{K}^{\mathrm{XX}} = \psi(\mathbf{X})^\top\psi(\mathbf{X}) \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ and $k^{\mathrm{XX}}_{\mathrm{ij}} = k^{\mathrm{X}}(\mathbf{x}_{\mathrm{i}},\mathbf{x}_{\mathrm{j}})$. Besides, the MMD matrix $\mathbf{M}^{\mathrm{M}}$ can be computed as below,

$$\mathrm{m}^{\mathrm{M}}_{\mathrm{ij}} = \begin{cases} 1/(\mathrm{n^s n^s}), & \mathbf{x}_{\mathrm{i}},\mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{s}} \\ 1/(\mathrm{n^t n^t}), & \mathbf{x}_{\mathrm{i}},\mathbf{x}_{\mathrm{j}} \in \mathcal{D}^{\mathrm{t}} \\ -1/(\mathrm{n^s n^t}), & \text{otherwise.} \end{cases} \tag{31}$$

Moreover, the MMD in a projected RKHS is tr($\mathbf{B}^\top\mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{m}}\mathbf{K}^{\mathrm{XX}}\mathbf{B}$).

C.2 Class-Wise Maximum Mean Discrepancy

The class-wise maximum mean discrepancy (CMMD) [20] constructs the sum of MMD for each specific class as the following equation,

$$
\begin{aligned}
&\mathbb{D}_{\mathcal{H}}\Big(\mathcal{P}^{s}(\mathbf{X}^{s}|\mathbf{Y}^{s}), \mathcal{P}^{t}(\mathbf{X}^{t}|\mathbf{Y}^{t})\Big) \\
&= \textstyle\sum_{c=1}^{C} \Big|\Big|\mathbb{E}\Big(\psi(\mathbf{X}^{s,c})\Big) - \mathbb{E}\Big(\psi(\mathbf{X}^{t,c})\Big)\Big|\Big|_{\mathcal{H}}^{2} = \textstyle\sum_{c=1}^{C} \Big|\Big|\mu_{\mathbf{X}^{s,c}} - \mu_{\mathbf{X}^{t,c}}\Big|\Big|_{\mathcal{H}}^{2} \\
&= \textstyle\sum_{c=1}^{C} \Big|\Big|\frac{1}{n^{s,c}} \sum_{i=1}^{n^{s,c}} \psi(\mathbf{x}_{i}) - \frac{1}{n^{t,c}} \sum_{j=1}^{n^{t,c}} \psi(\mathbf{x}_{j})\Big|\Big|_{\mathcal{H}}^{2} = \textstyle\sum_{c=1}^{C} \mathrm{tr}(\mathbf{K}^{XX}\mathbf{M}^{C,c}),
\end{aligned}
\tag{32}
$$

where the MMD matrix $\mathbf{M}^{C,c}$ can be computed as below,

$$
m_{ij}^{C,c} = \begin{cases}
1/(n^{s,c}n^{s,c}), & \mathbf{x}_{i} \in \mathcal{D}^{s,c}, \mathbf{x}_{j} \in \mathcal{D}^{s,c} \\
1/(n^{t,c}n^{t,c}), & \mathbf{x}_{i} \in \mathcal{D}^{t,c}, \mathbf{x}_{j} \in \mathcal{D}^{t,c} \\
-1/(n^{s,c}n^{t,c}), & \mathbf{x}_{i} \in \mathcal{D}^{s,c}, \mathbf{x}_{j} \in \mathcal{D}^{t,c} \\
-1/(n^{t,c}n^{s,c}), & \mathbf{x}_{i} \in \mathcal{D}^{t,c}, \mathbf{x}_{j} \in \mathcal{D}^{s,c} \\
0, & \text{otherwise.}
\end{cases}
\tag{33}
$$

Similarly, the CMMD in a projected RKHS is $\sum_{c=1}^{C} \mathrm{tr}(\mathbf{B}^{\top}\mathbf{K}^{XX}\mathbf{M}^{C,c}\mathbf{K}^{XX}\mathbf{B})$.

