SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Maximizing Conditional Independence for Unsupervised Domain Adaptation

Yi-Ming Zhai[1], Chuan-Xian Ren[1*], You-Wei Luo[1] & Dao-Qing Dai[1]

[1]*School of Mathematics, Sun Yat-Sen University, Guangzhou* 510275*, China*

Appendix A Main Proofs

Appendix A.1 Class-conditional Distribution Alignment

Theorem 1. Assume that the product $k_{\ddot{\mathcal{X}}}k_{\mathcal{Z}}$ is a characteristic kernel on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$, and $\mathcal{F}_{\mathcal{Y}}+\mathbb{R}$ is dense in $L^2(P_{\mathcal{Y}})$. For any conditional distributions $P^S_{X|Y}$, $P^T_{X|Y}\in \Pr^S(\mathcal{X}|\mathcal{Y})$, we have

$$V_{\ddot{Z}\ddot{X}|Y}=0\Longrightarrow P^S_{X|Y}=P^T_{X|Y}.$$

Proof. Based on Lemma 1, we have

$$V_{\ddot{Z}\ddot{X}|Y}=0\Longrightarrow X\perp\!\!\!\perp Z\mid Y.$$

With the property of conditional independence, we have

$$P(X\mid Z,Y)=P(X\mid Y).$$

Thus,

$$P^S_{X|Y}=P^T_{X|Y}.$$

Appendix A.2 Empirical Estimation

The empirical estimation of the dependence can be expressed as

$$\hat{I}_n^{NOCCO}(X,Z)=\mathrm{Tr}(\mathbf{R_Z}\mathbf{I}_n\mathbf{R_X}\mathbf{I}_n).$$

Proof. Provided that V_{YX} is Hilbert-Schmidt, the dependence of X and Z is measured by

$$I^{NOCCO}(X,Z)=\|V_{ZX}\|^2_{HS}. \tag{A1}$$

Denote $\mathbf{X}\in\mathbb{R}^{d_1\times n}$ and $\mathbf{Z}\in\mathbb{R}^{d_2\times n}$ as the observations of random variables X and Z, respectively. The kernel feature maps $\phi(\cdot)$ and $\lambda(\cdot)$ are used to map $\mathbf{X}$ and $\mathbf{Z}$ into the Reproducing Kernel Hilbert Space (RKHS). Then, we obtain kernel matrices

$$\mathbf{K_X}=\mathbf{\Phi}^T\mathbf{\Phi},\quad \mathbf{K_Z}=\mathbf{\Lambda}^T\mathbf{\Lambda},$$

where $\mathbf{K_X}$, $\mathbf{K_Z}\in\mathbb{R}^{n\times n}$, and

$$\mathbf{\Phi}=[\phi(\mathbf{x}_1),\phi(\mathbf{x}_2),\cdots,\phi(\mathbf{x}_n)],$$
$$\mathbf{\Lambda}=[\lambda(\mathbf{z}_1),\lambda(\mathbf{z}_2),\cdots,\lambda(\mathbf{z}_n)].$$

Thus, the empirical estimation of the cross-covariance operator Σ_{ZX} is written as

$$\widehat{\Sigma}^{(n)}_{ZX}=\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n},$$

where $\mathbf{H}_n=\mathbf{I}_n-\frac{\mathbf{1}\mathbf{1}^T}{n}$ is the centering matrix and $\mathbf{1}\in\mathbb{R}^n$ is the all-ones vector. Similarly, the estimations of covariance operators $\widehat{\Sigma}^{(n)}_{XX}$ and $\widehat{\Sigma}^{(n)}_{ZZ}$ can be written as

$$\widehat{\Sigma}^{(n)}_{XX}=\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n},\qquad \widehat{\Sigma}^{(n)}_{ZZ}=\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n}.$$

* Corresponding author (email: rchuanx@mail.sysu.edu.cn)

Thus, by regularizing the singular covariance operators with ε_n [51], the normalized cross-covariance operator (NOCCO) V_{ZX} can be estimated by

$$\widehat{V}_{ZX}^{(n)} = \left(\Sigma_{ZZ}^{(n)} + \varepsilon_n I\right)^{-1/2} \widehat{\Sigma}_{ZX}^{(n)} \left(\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n I\right)^{-1/2} \tag{A2}$$

$$= \mathrm{Tr}\left(\left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2}\right). \tag{A3}$$

