

COD: Learning Conditional Invariant Representation for Domain Adaptation Regression

Hao-Ran Yang¹, Chuan-Xian Ren^{1*}, and You-Wei Luo^{1,2}

¹ School of Mathematics, Sun Yat-Sen University, China

² School of Mathematics, Jiaying University, China

{yanghr26@mail2, rchuanx@mail, luoyw28@mail2}.sysu.edu.cn

Abstract. Aiming to generalize the label knowledge from a source domain with continuous outputs to an unlabeled target domain, Domain Adaptation Regression (DAR) is developed for complex practical learning problems. However, due to the continuity problem in regression, existing conditional distribution alignment theory and methods with discrete prior, which are proven to be effective in classification settings, are no longer applicable. In this work, focusing on the feasibility problems in DAR, we establish the sufficiency theory for the regression model, which shows the generalization error can be sufficiently dominated by the cross-domain conditional discrepancy. Further, to characterize conditional discrepancy with continuous conditioning variable, a novel Conditional Operator Discrepancy (COD) is proposed, which admits the metric property on conditional distributions via the kernel embedding theory. Finally, to minimize the discrepancy, a COD-based conditional invariant representation learning model is proposed, and the reformulation is derived to show that reasonable modifications on moment statistics can further improve the discriminability of the adaptation model. Extensive experiments on standard DAR datasets verify the validity of theoretical results and the superiority over SOTA DAR methods.

Keywords: Domain Adaptation Regression, Optimal Transport, Invariant Representation Learning, Kernel Embedding Theory.

1 Introduction

Deep learning has shown a promising ability for fitting and achieved great success in various application tasks. However, large-scale data with sufficient annotations plays a vital role in training high-quality deep models while the label annotations for unlabeled data are usually expensive. Therefore, aiming to reduce training costs by transferring knowledge, Unsupervised Domain Adaptation (UDA) [2, 14, 17, 28] has received increasing attention in recent years. Basically, UDA aims to transfer task-related information from a labeled source domain

* Corresponding author.

to a target domain without annotations, where the domain gap is characterized as the distribution shift across domains [43]. Following this line, considerable methods have been proposed to alleviate the UDA problem by shift correction and distribution alignment, such as adversarial training for cross-domain samples [14, 52] and explicitly distribution matching by statistical discrepancy minimization [28, 49]. From a theoretical perspective, recent literature [6, 60] show that the marginal distribution P_X with covariate shift assumption is insufficient to characterize the cross-domain generalization error, while the conditional distribution alignment indeed serves as a sufficient condition to ensure small error and successful transfer. Mathematically, when the conditional distributions $P_{X|y}$ for every y are aligned, a well-trained source predictor can perform well on the target domain, which is described in Fig. 1. Inspired by theoretical advances, recent works [23, 41, 45] consider the statistical methods for conditional alignment, and show promising performance in mitigating negative transfer and improving recognition accuracy. However, most of the existing results are designed for classification settings with discrete prior on label variables, which induces the infeasibility for regression problems. Consequently, it restricts the application of UDA methods for wider real-world problems, *e.g.*, human pose estimation [36], product condition monitoring [27], depth estimation [42].

Considering the dilemma of UDA, pioneer works have explored the Domain Adaptation Regression (DAR) framework for solving the distribution shift in continuous label scenarios, where theoretical results [7, 35] and correction methods [15, 47, 51] are provided. To further relax the requirement of labeled training target data in pioneer DAR works, advanced works [4, 38] propose unsupervised DAR methods to learn transferable regression models without the help of labeled target data. Nonetheless, existing DAR theory and methods are directly facilitated by the covariate shift framework, where the marginal distribution alignment is considered. Therefore, it inevitably induces the negative transfer and misaligned local structure, which are shown in the results in classification settings [6, 60]. In fact, there is a significant obstacle for modeling conditional distribution alignment in regression scenarios, *i.e.*, the conditional shift correction induces infinite matching problems for $P_{X|Y=y}$ on all points y . The continuous label variable with infinite possible outcomes makes it infeasible to directly apply the existing conditional alignment methods for DAR. Thus, it is vital to understand the continuity problem in regression, and develop the generalization error theory and statistical methodology for DAR scenarios.

In this work, we focus on the theory guarantee and conditional alignment methodology for successful DAR. Theoretically, we explore the generalization upper bound for continuous output scenarios and show the cross-domain joint error can be sufficiently upper-bounded by the conditional discrepancy. It ensures the sufficiency of conditional alignment still holds for regression scenarios and provides reliable guarantees for theory-driven modeling. Methodologically, to deal with the continuous problem in conditional alignment, we treat the conditional distribution as finite statistical moments in Hilbert space. By exploring the distribution embedding theory of Reproducing Kernel Hilbert Space (RKHS), a novel

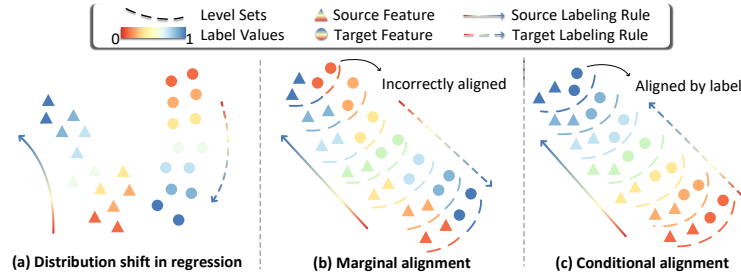


Fig. 1: Illustration of the conditional shift in regression setting, where label value $y \in [0, 1]$ and color gradients imply the continuity. **(a)** Before adaptation, distribution shift exists between source and target representations so the predictor trained on source domain cannot generalize to target domain. **(b)** After marginal distribution alignment, distributions of source and target domain are globally aligned. However, representations with different labels may be falsely aligned across domains, leading to significant inconsistency between cross-domain labeling rules. Thus, level sets provided by the source predictor are not suitable for target representations. **(c)** After conditional distribution alignment, label-wise matching is achieved, where the level sets of cross-domain labeling rules are consistent and the source predictor will be reliable.

conditional metric called Conditional Operator Discrepancy (COD) is proposed, which can be interpreted as the conditional Wasserstein distance in RKHS. With the sufficiency theory and metric-based learning principle for regression, a COD-based conditional invariant representation learning method is proposed, which aims to correct the conditional shift on the pushforward measures in latent representation space. By reformulating the empirical COD estimation, a modified principle is developed to further improve the discriminability during knowledge transfer. Generally, our contributions can be summarized as follows:

- The sufficiency theory for the conditional shift framework is established for DAR settings, which shows the cross-domain conditional discrepancy on continuous label variable serves as the major term in generalization error.
- A novel COD metric for characterizing conditional distribution discrepancy with continuous and infinite conditioning variable is explored, providing a feasible solution for conditional discrepancy optimization in DAR.
- A COD-based representation learning model with reformulation on moment statistics is proposed, which admits discriminability on task-related knowledge and shows superior performance in extensive experiments.