C.3 Weighted Class-Wise Maximum Mean Discrepancy

To deal with class imbalanced dataset, the weighted class-wise maximum mean discrepancy (WCMMD) introduces the class prior probability $\mathcal{P}(\mathbf{Y})$ into the CMMD [21], which pays more attention on the large-size categories and is formulated as the following equation,

$$
\begin{aligned}
&\textstyle\sum_{c=1}^{C} \Big|\Big|\frac{\mathcal{P}^{s}(\mathbf{y}^{s}=c)}{n^{s,c}} \sum_{i=1}^{n^{s,c}} \psi(\mathbf{x}_{i}) - \frac{\mathcal{P}^{t}(\mathbf{y}^{t}=c)}{n^{t,c}} \sum_{j=1}^{n^{t,c}} \psi(\mathbf{x}_{j})\Big|\Big|_{\mathcal{H}}^{2} \\
&= \textstyle\sum_{c=1}^{C} \Big|\Big|\frac{1}{n^{s}} \sum_{i=1}^{n^{s,c}} \psi(\mathbf{x}_{i}) - \frac{1}{n^{t}} \sum_{j=1}^{n^{t,c}} \psi(\mathbf{x}_{j})\Big|\Big|_{\mathcal{H}}^{2} = \textstyle\sum_{c=1}^{C} \mathrm{tr}(\mathbf{K}^{XX}\mathbf{M}^{WC,c}),
\end{aligned}
\tag{34}
$$

where $\mathbf{M}^{WC,c}$ can be computed with the following equation,

$$
m_{ij}^{WC,c} = \begin{cases}
1/(n^{s}n^{s}), & \mathbf{x}_{i} \in \mathcal{D}^{s,c}, \mathbf{x}_{j} \in \mathcal{D}^{s,c} \\
1/(n^{t}n^{t}), & \mathbf{x}_{i} \in \mathcal{D}^{t,c}, \mathbf{x}_{j} \in \mathcal{D}^{t,c} \\
-1/(n^{s}n^{t}), & \mathbf{x}_{i} \in \mathcal{D}^{s,c}, \mathbf{x}_{j} \in \mathcal{D}^{t,c} \\
-1/(n^{t}n^{s}), & \mathbf{x}_{i} \in \mathcal{D}^{t,c}, \mathbf{x}_{j} \in \mathcal{D}^{s,c} \\
0, & \text{otherwise.}
\end{cases}
\tag{35}
$$

Similarly, the WCMMD in a projected RKHS is $\sum_{c=1}^{C} \mathrm{tr}(\mathbf{B}^{\top}\mathbf{K}^{XX}\mathbf{M}^{WC,c}\mathbf{K}^{XX}\mathbf{B})$.

D The Uniformity of JMMD

Theorem 5 *The marginal, class conditional and weighted class conditional probability distribution distances are three special cases of JMMD with label reproducing kernels $\boldsymbol{K}^{1}$, $\boldsymbol{K}^{2}$ and $\boldsymbol{K}^{3}$. $\boldsymbol{K}^{1} = \boldsymbol{1}_{n\times n}$ is a matrix whose elements are all 1 with the size of $n \times n$, and $\boldsymbol{K}^{2}$, $\boldsymbol{K}^{3}$ are defined as below,*

$$\mathrm{k}^2_{\mathrm{ij}} = \begin{cases} (\mathrm{n^s n^s})/(\mathrm{n^{s,c} n^{s,c}}), & \mathbf{x}_\mathrm{i} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{s,c}} \\ (\mathrm{n^t n^t})/(\mathrm{n^{t,c} n^{t,c}}), & \mathbf{x}_\mathrm{i} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{t,c}} \\ (\mathrm{n^s n^t})/(\mathrm{n^{s,c} n^{t,c}}), & \mathbf{x}_\mathrm{i} \in \mathcal{D}^{\mathrm{s,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{t,c}} \\ (\mathrm{n^t n^s})/(\mathrm{n^{t,c} n^{s,c}}), & \mathbf{x}_\mathrm{i} \in \mathcal{D}^{\mathrm{t,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{s,c}} \\ 0, & \text{otherwise,} \end{cases} \tag{36}$$

$$\mathrm{k}^3_{\mathrm{ij}} = \begin{cases} 1, & \mathbf{x}_\mathrm{i} \in D^{\mathrm{s,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{s,c}} \\ 1, & \mathbf{x}_\mathrm{i} \in D^{\mathrm{t,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{t,c}} \\ 1, & \mathbf{x}_\mathrm{i} \in D^{\mathrm{s,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{t,c}} \\ 1, & \mathbf{x}_\mathrm{i} \in D^{\mathrm{t,c}}, \mathbf{x}_\mathrm{j} \in \mathcal{D}^{\mathrm{s,c}} \\ 0, & \text{otherwise,} \end{cases} \tag{37}$$

where the superscript 's/t,c' denotes data points from the c-th class in the source/target domain.