Based on (A3), we can provide an empirical estimation of (A1) as

$$\begin{aligned}
\hat{I}_n^{NOCCO} =& \|\widehat{V}_{ZX}^{(n)}\|_{HS}^2 \\
=& \mathrm{Tr}\left(\widehat{V}_{ZX}^{(n)*}\widehat{V}_{ZX}^{(n)}\right) \\
=& \mathrm{Tr}\left(\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n} \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2}\right) \\
=& \mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n} \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1}\right).
\end{aligned} \tag{A4}$$

Here we have

$$\left(\frac{\mathbf{\Phi}\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1} = \frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\left(\varepsilon_n n\mathbf{I}_n + \mathbf{K}_X\right)^{-1}\mathbf{\Phi}^T\right),$$

where $\mathbf{K}_X = \mathbf{\Phi}^T\mathbf{\Phi}$. $\mathbf{I}$ and $\mathbf{I}_n$ are different in this equation, the former can be regarded as an operator and the latter is the identity matrix with dimension n. $\mathbf{H}_n$ is a symmetric idempotent matrix, which satisfies $\mathbf{H}_n\mathbf{H}_n = \mathbf{H}_n$. Denote $\mathbf{G}_X = \mathbf{H}_n\mathbf{K}_X\mathbf{H}_n$. Therefore,

$$\begin{aligned}
\left(\widehat{\Sigma}_{XX}^{(n)} + \varepsilon_n\mathbf{I}\right)^{-1} &= \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1} \\
&= \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{H}_n^T\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1} \\
&= \frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{H}_n^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{H}_n\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right) \\
&= \frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{H}_n^T\mathbf{K}_X\mathbf{H}_n\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right) \\
&= \frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_X\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right).
\end{aligned} \tag{A5}$$

Thus, we can employ (A5) to calculate (A4) as

$$\begin{aligned}
\hat{I}_n^{NOCCO} =& \|\widehat{V}_{ZX}^{(n)}\|_{HS}^2 \\
=& \mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n}\frac{1}{\varepsilon_n}[\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_Z\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T]\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n}\frac{1}{\varepsilon_n}[\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_X\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T]\right) \\
=& \mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{H}_n\mathbf{\Lambda}^T}{n}\frac{1}{\varepsilon_n}[\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_Z\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T]\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{H}_n\mathbf{\Phi}^T}{n}\frac{1}{\varepsilon_n}[\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_X\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T]\right) \\
=& \mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{H}_n\mathbf{\Lambda}^T}{n}\frac{1}{\varepsilon_n}[\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_Z\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T]\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{H}_n\mathbf{\Phi}^T}{n}\frac{1}{\varepsilon_n}[\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_X\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T]\right) \\
=& \mathrm{Tr}\left(\mathbf{H}_n\mathbf{\Lambda}^T[\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_Z\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T]\frac{\mathbf{\Lambda}\mathbf{H}_n}{n\varepsilon_n}\mathbf{H}_n\mathbf{\Phi}^T[\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_X\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T]\frac{\mathbf{\Phi}\mathbf{H}_n}{n\varepsilon_n}\right) \\
=& \mathrm{Tr}\left(\frac{\mathbf{G}_Z - \mathbf{G}_Z\left(\mathbf{G}_Z + \varepsilon_n n\mathbf{I}_n\right)^{-1}\mathbf{G}_Z}{n\varepsilon_n}\frac{\mathbf{G}_X - \mathbf{G}_X\left(\mathbf{G}_X + \varepsilon_n n\mathbf{I}_n\right)^{-1}\mathbf{G}_X}{n\varepsilon_n}\right) \\
=& \mathrm{Tr}\left(\mathbf{G}_Z\left(\mathbf{G}_Z + \varepsilon_n n\mathbf{I}_n\right)^{-1}\mathbf{G}_X\left(\mathbf{G}_X + \varepsilon_n n\mathbf{I}_n\right)^{-1}\right) \\
=& \mathrm{Tr}\left(\mathbf{R}_Z\mathbf{R}_X\right) \\
=& \mathrm{Tr}(\mathbf{R}_\mathbf{Z}\mathbf{I}_n\mathbf{R}_\mathbf{X}\mathbf{I}_n).
\end{aligned}$$

Note: $\mathbf{I}_n - \left(\mathbf{G} + n\varepsilon_n\mathbf{I}_n\right)^{-1}\mathbf{G} = \varepsilon_n n\left(\mathbf{G} + \varepsilon_n n\mathbf{I}_n\right)^{-1}$.