2 Related Work

Domain Adaptation Classification. Based on the distribution shift assumption, most of the UDA methods can be roughly categorized into three types: methods that align marginal distribution, conditional distribution and joint distribution. Most UDA methods consider only the marginal distribution align-

ment. Discrepancy-based methods attempt to directly minimize the distribution discrepancy and different types of discrepancy measurement greatly enrich this type of methodology. For example, these methods adopt moment matching [28, 49] or extension under manifold framework [17, 33]. In particular, Optimal transport (OT) and Wasserstein distance are introduced to UDA as a sound measurement for domain shift, based on which a large number of methods have been proposed [8–10, 24, 46]. Zhang *et al.* [59] further develop both theory and methodology to RKHS and Luo *et al.* [31] extends Kernel Bures metric to conditional distribution alignment. Adversarial-based methods [14, 20, 21] learn domain-invariant features via two-player games. Specifically, they alternatively optimize the feature generator and the domain discriminator to achieve domain confusion. In order to obtain better performance, advanced methods have explored how to utilize the label information to achieve conditional distribution alignment [3, 32, 41, 44, 45, 50, 53, 61] or joint distribution alignment [29, 30]. However, UDA methods for shift correction are proposed for discrete and finite conditioning variables, which are inapplicable for regression.

Domain Adaptation Regression. DA for regression has received relatively little attention compared with classification. Mansour *et al.* [35] and Cortes *et al.* [7] introduce theoretical analysis for DA in regression scenarios. Zhao *et al.* [62] discuss the tradeoffs between accuracy and invariance in learning invariant representations in both classification and regression settings. A range of methods in the shallow regime are proposed to tackle the DAR problem. These methods can be roughly categorized into two types: importance weighting [15, 37] and learning domain-invariant representation [8, 39, 40]. Besides, some attempts have been made in deep regime to tackle regression problem [25, 47, 51, 56]. Unfortunately, most of them belong to semi-supervised DA, which means they need to use a small number of target domain labels. Recently, deep UDA methods for regression that get rid of target label are proposed [4, 11, 38, 57]. RSD [4] states that deep neural networks are less robust to feature scaling in regression problems and aligning the deep features directly will impede task performance for altering feature scale. Thus, a strategy of aligning orthogonal bases of both domains is adopted in RSD. Based on RSD, DARE-GRAM [38] aligns the inverse Gram Matrix of deep features of both domains, and ensures promising improvements in accuracy and stability. Meanwhile, several DAR methods for specific tasks have been explored [1, 18, 22, 26, 27, 55]. However, the theory and method for conditional shift correction are still unexplored for DAR scenarios, which hinders the in-depth reasoning and development for practical modeling.

3 COD: Theory and Methodology

3.1 Preliminary

Setup. In this paper, we discuss UDA method dealing with regression problems. Let X and Y be covariate and label variable taken values on \mathcal{X} and \mathcal{Y} . \mathbb{P} denotes the set of distributions and subscripts represent the corresponding random variable while the superscript represents the domain, *e.g.* $P_{XY}^s \in \mathbb{P}_{\mathcal{X}\mathcal{Y}}$

implies the joint distribution on source domain. In UDA settings, we are provided with labeled source samples $(x^s, y^s) \sim P_{XY}^s$ and unlabeled target samples $(x^t, y^t) \sim P_{XY}^t$, which means the target labels are not available during training. Samples of source and target domain follow different distributions, *i.e.* $P_{XY}^s \neq P_{XY}^t$. Denote $g : X \rightarrow Z$ as the representation extractor, $h : Z \rightarrow Y$ as the predictor, ε_s and ε_t as the source and target prediction error, respectively. UDA methods aim to generalize the learner $h \circ g$ across domains and minimize target risk ε_t by mitigating the distribution shift problem. In contrast to the discrete label space in classification settings, unsupervised DAR discusses UDA problems with continuous labels, which means the number of classes is uncountable, leading to great difficulty in aligning conditional distribution $P_{X|Y}$. *The proofs are provided in supplementary materials.*

OT in RKHS. Let $(\mathcal{X}, \mathcal{B})$ be a measure space with Borel σ -field \mathcal{B} . Denote $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ as the RKHS of \mathcal{X} , which is generated by the positive definite kernel $k_{\mathcal{X}}$. Elements from \mathcal{X} are mapped to $\mathcal{H}_{\mathcal{X}}$ via feature map $\phi(x) = k_{\mathcal{X}}(x, \cdot)$, which is assumed to satisfy the reproducing properties $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}_{\mathcal{X}}} = k_{\mathcal{X}}(x, x')$ and $\langle \phi(x), f \rangle_{\mathcal{H}_{\mathcal{X}}} = f(x), \forall f \in \mathcal{H}_{\mathcal{X}}$. The mean element μ_X in $\mathcal{H}_{\mathcal{X}}$ with law P_X is given by $\mu_X = \mathbb{E}_X[\phi(X)]$, and the covariance operator \mathcal{C}_{XX} is defined as $\mathcal{C}_{XX} = E_X(\phi(X) - \mu_X) \otimes (\phi(X) - \mu_X)$.

For any two distributions $P_X^s, P_X^t \in \mathbb{P}_{\mathcal{X}}$, the Kantorovitch optimal transport in RKHS $\mathcal{H}_{\mathcal{K}}$ is considered on pushforward measures $\phi_{\#}P_X^s$ and $\phi_{\#}P_X^t$:

$$d_{\text{KW}}^2(X^s, X^t) = \inf_{\pi_{\mathcal{K}} \in \Pi(\phi_{\#}P_X^s, \phi_{\#}P_X^t)} \int_{\mathcal{H}_{\mathcal{K}} \times \mathcal{H}_{\mathcal{K}}} \|X^s - X^t\|_{\mathcal{H}_{\mathcal{K}}}^2 d\pi_{\mathcal{K}}(X^s, X^t) \quad (1)$$

Zhang *et al.* [59] show that if the pushforward measures are Gaussian, the Kernel Gaussian Wasserstein (KGW) distance between them can be written as

$$d_{\text{KGW}}^2(X^s, X^t) = \|u_X^s - u_X^t\|_{\mathcal{H}_{\mathcal{K}}}^2 + \text{tr}(\mathcal{C}_{XX}^{ss} + \mathcal{C}_{XX}^{tt} - 2\mathcal{C}_{XX}^{st}) \quad (2)$$

where $\mathcal{C}_{XX}^{st} = \sqrt{\sqrt{\mathcal{C}_{XX}^{ss}} \mathcal{C}_{XX}^{tt} \sqrt{\mathcal{C}_{XX}^{ss}}}$ and $\mathcal{C}_{XX}^{ss}, \mathcal{C}_{XX}^{tt}$ are the covariance operators of P_X^s and P_X^t on $\mathcal{H}_{\mathcal{K}}$, respectively.