Proof:

As proved before, the formulations of concise JMMD are $\mathrm{tr}(\mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J}}))$ (in a RKHS) and $\mathrm{tr}(\mathbf{B}^\top \mathbf{K}^{\mathrm{XX}}(\mathbf{K}^{\mathrm{YY}} \odot \mathbf{M}^{\mathrm{J}})\mathbf{K}^{\mathrm{XX}}\mathbf{B})$ (in a projected RKHS). Moreover, the formulations of marginal probability distribution distance are $\mathrm{tr}(\mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{M}})$ (in a RKHS) and $\mathrm{tr}(\mathbf{B}^\top \mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{M}}\mathbf{K}^{\mathrm{XX}}\mathbf{B})$ (in a projected RKHS). The formulations of class conditional probability distribution distance are $\mathrm{tr}(\mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{C,c}})$ (in a RKHS) and $\mathrm{tr}(\mathbf{B}^\top \mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{C,c}}\mathbf{K}^{\mathrm{XX}}\mathbf{B})$ (in a projected RKHS). The formulations of weighted class conditional probability distribution distance are $\mathrm{tr}(\mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{WC,c}})$ (in a RKHS) and $\mathrm{tr}(\mathbf{B}^\top \mathbf{K}^{\mathrm{XX}}\mathbf{M}^{\mathrm{WC,c}}\mathbf{K}^{\mathrm{XX}}\mathbf{B})$ (in a projected RKHS). It is easy to verify that $\mathbf{K}^1 \odot \mathbf{M}^{\mathrm{J}} = \mathbf{M}^{\mathrm{M}}$, $\mathbf{K}^2 \odot \mathbf{M}^{\mathrm{J}} = \mathbf{M}^{\mathrm{C,c}}$ and $\mathbf{K}^3 \odot \mathbf{M}^{\mathrm{J}} = \mathbf{M}^{\mathrm{WC,c}}$. Therefore, the marginal, class conditional and weighted class conditional probability distribution distances are three special cases of JMMD with different label reproducing kernels $\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$. We will prove $\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$ are the reproducing kernels in next Subsection.

□

E Reproducing Kernels

Theorem 6 *$\boldsymbol{K}^1$, $\boldsymbol{K}^2$, $\boldsymbol{K}^3$ and $\boldsymbol{K}^4$ are the reproducing kernels, where $\boldsymbol{K}^1$, $\boldsymbol{K}^2$, $\boldsymbol{K}^3$ are defined in Theorem 5.*

Proof:

$\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$: According to the Mercer's theorem [23], we only have to prove that the Gram matrices $\mathbf{G}^1$, $\mathbf{G}^2$ and $\mathbf{G}^3$ corresponding to $\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$ are semi-positive definite matrices. In fact, the Gram matrices $\mathbf{G}^2$ and $\mathbf{G}^3$ can be decomposed into the following equation,

$$\mathbf{G}^2 = \textstyle\sum_{\mathrm{c}=1}^{\mathrm{C}} \mathbf{G}^{2,\mathrm{c}}, \quad \mathbf{G}^3 = \textstyle\sum_{\mathrm{c}=1}^{\mathrm{C}} \mathbf{G}^{3,\mathrm{c}}. \tag{38}$$

It is obvious that the sum of several semi-positive definite matrices is also a semi-positive definite matrix, thus we only have to prove that the Gram matrices $\mathbf{G}^1$, $\mathbf{G}^{2,\mathrm{c}}$ and $\mathbf{G}^{3,\mathrm{c}}$ are semi-positive definite. $\mathbf{G}^1$, $\mathbf{G}^{2,\mathrm{c}}$ and $\mathbf{G}^{3,\mathrm{c}}$ can be decomposed into the following equations,

$$\mathbf{G}^1 = \mathbf{p}^1\mathbf{p}^{1\top}, \, \mathbf{G}^{2,\mathrm{c}} = \mathbf{p}^{2,\mathrm{c}}\mathbf{p}^{2,\mathrm{c}\top}, \, \mathbf{G}^{3,\mathrm{c}} = \mathbf{p}^{3,\mathrm{c}}\mathbf{p}^{3,\mathrm{c}\top}, \tag{39}$$