Theorem 2. Denote $\mathbf{S} = \mathbf{I}_n - \mathbf{R}_\mathbf{Y}$. The empirical estimation of the conditional dependence is

$$\hat{I}_n^{COND}(X, Z|Y) = \mathrm{Tr}(\mathbf{R}_{\ddot{\mathbf{Z}}}\mathbf{S}\mathbf{R}_{\ddot{\mathbf{X}}}\mathbf{S}). \tag{A6}$$

Proof. Denote $\ddot{X} = (X, Y)$, $\ddot{Z} = (Z, Y)$ with the kernel product $k_{\ddot{\mathcal{X}}} = k_\mathcal{X}k_\mathcal{Y}$ and $k_{\ddot{\mathcal{Z}}} = k_\mathcal{Z}k_\mathcal{Y}$. Lemma 1 implies that $X \perp\!\!\!\perp Z \mid Y$ if and only if $V_{\ddot{Z}\ddot{X}|Y} = 0$. Provided that $V_{\ddot{Z}\ddot{X}|Y}$ is Hilbert-Schmidt, the conditional dependence of X and Z given Y is measured by

$$I^{COND}(X, Z \mid Y) = \|V_{\ddot{Z}\ddot{X}|Y}\|_{HS}^2. \tag{A7}$$

Denote $\mathbf{X} \in \mathbb{R}^{d_1 \times n}$, $\mathbf{Y} \in \mathbb{R}^{K \times n}$, and $\mathbf{Z} \in \mathbb{R}^{d_2 \times n}$ as the observations of random variables X, Y and Z, respectively. The kernel feature maps $\phi(\cdot)$, $\lambda(\cdot)$ and $\psi(\cdot)$ are used to map $\ddot{\mathbf{X}}$, $\ddot{\mathbf{Z}}$ and $\mathbf{Y}$ into RKHSs, where $\ddot{\mathbf{X}} = (\mathbf{X}_{re}; \mathbf{Y}) \in \mathbb{R}^{(d'+K)\times n}$, $\ddot{\mathbf{Z}} = (\mathbf{Z}; \mathbf{Y}) \in \mathbb{R}^{(d_2+K)\times n}$. Then, we obtain kernel matrices

$$\mathbf{K}_{\ddot{\mathbf{X}}} = \mathbf{\Phi}^T\mathbf{\Phi}, \quad \mathbf{K}_{\ddot{\mathbf{Z}}} = \mathbf{\Lambda}^T\mathbf{\Lambda}, \quad \mathbf{K}_{\mathbf{Y}} = \mathbf{\Psi}^T\mathbf{\Psi},$$

where $\mathbf{K}_{\ddot{\mathbf{X}}}$, $\mathbf{K}_{\ddot{\mathbf{Z}}}$, $\mathbf{K}_{\mathbf{Y}} \in \mathbb{R}^{n\times n}$, and

$$\begin{aligned}
\mathbf{\Phi} &= [\phi(\ddot{\mathbf{x}}_1), \phi(\ddot{\mathbf{x}}_2), \cdots, \phi(\ddot{\mathbf{x}}_n)],\\
\mathbf{\Lambda} &= [\lambda(\ddot{\mathbf{z}}_1), \lambda(\ddot{\mathbf{z}}_2), \cdots, \lambda(\ddot{\mathbf{z}}_n)],\\
\mathbf{\Psi} &= [\psi(\mathbf{y}_1), \psi(\mathbf{y}_2), \cdots, \psi(\mathbf{y}_n)].
\end{aligned}$$