Remark 1. Note that KGW consists of two terms, *i.e.*, the Maximum Mean Discrepancy (MMD) on first-order statistics and kernel Bures metric on second-order statistics. It provides a basic innovation for exploring conditional discrepancy metric by considering the conditional statistics in RKHS, *i.e.*, first-order statistics in literature [48] and second-order statistics in literature [13].

3.2 Generalization Error Analysis for DAR

Most of DAR methods focus on aligning the marginal distribution P_X . However, Zhao *et al.* [60] have shown that in the presence of label shift, only aligning the marginal distribution does not guarantee satisfying generalization performance on target domain. In fact, [6, 60] have proven the following theorem (1) for classification problems:

Theorem 1 ([60]). For representation variable $Z = g(X)$, suppose inequality $d_{JS}(P_Y^s, P_Y^t) \geq d_{JS}(P_Z^s, P_Z^t)$ holds, then for predictor $h : Z \rightarrow Y$, we have

$$\varepsilon_s(h \circ g) + \varepsilon_t(h \circ g) \geq \frac{1}{2} \left(\sqrt{d_{JS}(P_Y^s, P_Y^t)} - \sqrt{d_{JS}(P_Z^s, P_Z^t)} \right)^2, \quad (3)$$

where d_{JS} denotes the Jensen-Shannon divergence.

Assuming the label distributions differ between source and target domain, i.e., $d_{JS}(P_Y^s, P_Y^t) > 0$, the lower bound shows that good alignment of representation distributions leads to bad joint error. In other words, when label shift exists, which is rather common in realistic scenarios, only aligning the marginal distribution is not enough for low target error.

To deal with the insufficiency problem shown in the lower bound analysis, it is intuitive to introduce label information into generalization analysis, e.g., the balanced error rate for classification scenarios [6]. Analogously, for the continuous regression scenarios, we first define the continuous balanced error rate for quantifying maximum conditional risk in DAR.

Definition 1. The continuous balanced error rate of predictor h on distribution P is defined as:

$$\text{cBER}_P(h, Y) := \sup_{(x, y) \in \mathcal{X} \oplus \mathcal{Y}} P(h(x) \neq Y \mid Y = y). \quad (4)$$

Intuitively, the continuous balanced error rate can be regarded as the maximum potential risk for predicting sample x with label y . Based on cBER and conditional shift correction with representation transformation $Z = g(X)$, a sufficient condition for DAR can be induced as follows.

Theorem 2. For representation variable $Z^s = g(X^s)$ and $Z^t = g(X^t)$, if $g(\cdot)$ satisfies conditional invariant property, i.e.,

$$P_{Z^s|Y=y}^s = P_{Z^t|Y=y}^t, \quad \forall y \in \mathcal{Y}, \quad (5)$$

then for any predictor $h : Z \rightarrow Y$, we have

$$\varepsilon_s(h) + \varepsilon_t(h) \leq 2\text{cBER}_{P^s}(h, Y). \quad (6)$$

Therefore, the key for DA, no matter for classification or regression problems, is to achieve conditional distribution alignment. In classification settings, conditional alignment can be achieved by conducting class-wise marginal distribution alignment with the help of target pseudo label. However, in regression scenarios, the label variable is continuous, which means previous methods correcting conditional shift in classification settings are no longer applicable. Obviously, it is impossible to align the $P_{X|y}$ for every single y when y is a continuous variable and the possible number of y is infinite. Thus, it is important and meaningful to explore a new framework that is capable of conditional distribution alignment for continuous label variable.

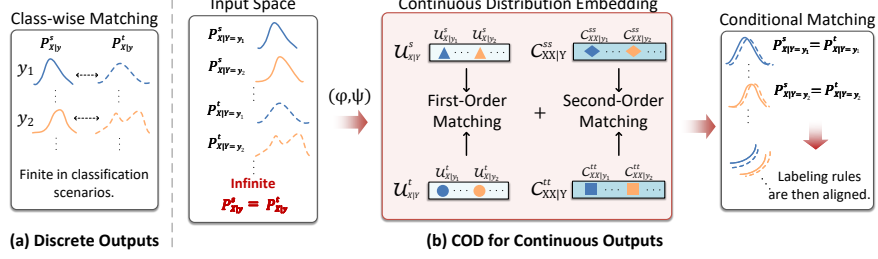


Fig. 2: Illustration of COD metric. **(a)** Classification DA methods usually correct conditional shift by class-wise computations on discrete clusters $P_{X|y}$, which is infeasible for regression due to the infinite slices $P_{X|y}$. **(b)** In COD, the continuous conditional distributions are embedded into the RKHS and characterized by finite statistical moments in kernel spaces, which do not rely on specific conditions y . Under the guarantees of distribution embedding property, the conditional alignment is equivalent to the matching on first-order and second-order statistics, *i.e.*, mean embedding operator $\mathcal{U}_{X|Y}$ and covariance operator $\mathcal{C}_{XX|Y}$. When COD is minimized, conditional alignment is achieved, and the level sets of cross-domain labeling rules are then aligned.

3.3 COD metric for DAR

Theory for COD. Due to the continuity of label variable, it is impractical to correct conditional shift by aligning marginal distribution of $P_{X|y}$ for every single y . For this inevitable obstacle, an intuitive idea is can we align all $P_{X|y}$ as a whole? *i.e.*, $P_{X|Y}$, rather than align $P_{X|y}$ one by one? It is more than attractive that if we can characterize $P_{X|Y}$ with a well-defined discrepancy, then all $P_{X|y}$ can be aligned by aligning the whole $P_{X|Y}$.

In kernel embedding theory, a distribution can be characterized by a point in RKHS. Then $P_{X|Y}$ can be seen as a point set in RKHS in which every element corresponds to a distribution $P_{X|y}$ for a fixed y . From another perspective, we can consider $P_{X|Y}$ as the distribution in RKHS and measure the cross-domain conditional discrepancy by defining the COD distance on their statistical moments in RKHS. Following the definitions of conditional statistics in pioneer works, *i.e.*, the conditional mean operator as $\mathcal{U}_{X|Y} := \mathcal{C}_{XY}\mathcal{C}_{YY}^{-1}$ [48] and conditional covariance operator as $\mathcal{C}_{XX|Y} = \mathcal{C}_{XX} - \mathcal{C}_{XY}\mathcal{C}_{YY}^{-1}\mathcal{C}_{YX}$ [13], we present the definition of COD as follows.

Definition 2 (COD). Given conditional distributions $P_{X|Y}^s$ and $P_{X|Y}^t$, denote $\mathcal{C}_{XX|Y}^{ss}$ and $\mathcal{C}_{XX|Y}^{tt}$ as their corresponding covariance operators. The COD is defined as

$$d_{\text{COD}}^2(P_{X|Y}^s, P_{X|Y}^t) = \|\mathcal{U}_{X|Y}^s - \mathcal{U}_{X|Y}^t\|_{\mathcal{H}_k}^2 + \text{tr}(\mathcal{C}_{XX|Y}^{ss} + \mathcal{C}_{XX|Y}^{tt} - 2\mathcal{C}_{XX|Y}^{st}), \quad (7)$$

where $\mathcal{C}_{XX|Y}^{st} = \sqrt{\mathcal{C}_{XX|Y}^{ss}\mathcal{C}_{XX|Y}^{tt}}\sqrt{\mathcal{C}_{XX|Y}^{ss}}$.