where $\mathbf{p}^1 = \mathbf{1}_\mathrm{n}$ is a column vector whose elements are all 1, and $\mathbf{p}^{2,\mathrm{c}} \in \mathbb{R}^\mathrm{n}$, $\mathbf{p}^{3,\mathrm{c}} \in \mathbb{R}^\mathrm{n}$ could be defined as below,

$$p_i^{2,c} = \begin{cases} n^s/n^{s,c}, & \mathbf{x}_i \in \mathcal{D}^{s,c} \\ n^t/n^{t,c}, & \mathbf{x}_i \in \mathcal{D}^{t,c} \\ 0, & \text{otherwise}, \end{cases} \tag{40}$$

$$p_i^{3,c} = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{D}^{s,c} \\ 1, & \mathbf{x}_i \in \mathcal{D}^{t,c} \\ 0, & \text{otherwise}, \end{cases} \tag{41}$$

where $p_i^{2/3,c}$ is the value of the i-th component of $\mathbf{l}^{2/3,c}$. For $\forall\, \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq 0$, we have,

$$\begin{aligned} \mathbf{x}^\top \mathbf{G}^1 \mathbf{x} &= \mathbf{x}^\top \mathbf{p}^1 \mathbf{p}^{1\top} \mathbf{x} = \mathbf{x}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{x} \\ &= (x_1 + x_2 + ... + x_n)^2 \geq 0, \end{aligned} \tag{42}$$

$$\begin{aligned} &\mathbf{x}^\top \mathbf{G}^{2,c} \mathbf{x} = \mathbf{x}^\top \mathbf{p}^{2,c} \mathbf{p}^{2,c\top} \mathbf{x} \\ &= (x_1 p_1^{2,c} + x_2 p_2^{2,c} + ... + x_n p_n^{2,c})^2 \geq 0. \end{aligned} \tag{43}$$

and,

$$\begin{aligned} &\mathbf{x}^\top \mathbf{G}^{3,c} \mathbf{x} = \mathbf{x}^\top \mathbf{p}^{3,c} \mathbf{p}^{3,c\top} \mathbf{x} \\ &= (x_1 p_1^{3,c} + x_2 p_2^{3,c} + \cdots + x_n p_n^{3,c})^2 \geq 0. \end{aligned} \tag{44}$$

Therefore, $\mathbf{G}^1$, $\mathbf{G}^2$ and $\mathbf{G}^3$ are semi-positive definite matrices. Then, $\mathbf{K}^1$, $\mathbf{K}^2$ and $\mathbf{K}^3$ are the reproducing kernels.

$\mathbf{K}^4$: For $\forall\, \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq 0$, due to $w_{ij} \geq 0$, we have,

$$\mathbf{x}^\top \mathbf{G}^4 \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - x_j)^2 \geq 0, \tag{45}$$

where $\mathbf{G}^4$ is the Gram matrix of $\mathbf{K}^4$ and $\mathbf{W}$ is defined as below,

$$w_{ij} = \begin{cases} (n^s n^s)/(n^{s,c} n^{s,c}), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ (n^t n^t)/(n^{t,c} n^{t,c}), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ 0, & \text{otherwise}. \end{cases} \tag{46}$$

Therefore, $\mathbf{G}^4$ is a semi-positive matrix and $\mathbf{K}^4$ is the reproducing kernel.

□

F Experiments

We run the JDA+JMMD/HSIC/Our(JMMD-HSIC) and adopt the classifiers of 1-nearest neighbor (1-NN), support vector machines (SVM[*]), label propagation (LP[†]) [60] and nearest class prototype

[*] https://www.csie.ntu.edu.tw/ cjlin/libsvm

[†] https://www.escience.cn/people/fpnie/index.html

Table 5: Ablation study using different classifiers/labels on the Office10-Caltech10 dataset with SURF features.