Similarly with the estimation of the NOCCO, we have

$$\begin{aligned}
\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}} &= \left(\widehat{\Sigma}^{(n)}_{\ddot{Z}\ddot{Z}} + \varepsilon_n I\right)^{-1/2} \widehat{\Sigma}^{(n)}_{\ddot{Z}\ddot{X}} \left(\widehat{\Sigma}^{(n)}_{\ddot{X}\ddot{X}} + \varepsilon_n I\right)^{-1/2}\\
&= \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n I\right)^{-1/2} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n I\right)^{-1/2},
\end{aligned}$$

$$\begin{aligned}
\widehat{V}^{(n)}_{\ddot{Z}Y} &= \left(\widehat{\Sigma}^{(n)}_{\ddot{Z}\ddot{Z}} + \varepsilon_n I\right)^{-1/2} \widehat{\Sigma}^{(n)}_{\ddot{Z}Y} \left(\widehat{\Sigma}^{(n)}_{YY} + \varepsilon_n I\right)^{-1/2}\\
&= \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n I\right)^{-1/2} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Psi}^T}{n} \left(\frac{\mathbf{\Psi}\mathbf{H}_n\mathbf{\Psi}^T}{n} + \varepsilon_n I\right)^{-1/2},
\end{aligned}$$

$$\begin{aligned}
\widehat{V}^{(n)}_{Y\ddot{X}} &= \left(\widehat{\Sigma}^{(n)}_{YY} + \varepsilon_n I\right)^{-1/2} \widehat{\Sigma}^{(n)}_{Y\ddot{X}} \left(\widehat{\Sigma}^{(n)}_{\ddot{X}\ddot{X}} + \varepsilon_n I\right)^{-1/2}\\
&= \left(\frac{\mathbf{\Psi}\mathbf{H}_n\mathbf{\Psi}^T}{n} + \varepsilon_n I\right)^{-1/2} \frac{\mathbf{\Psi}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n I\right)^{-1/2}.
\end{aligned}$$

Based on the estimation of the cross-covariance operators $\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}}$, $\widehat{V}^{(n)}_{\ddot{Z}Y}$, and $\widehat{V}^{(n)}_{Y\ddot{X}}$, the estimation of $V_{\ddot{Z}\ddot{X}|Y}$ is written as

$$\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}|Y} = \widehat{V}^{(n)}_{\ddot{Z}\ddot{X}} - \widehat{V}^{(n)}_{\ddot{Z}Y}\widehat{V}^{(n)}_{Y\ddot{X}}.$$

Therefore, the empirical estimation of (A7) is

$$\begin{aligned}
\hat{I}^{COND}_n &\equiv \| \widehat{V}^{(n)}_{\ddot{Z}\ddot{X}|Y} \|^2_{HS}\\
&= \mathrm{Tr}\left(\widehat{V}^{(n)^*}_{\ddot{Z}\ddot{X}|Y} \widehat{V}^{(n)}_{\ddot{Z}\ddot{X}|Y}\right)\\
&= \mathrm{Tr}\left(\left(\widehat{V}^{(n)^*}_{\ddot{Z}\ddot{X}} - \widehat{V}^{(n)^*}_{Y\ddot{X}}\widehat{V}^{(n)^*}_{\ddot{Z}Y}\right)\left(\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}} - \widehat{V}^{(n)}_{\ddot{Z}Y}\widehat{V}^{(n)}_{Y\ddot{X}}\right)\right)\\
&= \mathrm{Tr}\left(\widehat{V}^{(n)^*}_{\ddot{Z}\ddot{X}}\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}} - \widehat{V}^{(n)^*}_{Y\ddot{X}}\widehat{V}^{(n)^*}_{\ddot{Z}Y}\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}} - \widehat{V}^{(n)^*}_{\ddot{Z}\ddot{X}}\widehat{V}^{(n)}_{\ddot{Z}Y}\widehat{V}^{(n)}_{Y\ddot{X}} + \widehat{V}^{(n)^*}_{Y\ddot{X}}\widehat{V}^{(n)^*}_{\ddot{Z}Y}\widehat{V}^{(n)}_{\ddot{Z}Y}\widehat{V}^{(n)}_{Y\ddot{X}}\right),
\end{aligned} \tag{A8}$$