Theoretically, it can be proved that COD indeed admits the metric property on conditional distributions, *i.e.*, two conditional distributions are the same if

and only if the COD metric between them is zero. An intuitive illustration of COD is shown in Fig. 2. Formally, COD admits two key properties.

Proposition 1. (a) COD defines a metric between Gaussian measures.

(b) Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be the locally compact and Hausdorff measurable space and k be c_0 -universal kernel. Assuming that $(\phi(X), \psi(Y))$ is a Gaussian random variable in $\mathcal{H}_{\mathcal{X}} \oplus \mathcal{H}_{\mathcal{Y}}$. For any $P_{X|Y}^s, P_{X|Y}^t \in \mathbb{P}_{\mathcal{X}|Y}$, we have

$$d_{\text{COD}}^2(P_{X|Y}^s, P_{X|Y}^t) = 0 \implies P_{X|Y}^s = P_{X|Y}^t. \quad (8)$$

Empirical Estimation of COD. Based on the above properties, we can perform conditional distribution alignment by minimizing the COD metric between source and target domain. To do this, we shall derive the empirical estimation of COD, which can be obtained with the help of the kernel trick.

Let $(\mathbf{x}_i^s, \mathbf{y}_i^s)_{i=1}^{n_s}$ and $(\mathbf{x}_i^t, \mathbf{y}_i^t)_{i=1}^{n_t}$ be two set of samples drawn i.i.d. from source and target domain. For simplicity, n_s and n_t are both set to n . In kernel method, samples are mapped to RKHS $\mathcal{H}_{\mathcal{X}} \oplus \mathcal{H}_{\mathcal{Y}}$ by the implicit feature map (ϕ, ψ) . Denote the feature map matrices by (Φ^s, Ψ^s) and (Φ^t, Ψ^t) , the explicit kernel matrices $\mathbf{K}_X^{ss} = \Phi^s \Phi^{sT}$, $\mathbf{K}_Y^{ss} = \Psi^s \Psi^{sT}$ are computed as $(\mathbf{K}_X^{ss})_{ij} = k_{\mathcal{X}}(\mathbf{x}_i^s, \mathbf{x}_j^s)$, $(\mathbf{K}_Y^{ss})_{ij} = k_{\mathcal{Y}}(\mathbf{y}_i^s, \mathbf{y}_j^s)$, respectively. And so as \mathbf{K}_X^{tt} , \mathbf{K}_X^{st} and \mathbf{K}_Y^{tt} . Let \mathbf{I}_n be the n -dimensional identity matrix, $\mathbf{1}_n$ be the n -dimensional vector with all elements equal to 1. The n -dimensional centering matrix can be defined as $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$. Then the centralized kernel matrix is defined as $\mathbf{G} = \mathbf{H}_n \mathbf{K} \mathbf{H}_n$.

With above notations, the covariance operator \mathcal{C}_{XY} can be empirically estimated as $\hat{\mathcal{C}}_{XY} = \frac{1}{n} (\Phi - \hat{\mu}_X \mathbf{1}^T) (\Psi - \hat{\mu}_Y \mathbf{1}^T)^T = \frac{1}{n} \Phi \mathbf{H} \Psi^T$, then the conditional moment statistics can be estimated as

$$\hat{\mathcal{U}}_{X|Y} = \left(\frac{1}{n} \Phi \mathbf{H} \Psi^T \right) \left(\frac{1}{n} \Psi \mathbf{H} \Psi^T \right)^{-1}, \quad \hat{\mathcal{C}}_{X|Y} = \frac{1}{n} \Phi \mathbf{H}_n \mathbf{A} (\Phi \mathbf{H}_n \mathbf{A})^T, \quad (9)$$

where \mathbf{A} satisfies that $\mathbf{B} = \mathbf{A} \mathbf{A}^T$ and $B \triangleq \mathbf{I}_n - \frac{1}{n\varepsilon} \left[\mathbf{G}_Y - \mathbf{G}_Y (\mathbf{G}_Y + \varepsilon n \mathbf{I}_n)^{-1} \mathbf{G}_Y \right]$. Based on these estimations, the closed-form empirical estimation of the COD metric can be derived as follows.

Proposition 2. The empirical Conditional Operator Discrepancy distance is

$$\begin{aligned} \hat{d}_{\text{COD}}^2 &= \text{tr}(\mathbf{K}_Y^{ss} (\mathbf{K}_Y^{ss})^{-1} \mathbf{K}_X^{ss} (\mathbf{K}_Y^{ss})^{-1}) + \text{tr}(\mathbf{K}_Y^{tt} (\mathbf{K}_Y^{tt})^{-1} \mathbf{K}_X^{tt} (\mathbf{K}_Y^{tt})^{-1}) \\ &\quad - 2 \text{tr}(\mathbf{K}_Y^{ts} (\mathbf{K}_Y^{ss})^{-1} \mathbf{K}_{X^s X^t} (\mathbf{K}_Y^{tt})^{-1}) \\ &\quad + \varepsilon \text{tr} \left[\mathbf{G}_X^s (\varepsilon n \mathbf{I}_n + \mathbf{G}_Y^s)^{-1} \right] + \varepsilon \text{tr} \left[\mathbf{G}_X^t (\varepsilon n \mathbf{I}_n + \mathbf{G}_Y^t)^{-1} \right] \\ &\quad - \frac{2}{n} \left\| (\mathbf{H}_n \mathbf{A}^t)^T \mathbf{K}_X^{ts} (\mathbf{H}_n \mathbf{A}^s) \right\|_*, \end{aligned} \quad (10)$$

where $\|\cdot\|_*$ is the nuclear norm. Empirically, \mathbf{K}_Y^{-1} is approximated by the regularized formulation $(\mathbf{K}_Y + \lambda \mathbf{I})^{-1}$.

Discriminability of empirical COD. As discussed above, the first term of COD measures the distance of the first-order statistic between conditional distributions in RKHS. However, recent study [5, 54] shows that the mean matching

may reduce the discriminability when bringing the two distributions close, which undoubtedly degrades the performance of the algorithm. Therefore, it is worth making an in-depth analysis of the discriminability of COD as well.

Without loss of generality, we first consider Kronechker Delta kernel on conditioning variable Y , *i.e.*, $k(\mathbf{y}_1, \mathbf{y}_2) = \delta(\mathbf{y}_1, \mathbf{y}_2)$. More explicitly, $k(\mathbf{y}_1, \mathbf{y}_2)$ equals to 1 if $\mathbf{y}_1 = \mathbf{y}_2$ or 0 otherwise. For empirical scenarios with finite samples, the numbers of samples and observed labels are finite, which implies the number of distinct label c will be obviously no more than sample size n . For convenience, we call two samples are in the same class when their label values are the same. Consequently, the first term in COD can be reformulated as follows.