Classifier	1-NN	SVM		LP		NCP	
Original Features	40.9	47.7		48.3		45.7	
Label	Hard	Hard	Soft	Hard	Soft	Hard	Soft
JDA+JMMD	47.7	50.2	49.2	54.2	52.8	47.8	41.4
JDA+HSIC	47.9	49.0	48.5	54.0	52.8	48.8	46.8
JDA+Our	**49.6**	**50.9**	**49.9**	**55.4**	**54.3**	**49.6**	**47.1**

Table 6: Comparison average results of our proposed SPL+JMMD-HSIC with state-of-the-art DA methods on Office10-Caltech10 dataset with DECAF-6 features. A, C, D, W in the second row denotes domains of Amazon, Caltech, Dslr, and Webcam, respectively.

Source	Venue	Amazon			Caltech			Dslr			Webcam			Avg.
Target		C	D	W	A	D	W	A	C	W	A	C	D	
PGCD [54]	TIP'23	86.5	90.4	84.1	92.5	92.4	91.2	92.5	87.6	100.0	91.6	85.3	100.0	91.2
RMMD-II [55]	TNNLS'23	88.4	91.7	92.9	93.4	96.8	95.9	93.6	88.9	100.0	92.2	88.9	100.0	93.6
SPL [50]	AAAI'20	87.4	89.2	95.3	92.7	98.7	93.2	92.9	88.6	98.6	92.0	87.0	100.0	93.0
SPL+Our	-	90.0	96.8	93.9	93.7	99.4	93.9	93.8	90.3	100.0	93.3	89.4	99.4	94.5
OGL^{2}P [51]	TIP'25	89.7	97.5	91.9	94.3	98.7	95.9	94.2	90.2	99.3	94.6	89.5	100.0	94.6
OGL^{2}P+Our	-	90.2	97.8	93.2	95.3	99.2	95.7	94.6	90.5	100.0	94.2	89.3	100.0	**95.0**

Table 7: Comparison of average results of our proposed SPL+JMMD-HSIC with state-of-the-art DA methods on ImageCLEF-DA dataset with ResNet-50 features. C, I, P in the second row denotes domains of Caltech-256, ImageNet ILSVRC, and Pascal VOC, respectively.

Source	Venue	Caltech -256		ImageNet ILSVRC		Pascal VOC		Avg.
Target		I	P	C	P	C	I	
RMMD-I [55]	TNNLS'23	93.2	78.3	95.7	79.5	95.5	92.0	89.0
RSDA-MSTN [58]	TPAMI'24	93.3	79.3	97.8	80.5	96.8	94.2	90.3
SPL [50]	AAAI'20	95.7	80.5	96.7	78.3	96.3	94.5	90.3
SPL+Our	-	96.3	81.4	96.7	80.5	96.7	95.0	91.1
OGL^{2}P [51]	TIP'25	95.8	81.2	96.8	82.2	97.2	95.7	91.5
OGL^{2}P+Our	-	96.5	81.8	97.4	83.5	97.8	96.0	**92.2**

(NCP[‡]) [50] on the Office-Caltech10 dataset with SURF features (average classification results on 12 DA tasks). As can be seen from Tab. 5, JMMD and HSIC perform better than the original features as JMMD matches the distributions of the source domain and target domain, and HSIC enhances domain-specific discriminative structures. The proposed JMMD-HSIC could achieve the best results no matter what classifiers or labels are, which shows the effectiveness of JMMD-HSIC and indicates that it is necessary to jointly consider JMMD and HSIC for a better DA capacity. Here, the symbol 'Soft' denotes the probability soft label and the symbol 'Hard' is the hard (one-hot) label. Notably, 1-NN could not produce a soft label thus only 'Hard' is reported. It can be seen that the performance of the 'Soft' label is even worse than that of the 'Hard' label, and it may be because the performance heavily depends on the quality of predicted soft labels of the target domain [62, 63].

We compare our proposed approach with existing state-of-the-art shallow (SPL [50], PGCD [54], RMMD [55]) and deep DA approaches (RSDA-MSTN [58], OGL^{2}P [51]) on D^1 and D^2. As can be seen from Tabs. 6 and 7, our proposed approach is better than the baseline methods SPL and OGL^{2}P on average, and has achieved 1.5%/0.4% and 0.8/0.7% improvements on the two datasets, respectively. Besides, OGL^{2}P+JMMD-HSIC could achieve the best average results among all compared approaches, which has achieved 0.4% and 0.7% improvements compared with the second-best methods, *i.e.*, OGL^{2}P. Generally speaking, these results can show the effectiveness and competitiveness of our proposed JMMD-HSIC.

[‡] https://github.com/hellowangqian/domainadaptation-capls