where

$$\begin{aligned}
&\mathrm{Tr}\left(\widehat{V}^{(n)^*}_{\ddot{Z}\ddot{X}}\widehat{V}^{(n)}_{\ddot{Z}\ddot{X}}\right)\\
=&\mathrm{Tr}\left(\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n} \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1/2}\right)\\
=&\mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n}\left(\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Lambda}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1} \frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n} \left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Phi}^T}{n} + \varepsilon_n\mathbf{I}\right)^{-1}\right)\\
=&\mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{\Lambda}^T}{n}\frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{Z}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T\right)\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{\Phi}^T}{n}\frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{X}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right)\right)\\
=&\mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{H}_n\mathbf{\Lambda}^T}{n}\frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{Z}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T\right)\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{H}_n\mathbf{\Phi}^T}{n}\frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{X}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right)\right)\\
=&\mathrm{Tr}\left(\frac{\mathbf{\Phi}\mathbf{H}_n\mathbf{H}_n\mathbf{\Lambda}^T}{n}\frac{1}{\varepsilon_n}\left(\mathbf{I} - \lambda\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{Z}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T\right)\frac{\mathbf{\Lambda}\mathbf{H}_n\mathbf{H}_n\mathbf{\Phi}^T}{n}\frac{1}{\varepsilon_n}\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_X\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right)\right)\\
=&\mathrm{Tr}\left(\mathbf{H}_n\mathbf{\Lambda}^T\left(\mathbf{I} - \mathbf{\Lambda}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{Z}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Lambda}^T\right)\frac{\mathbf{\Lambda}\mathbf{H}_n}{n\varepsilon_n}\mathbf{H}_n\mathbf{\Phi}^T\left(\mathbf{I} - \mathbf{\Phi}\mathbf{H}_n\left(\varepsilon_n n\mathbf{I}_n + \mathbf{G}_{\ddot{X}}\right)^{-1}\mathbf{H}_n^T\mathbf{\Phi}^T\right)\frac{\mathbf{\Phi}\mathbf{H}_n}{n\varepsilon_n}\right)\\
=&\mathrm{Tr}\left(\frac{\mathbf{G}_{\ddot{Z}} - \mathbf{G}_{\ddot{Z}}\left(\mathbf{G}_{\ddot{Z}} + \varepsilon_n n\mathbf{I}_n\right)^{-1}\mathbf{G}_{\ddot{Z}}}{n\varepsilon_n}\frac{\mathbf{G}_{\ddot{X}} - \mathbf{G}_{\ddot{X}}\left(\mathbf{G}_{\ddot{X}} + \varepsilon_n n\mathbf{I}_n\right)^{-1}\mathbf{G}_{\ddot{X}}}{n\varepsilon_n}\right)\\
=&\mathrm{Tr}\left(\mathbf{G}_{\ddot{Z}}\left(\mathbf{G}_{\ddot{Z}} + \varepsilon_n n\mathbf{I}_n\right)^{-1}\mathbf{G}_{\ddot{X}}\left(\mathbf{G}_{\ddot{X}} + \varepsilon_n n\mathbf{I}_n\right)^{-1}\right)\\
=&\mathrm{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{R}_{\ddot{X}}\right),
\end{aligned}$$