Proposition 3. *Under the conditions mentioned above, the empirical estimation of d_{CMMD} can be written as*

$$\begin{aligned} \widehat{d}_{CMMD}^2 &= \sum_{p=1}^c \frac{1}{(\lambda + n_p)^2} \left(\sum_{i,j} k(\mathbf{x}_{(p),i}^s, \mathbf{x}_{(p),j}^s) \right) \\ &\quad + \sum_{p=1}^c \frac{1}{(\lambda + n_p)^2} \left(\sum_{i,j} k(\mathbf{x}_{(p),i}^t, \mathbf{x}_{(p),j}^t) \right) \\ &\quad - 2 \sum_{p=1}^c \frac{1}{(\lambda + n_p)^2} \left(\sum_{i,j} k(\mathbf{x}_{(p),i}^s, \mathbf{x}_{(p),j}^t) \right), \end{aligned} \quad (11)$$

where n_i denotes the sample-size of i -th label such that $n = \sum_{i=1}^c n_i$, $\mathbf{x}_{(p),i}$ denotes the i -th sample of p -th label.

Eq. (11) contains three terms. The third term characterizes the cross-domain intra-class similarity. When Eq. (11) is minimized, the third term is maximized so that the cross-domain sample with the same label will be aligned, leading to conditional alignment as expected. However, the first two terms of the equation represent the intra-class similarity of both domains. Minimizing them will decrease the intra-class similarity, which means intra-class samples will be pushed away from each other and the discriminability of representations is degraded.

Remark 2. Note the Kronechker Delta kernel is defined as

$$k(\mathbf{y}_1, \mathbf{y}_2) = \begin{cases} 1 & \text{if } \mathbf{y}_1 = \mathbf{y}_2 \\ 0 & \text{otherwise} \end{cases}. \quad (12)$$

When it comes to the Gaussian kernel for empirical modeling, the case of $\mathbf{y}_1 = \mathbf{y}_2$ is the same; for other cases, the kernel values can be arbitrarily close to zero, *i.e.*, for any $\epsilon > 0$, there always width σ_0 , such that

$$k(\mathbf{y}_1, \mathbf{y}_2) = \exp(-\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2 / \sigma_0^2) < \epsilon. \quad (13)$$

Thus, it is generally reasonable to use Gaussian kernel, which admits a similar property to Kronechker Delta kernel, for empirical modeling.

Remark 3. Note that although we treat the number of distinct labels as finite numbers in mini-batch training, it does not mean class-wise matching can be conducted. In fact, even though the pseudo target label has finite range in mini-batch, it still admits the continuous property, which means we can not cluster samples as usually done in classification methods. When two samples have predicted labels with little numerical difference, they can be regarded as samples of the same class in classification settings. Nonetheless, this kind of clustering is infeasible in regression settings since the labels are continuous. Therefore, class-wise alignment is not practical even in mini-batch training.

With these observation, it is reasonable to reformulate Eq. (11) to promote the discriminability of representations:

$$\begin{aligned} \widehat{d}_{\text{CMMD}_{\text{mod}}} = & -\text{tr}(\mathbf{K}_Y^{tt}(\mathbf{K}_Y^{tt})^{-1}\mathbf{K}_X^{tt}(\mathbf{K}_Y^{tt})^{-1}) - \text{tr}(\mathbf{K}_Y^{tt}(\mathbf{K}_Y^{tt})^{-1}\mathbf{K}_X^{tt}(\mathbf{K}_Y^{tt})^{-1}) \\ & - 2\text{tr}(\mathbf{K}_{Y^t Y^s}(\mathbf{K}_Y^{ss})^{-1}\mathbf{K}_{X^s X^t}(\mathbf{K}_Y^{tt})^{-1}) \end{aligned} \quad (14)$$

Now minimizing Eq. (14) will maximize intra-class similarity regardless of domain, yielding conditional matching with great discriminability. Then the empirical COD is modified as

$$\begin{aligned} \widehat{d}_{\text{COD}_{\text{mod}}} = & -\text{tr}(\mathbf{K}_Y^{tt}(\mathbf{K}_Y^{tt})^{-1}\mathbf{K}_X^{tt}(\mathbf{K}_Y^{tt})^{-1}) - \text{tr}(\mathbf{K}_Y^{tt}(\mathbf{K}_Y^{tt})^{-1}\mathbf{K}_X^{tt}(\mathbf{K}_Y^{tt})^{-1}) \\ & - 2\text{tr}(\mathbf{K}_{Y^t Y^s}(\mathbf{K}_Y^{ss})^{-1}\mathbf{K}_{X^s X^t}(\mathbf{K}_Y^{tt})^{-1}) \\ & + \varepsilon \text{tr} \left[\mathbf{G}_X^s (\varepsilon n \mathbf{I}_n + \mathbf{G}_Y^s)^{-1} \right] + \varepsilon \text{tr} \left[\mathbf{G}_X^t (\varepsilon n \mathbf{I}_n + \mathbf{G}_Y^t)^{-1} \right] \\ & - \frac{2}{n} \left\| (\mathbf{H}_n \mathbf{A}^t)^T \mathbf{K}_X^{ts} (\mathbf{H}_n \mathbf{A}^s) \right\|_* \end{aligned} \quad (15)$$

3.4 COD-based DAR Modeling

According to theorem (2), when the conditional distributions $P_{X^s|Y}$ and $P_{X^t|Y}$ are aligned, a well-trained source predictor can gain promising performance on target domain. However, when the shift is significant, it is difficult for predictor to produce high-quality pseudo labels at the beginning. For this reason, the model is likely to miss the optimal optimization direction in the early training stage and thus fall into sub-optimal solutions. Therefore, the marginal alignment can improve the reliability of initial pseudo labels, and further benefit the conditional alignment. Overall, the learning objective of COD can be formulated as

$$\min_{g,h} \mathcal{L}_{\text{src}} + \lambda_1 \widehat{d}_{\text{COD}_{\text{mod}}} + \lambda_2 \widehat{d}_{\text{KGW}}, \quad (16)$$

where λ_1, λ_2 are the trade-off parameters. For regression task, the source risk \mathcal{L}_{src} is commonly set as Mean Square Error (MSE) over continuous outputs:

$$\mathcal{L}_{\text{src}} = \frac{1}{n^s} \sum_{i=1}^{n^s} \|\hat{\mathbf{y}}_i^s - \mathbf{y}_i^s\|_2^2 \quad (17)$$

where $\hat{\mathbf{y}}_i^s$ is the predicted value of the i -th source sample. Note that the marginal alignment term can be replaced by any other discrepancy metric, while we adopt KGW here to preserve the consistency with OT framework.