$$
\begin{aligned}
&\operatorname{Tr}\left(\widehat{V}_{Y\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)*}\widehat{V}_{\ddot{Z}\ddot{X}}^{(n)}\right)\\
&=\operatorname{Tr}\left(\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1/2}\frac{\mathbf{\Phi H}_n\mathbf{\Psi}^T}{n}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Psi H}_n\mathbf{\Lambda}^T}{n}\left(\frac{\mathbf{\Lambda H}_n\mathbf{\Lambda}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Lambda H}_n\mathbf{\Phi}^T}{n}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1/2}\right)\\
&=\operatorname{Tr}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Psi}^T}{n}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Psi H}_n\mathbf{\Lambda}^T}{n}\left(\frac{\mathbf{\Lambda H}_n\mathbf{\Lambda}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Lambda H}_n\mathbf{\Phi}^T}{n}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\right)\\
&=\operatorname{Tr}\left(\left(\frac{\mathbf{G}_Y-\mathbf{G}_Y\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_Y\right)^{-1}\mathbf{G}_Y}{\varepsilon_n n}\right)\left(\frac{\mathbf{G}_{\ddot{Z}}-\mathbf{G}_{\ddot{Z}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{Z}}\right)^{-1}\mathbf{G}_{\ddot{Z}}}{\varepsilon_n n}\right)\left(\frac{\mathbf{G}_{\ddot{X}}-\mathbf{G}_{\ddot{X}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{X}}\right)^{-1}\mathbf{G}_{\ddot{X}}}{\varepsilon_n n}\right)\right)\\
&=\operatorname{Tr}\left(\left(\mathbf{G}_Y\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_Y\right)^{-1}\right)\left(\mathbf{G}_{\ddot{Z}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{Z}}\right)^{-1}\right)\left(\mathbf{G}_{\ddot{X}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{X}}\right)^{-1}\right)\right)\\
&=\operatorname{Tr}\left(\mathbf{R}_Y\mathbf{R}_{\ddot{Z}}\mathbf{R}_{\ddot{X}}\right),
\end{aligned}
$$

$$
\begin{aligned}
&\operatorname{Tr}\left(\widehat{V}_{\ddot{Z}\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)}\widehat{V}_{Y\ddot{X}}^{(n)}\right)\\
&=\operatorname{Tr}\left(\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1/2}\frac{\mathbf{\Phi H}_n\mathbf{\Lambda}^T}{n}\left(\frac{\mathbf{\Lambda H}_n\mathbf{\Lambda}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Lambda H}_n\mathbf{\Psi}^T}{n}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Psi H}_n\mathbf{\Phi}^T}{n}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1/2}\right)\\
&=\operatorname{Tr}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Lambda}^T}{n}\left(\frac{\mathbf{\Lambda H}_n\mathbf{\Lambda}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Lambda H}_n\mathbf{\Psi}^T}{n}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Psi H}_n\mathbf{\Phi}^T}{n}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\right)\\
&=\operatorname{Tr}\left(\left(\frac{\mathbf{G}_{\ddot{Z}}-\mathbf{G}_{\ddot{Z}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{Z}}\right)^{-1}\mathbf{G}_{\ddot{Z}}}{\varepsilon_n n}\right)\left(\frac{\mathbf{G}_Y-\mathbf{G}_Y\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_Y\right)^{-1}\mathbf{G}_Y}{\varepsilon_n n}\right)\left(\frac{\mathbf{G}_{\ddot{X}}-\mathbf{G}_{\ddot{X}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{X}}\right)^{-1}\mathbf{G}_{\ddot{X}}}{\varepsilon_n n}\right)\right)\\
&=\operatorname{Tr}\left(\left(\mathbf{G}_{\ddot{Z}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{Z}}\right)^{-1}\right)\left(\mathbf{G}_Y\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_Y\right)^{-1}\right)\left(\mathbf{G}_{\ddot{X}}\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_{\ddot{X}}\right)^{-1}\right)\right)\\
&=\operatorname{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\mathbf{R}_{\ddot{X}}\right),
\end{aligned}
$$

$$
\begin{aligned}
&\operatorname{Tr}\left(\widehat{V}_{Y\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)}\widehat{V}_{Y\ddot{X}}^{(n)}\right)\\
&=\operatorname{Tr}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Psi H}_n\mathbf{\Lambda}^T}{n}\left(\frac{\mathbf{\Lambda H}_n\mathbf{\Lambda}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Lambda H}_n\mathbf{\Psi}^T}{n}\left(\frac{\mathbf{\Psi H}_n\mathbf{\Psi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\frac{\mathbf{\Psi H}_n\mathbf{\Phi}^T}{n}\left(\frac{\mathbf{\Phi H}_n\mathbf{\Phi}^T}{n}+\varepsilon_n\mathbf{I}\right)^{-1}\right)\\
&=\operatorname{Tr}\left(\mathbf{R}_Y\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\mathbf{R}_{\ddot{X}}\right).
\end{aligned}
$$

Note: $\mathbf{I}_n-\left(\mathbf{G}+n\varepsilon_n\mathbf{I}_n\right)^{-1}\mathbf{G}=\varepsilon_n n\left(\mathbf{G}+\varepsilon_n n\mathbf{I}_n\right)^{-1}$.