4 Experiments

We evaluate our method with several state-of-the-art domain adaptation methods on three regression benchmark datasets, implementation details and additional illustrations are provided in supplementary material.

dSprites [19] is a standard 2D synthetic dataset for deep representation learning, which contains three domains each with 737,280 images: Color (**C**), Noisy (**N**) and Scream (**S**). Following [4], we consider three factors for regression tasks: scale and two plane coordinates, while the orientation factor is excluded for its complexity. We evaluate all methods on six transfer tasks: $\mathbf{C} \rightarrow \mathbf{N}$, $\mathbf{C} \rightarrow \mathbf{S}$, $\mathbf{N} \rightarrow \mathbf{C}$, $\mathbf{N} \rightarrow \mathbf{S}$, $\mathbf{S} \rightarrow \mathbf{C}$ and $\mathbf{S} \rightarrow \mathbf{N}$. Following previous works [4,38], the sum of Mean Absolute Error (MAE) on all three sub-regression tasks is reported as the evaluation metric.

Biwi kinect [12] is a real-world dataset containing two domains according to gender: Female (**F**) with 5874 images and Male (**M**) with 9804 images. The three factors that can be employed for regression tasks are pitch, yaw and roll. We evaluate all methods on two transfer tasks: $\mathbf{F} \rightarrow \mathbf{M}$ and $\mathbf{M} \rightarrow \mathbf{F}$. The sum of MAE on three sub-regression tasks is reported.

MPI3D [16] is a simulation-to-real dataset of 3D objects which consist of three different domain: Toy (**T**), Realistic (**RC**) and Real (**RL**). Each domain contains 1,036,800 images. There are two factors in MPI3D that can be employed for regression tasks: a rotation about a vertical axis and a second rotation about a horizontal axis. We evaluate all methods on six transfer tasks: $\mathbf{RL} \rightarrow \mathbf{RC}$, $\mathbf{RL} \rightarrow \mathbf{T}$, $\mathbf{RC} \rightarrow \mathbf{T}$, $\mathbf{RC} \rightarrow \mathbf{RL}$, $\mathbf{T} \rightarrow \mathbf{RL}$ and $\mathbf{T} \rightarrow \mathbf{RC}$. The sum of MAE on two sub-regression tasks is reported.

Comparison methods. SOTA methods for UDA and DAR are selected: TCA [40], MCD [45], DAN [28], DANN [14], JDOT [8], AFN [58], RSD [4], DARE-GRAM [38]. Following previous works [4,38], Mean Absolute Error (MAE) is used as our evaluation metric across all regression tasks. In comparison experiment, the proposed DAR model in Eq. (16) is denoted as COD for convenience.

4.1 Results and Analysis

Comparison on dSprites. As shown in Tab. 1, COD achieves the best performance among all competing methods in nearly all tasks of dSprites. In $\mathbf{C} \rightarrow \mathbf{N}$ and $\mathbf{C} \rightarrow \mathbf{S}$, in which the Resnet-18 and UDA methods exhibit terrible performance because of the significant domain gap. Although SOTA DAR methods have greatly boosted the performance compared with classification methods, COD still achieves a significant improvement, reducing the error to only 40% of that produced by SOTA methods. These results demonstrate that the conditional invariant representations learned by COD indeed ensure better discriminability.

Comparison on Biwi Kinect. Since Biwi Kinect is the closest to real-world scenarios among the three benchmark datasets, the performance improvement of model is more challenging. In Tab. 2, the performance gap between UDA methods and DAR methods is relatively smaller, which implies the SOTA DAR

Table 1: Domain adaptation regression results (MAE) on **dSprites** (ResNet-18).

Method	C→N	C→S	N→C	N→S	S→C	S→N	Avg.
Resnet-18	0.94	0.90	0.16	0.65	0.08	0.26	0.498
TCA	0.94	0.87	0.19	0.66	0.10	0.23	0.498
MCD	0.81	0.81	0.17	0.65	0.07	0.19	0.450
JDOT	0.86	0.79	0.19	0.64	0.10	0.23	0.468
AFN	1.00	0.96	0.16	0.62	0.08	0.32	0.523
DAN	0.70	0.77	0.12	0.50	0.06	0.11	0.377
DANN	0.47	0.46	0.16	0.65	0.05	0.10	0.315
RSD	0.31	0.31	0.12	0.53	0.07	0.08	0.237
DARE-GRAM	0.30	0.20	0.11	0.25	0.05	0.07	0.164
COD	0.12	0.16	0.10	0.23	0.08	0.12	0.134

Table 2: Domain adaptation regression results (MAE) on **Biwi Kinect** (ResNet-18).

Method	M→F	F→M	Avg.
Resnet-18	0.29	0.38	0.335
TCA	0.31	0.39	0.350
MCD	0.31	0.37	0.340
JDOT	0.29	0.39	0.340
AFN	0.32	0.41	0.365
DAN	0.28	0.37	0.325
DANN	0.30	0.37	0.335
RSD	0.26	0.30	0.280
DARE-GRAM	0.23	0.29	0.260
COD	0.20	0.21	0.205

methods with marginal alignment indeed show limitations in the difficult transfer tasks. In such a scenario, the sufficiency of condition alignment is directly validated, where the MAE of COD is significantly lower than other SOTA methods. Thus, for real-world regression applications with complex shift, the COD model indeed shows superiority in both theoretical and methodological aspects.

Comparison on MPI3D. In MPI3D, the resolution remains the same with dSprites, while the main challenge is the large sample size. In Tab. 3, the MAE of COD is competitive with the SOTA method DARE-GRAM in the first four tasks, while higher in the last two tasks. However, note that COD still achieves the second-best result in averaged MAE, and obtains the lowest MAE in task RL→T. In conclusion, the overall MAE of COD on three benchmarks is still lower, which demonstrates COD is generally superior to SOTA methods.

Ablation study. The ablation experiment was conducted on the most challenging tasks from all three datasets: **C** → **S** and **N** → **S** from dSprites, **T** → **RL** and **T** → **RC** from MPI3D, and both **M** → **F** and **F** → **M** from Biwi Kinect. The major components of proposed method, *i.e.*, MSE objective \mathcal{L}_{src} , KGW distance d_{KGW} , COD distance d_{COD} and modified COD $d_{COD_{mod}}$, are evaluated, where the results are shown in Tab. 4.

Table 3: Domain adaptation regression results (MAE) on **MPI3D** (ResNet-18).

Method	RL→ RC	RL→ T	RC→ RL	RC→ T	T→ RL	T→ RC	Avg.
Resnet-18	0.17	0.44	0.19	0.45	0.51	0.50	0.377
TCA	0.17	0.42	0.19	0.42	0.50	0.50	0.373
MCD	0.13	0.40	0.15	0.45	0.52	0.50	0.358
JDOT	0.16	0.41	0.16	0.41	0.47	0.47	0.353
AFN	0.18	0.45	0.20	0.46	0.53	0.53	0.390
DAN	0.12	0.35	0.12	0.27	0.40	0.41	0.278
DANN	0.09	0.24	0.11	0.41	0.48	0.37	0.283
RSD	0.09	0.19	0.08	0.15	0.36	0.36	0.205
DARE-GRAM	0.09	0.15	0.10	0.14	0.24	0.24	0.160
COD	0.10	0.11	0.13	0.15	0.37	0.29	0.192

Table 4: Ablation study on different metrics for DAR.