Thus, we can calculate (A8) as

$$
\begin{aligned}
\hat{I}_n^{COND}&=\operatorname{Tr}\left(\widehat{V}_{\ddot{Z}\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}\ddot{X}}^{(n)}-\widehat{V}_{Y\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)*}\widehat{V}_{\ddot{Z}\ddot{X}}^{(n)}-\widehat{V}_{\ddot{Z}\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)}\widehat{V}_{Y\ddot{X}}^{(n)}+\widehat{V}_{Y\ddot{X}}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)*}\widehat{V}_{\ddot{Z}Y}^{(n)}\widehat{V}_{Y\ddot{X}}^{(n)}\right)\\
&=\operatorname{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{R}_{\ddot{X}}-\mathbf{R}_Y\mathbf{R}_{\ddot{Z}}\mathbf{R}_{\ddot{X}}-\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\mathbf{R}_{\ddot{X}}+\mathbf{R}_Y\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\mathbf{R}_{\ddot{X}}\right)\\
&=\operatorname{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{R}_{\ddot{X}}-\mathbf{R}_{\ddot{Z}}\mathbf{R}_{\ddot{X}}\mathbf{R}_Y-\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\mathbf{R}_{\ddot{X}}+\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\mathbf{R}_{\ddot{X}}\mathbf{R}_Y\right)\\
&=\operatorname{Tr}\left(\left(\mathbf{R}_{\ddot{Z}}-\mathbf{R}_{\ddot{Z}}\mathbf{R}_Y\right)\left(\mathbf{R}_{\ddot{X}}-\mathbf{R}_{\ddot{X}}\mathbf{R}_Y\right)\right)\\
&=\operatorname{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{S}\mathbf{R}_{\ddot{X}}\mathbf{S}\right)
\end{aligned}
$$

where $\mathbf{S}=\mathbf{I}_n-\mathbf{R}_Y$.

Proposition 1. Assuming that $k_{\ddot{\mathcal{X}}}$ and $k_{\ddot{\mathcal{Z}}}$ are radial kernels. If Y is a constant random variable, then the empirical estimations of the conditional dependence and dependence are equal. Then,

$$\hat{I}_n^{NOCCO}(X,Z)=\hat{I}_n^{COND}(X,Z|Y),$$

where the radial kernel is $k(x,y)=k(\|x-y\|)$.

Proof. Recall that

$$
\begin{aligned}
\hat{I}_n^{NOCCO}(X,Z)&=\operatorname{Tr}\left(\mathbf{R}_Z\mathbf{I}_n\mathbf{R}_X\mathbf{I}_n\right),\\
\hat{I}_n^{COND}(X,Z|Y)&=\operatorname{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{S}\mathbf{R}_{\ddot{X}}\mathbf{S}\right).
\end{aligned}
$$

The main differences are the weighting matrix $\mathbf{S}$ and the kernel-based matrix $\mathbf{R}_{\ddot{X}}$ with extended-variable Y. Note that Y takes a constant value, regardless of any event that occurs. Since radial kernel $k(\mathbf{x}_1,\mathbf{x}_2)=k(\|\mathbf{x}_1-\mathbf{x}_2\|)$ is employed, then we have

$$
\mathbf{K_Y}=\begin{bmatrix}k(0)&k(0)&\cdots&k(0)\\k(0)&k(0)&\cdots&k(0)\\\vdots&\vdots&\ddots&\vdots\\k(0)&k(0)&\cdots&k(0)\end{bmatrix}_{n\times n}
$$

and the centered Gram matrices $\mathbf{G_Y}=\mathbf{0}$. Thus, we obtain the first identity $\mathbf{S}=\mathbf{I}_n-\mathbf{R}_Y=\mathbf{I}_n-\mathbf{G}_Y\left(\varepsilon_n n\mathbf{I}_n+\mathbf{G}_Y\right)^{-1}=\mathbf{I}_n$. For the extended variable $\ddot{X}$, its kernel follows that

$$
\|\ddot{\mathbf{x}}_i-\ddot{\mathbf{x}}_j\|=\left\|\begin{bmatrix}\mathbf{x}_i-\mathbf{x}_j\\\mathbf{y}-\mathbf{y}\end{bmatrix}\right\|=\left\|\begin{bmatrix}\mathbf{x}_i-\mathbf{x}_j\\\mathbf{0}\end{bmatrix}\right\|=\|\mathbf{x}_i-\mathbf{x}_j\|,
$$