MSE	Objectives			dSprites		Biwi Kinect		MPI3D	
	d_{KGW}	d_{COD}	$d_{\text{COD}_{\text{mod}}}$	C→ S	N→ S	T→ RL	T→ RC	M→ F	F→ M
✓				0.90	0.65	0.29	0.38	0.51	0.50
✓	✓			0.18	0.26	0.21	0.25	0.41	0.34
✓		✓		0.26	0.45	0.21	0.22	0.40	0.31
✓	✓	✓		0.17	0.25	0.21	0.22	0.41	0.29
✓	✓		✓	0.16	0.23	0.20	0.21	0.37	0.29

From the 2nd and 3rd rows, it can be observed that both OT-based metrics show promising results on DAR datasets, where the conditional alignment with COD shows better performance than marginal alignment with KGW on MPI3D and Biwi Kinect datasets. On the other hand, since dSprites induce more difficult transfer problems, the baseline performance is generally lower, and the pseudo labels for conditional alignment are less reliable. Therefore, the combination of KGW and COD is necessary to ensure higher performance on challenging tasks, where the results in the 4th row validate the mutually beneficial relation between these two objectives. Finally, the 5th row shows the effectiveness of rethinking the discriminability of COD, which implies the modifications on the moment statistics indeed ensure lower MAE values on all datasets. Overall, these results demonstrate that conditional alignment is indeed crucial for successful DAR, and the theoretical results and proposed method are generally valid.

Feature Visualization. t-SNE [34] is employed to visualize the 2-D features of different alignments on dSprites C→N task. **(a)** Before adaptation, distributions of source and target domain differ severely, as shown in Fig. 3(a), which implies predictor is infeasible for the target samples with a significant domain gap. **(b)** After marginal distribution alignment via the SOTA DARE-GRAM method, source and target representations are aligned globally, and the cross-domain distributions are overlapped on the whole. However, there are still misaligned local patterns where source and target samples of different colors are matched.

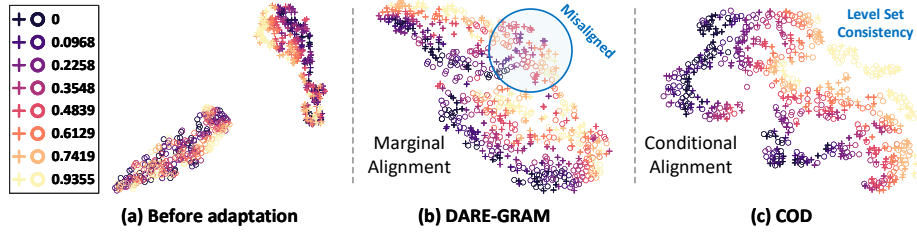


Fig. 3: t-SNE visualization of learned representations. Label values are denoted by color gradients. Eight label values are selected from the range of the variable for visualization. '+' : source samples; 'o' : target samples.

Therefore, it is hard to fit clear level sets in Fig. 3 (b). (c) After conditional distribution alignment via COD, source and target distributions are aligned not only globally but also locally. As shown in Fig. 3 (c), representations are matched according to label values so the change of color shows a clear gradient and level sets suitable for both domains can be fitted.

Hyper-parameter. We investigate the sensitivity of hyper-parameters on dSprites $C \rightarrow N$. MAE values by varying λ_1 and λ_2 , *i.e.*, parameters of modified COD and KGW, are shown in Fig. 4. The results show that the performance of the COD-based DAR model is generally robust to the different choice of parameters. In fact, the MAE values of the COD-based model are stable in random experiments, which implies the proposed method is both robust from the view of randomness and parameter setting. In conclusion, the stable performance validates that the proposed model is also empirically reliable.

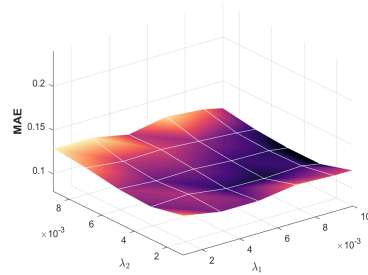


Fig. 4: MAE values under different settings of hyper-parameters.

5 Conclusion

In this work, we studied the limitations of existing DAR research, *i.e.*, the misaligned intrinsic discriminant structure and negative transfer of marginal distribution matching framework. To deal with the challenges of conditional shift with continuous output space, we provide feasible solutions from both theoretical and methodological views, where the main results show that: 1) the cross-domain generalization regression error can be sufficiently bounded by the conditional discrepancy on continuous labels; 2) the proposed COD serves as a well-defined metric for conditional discrepancy measures and empirical modeling. With these guarantees, a COD-based DAR method is proposed with a discriminability enhancement mechanism. Numerical validations on standard DAR datasets validate the superiority of COD.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (Grant No. 62376291), in part by Guangdong Basic and Applied Basic Research Foundation (2023B1515020004), in part by Science and Technology Program of Guangzhou (2024A04J6413), in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (24xkjc013), in part by Guangdong Province Key Laboratory of Computational Science at Sun Yat-sen University (2020B1212060032), and in part by Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education.

References

1. Bao, Y., Liu, Y., Wang, H., Lu, F.: Generalizing gaze estimation with rotation consistency. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4207–4216 (2022)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. *Mach. Learn.* **79**, 151–175 (2010)
3. Chen, L., Chen, H., Wei, Z., Jin, X., Tan, X., Jin, Y., Chen, E.: Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7181–7190 (2022)
4. Chen, X., Wang, S., Wang, J., Long, M.: Representation subspace distance for domain adaptation regression. In: Int. Conf. Mach. Learn. pp. 1749–1759 (2021)
5. Chen, Y., Song, S., Li, S., Wu, C.: A graph embedding framework for maximum mean discrepancy-based domain adaptation algorithms. *IEEE Trans. Image Process.* **29**, 199–213 (2019)
6. Tachet des Combes, R., Zhao, H., Wang, Y.X., Gordon, G.J.: Domain adaptation with conditional distribution matching and generalized label shift. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 19276–19289 (2020)
7. Cortes, C., Mohri, M.: Domain adaptation in regression. In: ICALT. pp. 308–323 (2011)
8. Courty, N., Flamary, R., Habrard, A., Rakotomamonjy, A.: Joint distribution optimal transportation for domain adaptation. *Adv. Neural Inform. Process. Syst.* **30** (2017)
9. Courty, N., Flamary, R., Tuia, D.: Domain adaptation with regularized optimal transport. In: ECML PKDD. pp. 274–289 (2014)
10. Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: Eur. Conf. Comput. Vis. pp. 447–463 (2018)
11. Dhaini, M., Berar, M., Honeine, P., Van Exem, A.: Unsupervised domain adaptation for regression using dictionary learning. *Knowl.-Based Syst.* **267**, 110439 (2023)
12. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. *Int. J. Comput. Vis.* **101**, 437–458 (2013)
13. Fukumizu, K., Bach, F.R., Jordan, M.I.: Kernel dimension reduction in regression. *Ann. Stat.* **37**(4), 1871–1905 (2009)
14. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**(59), 1–35 (2016)

15. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(12), 2234–2240 (2007)
16. Gondal, M.W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., Bauer, S.: On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Adv. Neural Inform. Process. Syst.* **32** (2019)
17. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2066–2073 (2012)
18. Goswami, S., Kontolati, K., Shields, M.D., Karniadakis, G.E.: Deep transfer operator learning for partial differential equations under conditional shift. *Nat. Mach. Intell.* **4**(12), 1155–1164 (2022)
19. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: *Int. Conf. Learn. Represent.* (2016)
20. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: *Int. Conf. Mach. Learn.* pp. 1989–1998 (2018)
21. Hu, J., Zhong, H., Yang, F., Gong, S., Wu, G., Yan, J.: Learning unbiased transferability for domain adaptation by uncertainty modeling. In: *Eur. Conf. Comput. Vis.* pp. 223–241 (2022)
22. Jiang, J., Ji, Y., Wang, X., Liu, Y., Wang, J., Long, M.: Regressive domain adaptation for unsupervised keypoint detection. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 6780–6789 (2021)
23. Kirchmeyer, M., Rakotomamonjy, A., de Bezenac, E., Gallinari, P.: Mapping conditional distributions for domain adaptation under generalized target shift. In: *Int. Conf. Learn. Represent.* (2022)
24. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced wasserstein discrepancy for unsupervised domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 10285–10295 (2019)
25. Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., Zhao, H.: Learning invariant representations and risks for semi-supervised domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 1104–1113 (2021)
26. Li, X., Li, J., Zuo, L., Zhu, L., Shen, H.T.: Domain adaptive remaining useful life prediction with transformer. *IEEE Trans. Instrum.* **71**, 1–13 (2022)
27. Li, Z., Zhang, K., Liu, Y., Zou, Y., Ding, G.: A novel remaining useful life transfer prediction method of rolling bearings based on working conditions common benchmark. *IEEE Trans. Instrum.* **71**, 1–9 (2022)
28. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: *Int. Conf. Mach. Learn.* pp. 97–105 (2015)
29. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. *Adv. Neural Inform. Process. Syst.* **31** (2018)
30. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: *Int. Conf. Mach. Learn.* pp. 2208–2217 (2017)
31. Luo, Y.W., Ren, C.X.: Conditional bures metric for domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 13989–13998 (2021)
32. Luo, Y.W., Ren, C.X.: Mot: Masked optimal transport for partial domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 3531–3540 (2023)
33. Luo, Y.W., Ren, C.X., Dai, D.Q., Yan, H.: Unsupervised domain adaptation via discriminative manifold propagation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(3), 1653–1669 (2022)

34. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**(11) (2008)
35. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430* (2009)
36. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., den Hengel, A.v.: Poseur: Direct human pose regression with transformers. In: *Eur. Conf. Comput. Vis.* pp. 72–88 (2022)
37. de Mathelin, A., Richard, G., Deheeger, F., Mougeot, M., Vayatis, N.: Adversarial weighting for domain adaptation in regression. In: *ICTAI*. pp. 49–56 (2021)
38. Nejjar, I., Wang, Q., Fink, O.: Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 11744–11754 (2023)
39. Nikzad-Langerodi, R., Zellinger, W., Saminger-Platz, S., Moser, B.A.: Domain adaptation for regression under beer–lambert’s law. *Knowl.-Based Syst.* **210**, 106447 (2020)
40. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **22**(2), 199–210 (2010)
41. Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation. In: *AAAI*. vol. 32 (2018)
42. Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R.: Rethinking depth estimation for multi-view stereo: A unified representation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 8645–8654 (2022)
43. Quinero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: *Dataset shift in machine learning*. Mit Press (2008)
44. Ren, C.X., Luo, Y.W., Dai, D.Q.: Buresnet: Conditional bures metric for transferable representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(4), 4198–4213 (2023)
45. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 3723–3732 (2018)
46. Shen, J., Qu, Y., Zhang, W., Yu, Y.: Wasserstein distance guided representation learning for domain adaptation. In: *AAAI*. vol. 32 (2018)
47. Singh, A., Chakraborty, S.: Deep domain adaptation for regression. *Development and Analysis of Deep Learning Architectures* pp. 91–115 (2020)
48. Song, L., Huang, J., Smola, A., Fukumizu, K.: Hilbert space embeddings of conditional distributions with applications to dynamical systems. In: *Int. Conf. Mach. Learn.* pp. 961–968 (2009)
49. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: *AAAI*. vol. 30 (2016)
50. Tang, H., Jia, K.: Discriminative adversarial domain adaptation. In: *AAAI*. vol. 34, pp. 5940–5947 (2020)
51. Teshima, T., Sato, I., Sugiyama, M.: Few-shot domain adaptation by causal mechanism transfer. In: *Int. Conf. Mach. Learn.* pp. 9458–9469 (2020)
52. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 7167–7176 (2017)
53. Wang, H., Shen, T., Zhang, W., Duan, L.Y., Mei, T.: Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: *Eur. Conf. Comput. Vis.* pp. 642–659 (2020)
54. Wang, W., Li, H., Ding, Z., Nie, F., Chen, J., Dong, X., Wang, Z.: Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Trans. Neural Netw. Learn. Syst.* (2021)

55. Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., Li, T.: Contrastive regression for domain adaptation on gaze estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19376–19385 (2022)
56. Wei, P., Ke, Y., Ong, Y.S., Ma, Z.: Adaptive transfer kernel learning for transfer gaussian process regression. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
57. Wu, J., He, J., Wang, S., Guan, K., Ainsworth, E.: Distribution-informed neural networks for domain adaptation regression. Adv. Neural Inform. Process. Syst. **35**, 10040–10054 (2022)
58. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: Int. Conf. Comput. Vis. pp. 1426–1435 (2019)
59. Zhang, Z., Wang, M., Nehorai, A.: Optimal transport in reproducing kernel hilbert spaces: Theory and applications. IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1741–1754 (2019)
60. Zhao, H., Combes, R.T.D., Zhang, K., Gordon, G.: On learning invariant representations for domain adaptation. In: Int. Conf. Mach. Learn. pp. 7523–7532 (2019)
61. Zhao, H., Coston, A., Adel, T., Gordon, G.J.: Conditional learning of fair representations. In: Int. Conf. Learn. Represent. (2020)
62. Zhao, H., Dan, C., Aragam, B., Jaakkola, T.S., Gordon, G.J., Ravikumar, P.: Fundamental limits and tradeoffs in invariant representation learning. J. Mach. Learn. Res. **23**(340), 1–49 (2022)