Thus, we obtain the second identity $\mathbf{K}_{\ddot{\mathbf{X}}} = \mathbf{K}_{\mathbf{X}}$ and $\mathbf{R}_{\ddot{X}} = \mathbf{R}_X$, where $\mathbf{R}_{\ddot{Z}} = \mathbf{R}_Z$ is similar. By combining these two identities, we have

$$\hat{I}_n^{COND}(X, Z|Y) \;=\; \mathrm{Tr}\left(\mathbf{R}_{\ddot{Z}}\mathbf{S}\mathbf{R}_{\ddot{X}}\mathbf{S}\right) \;=\; \mathrm{Tr}\left(\mathbf{R}_Z\mathbf{I}_n\mathbf{R}_X\mathbf{I}_n\right) \;=\; \hat{I}_n^{NOCCO}(X, Z).$$

Convergency. In the following, the convergency proof of COND $\widehat{V}_{XZ|Y}^{(n)}$ [38] is presented. Assume that V_{ZX}, V_{ZY}, and V_{YX} are Hilbert-Schmidt, and that the regularization constant ε_n satisfies $\varepsilon_n \to 0$ and $\varepsilon_n^3 n \to \infty$, then we have

$$\|\widehat{V}_{XZ|Y}^{(n)} - V_{XZ|Y}\|_{HS} \to 0 \quad (n \to \infty) \tag{A9}$$

in probability with rate $\varepsilon_n^{-\frac{3}{2}} n^{-\frac{1}{2}}$.

Proof. From the expressions in (A2) and (2), it is easy to obtain

$$\begin{aligned}
&\|\widehat{V}_{ZX|Y}^{(n)} - V_{ZX|Y}\|_{HS} \\
=&\|\widehat{V}_{ZX}^{(n)} - \widehat{V}_{ZY}^{(n)}\widehat{V}_{YX}^{(n)} - V_{ZX} + V_{ZY}V_{YX}\|_{HS} \\
=&\|\widehat{V}_{ZX}^{(n)} - V_{ZX} + V_{ZY}V_{YX} - V_{ZY}\widehat{V}_{YX}^{(n)} + V_{ZY}\widehat{V}_{YX}^{(n)} - \widehat{V}_{ZY}^{(n)}\widehat{V}_{YX}^{(n)}\|_{HS} \\
\leqslant&\|\widehat{V}_{ZX}^{(n)} - V_{ZX}\|_{HS} + \|V_{ZY}\|_{HS}\|V_{YX} - \widehat{V}_{YX}^{(n)}\|_{HS} + \|V_{ZY} - \widehat{V}_{ZY}^{(n)}\|_{HS}\|\widehat{V}_{YX}^{(n)}\|_{HS}
\end{aligned}$$

From the proof of Theorem 5 in [38], we have

$$\|\widehat{V}_{YX}^{(n)} - V_{YX}\|_{HS} = O_p(\varepsilon_n^{-3/2}n^{-1/2}) \quad \text{as} \quad n \to \infty.$$

Similarly, we have $\|\widehat{V}_{ZX}^{(n)} - V_{ZX}\|_{HS} = O_p(\varepsilon_n^{-3/2}n^{-1/2})$, $\|\widehat{V}_{ZY}^{(n)} - V_{ZY}\|_{HS} = O_p(\varepsilon_n^{-3/2}n^{-1/2})$ as $n \to \infty$. Since $\|V_{ZY}\|_{HS} \leqslant 1$ and $\|\widehat{V}_{YX}^{(n)}\|_{HS} \leqslant 1$, we have

$$\|\widehat{V}_{ZX|Y}^{(n)} - V_{ZX|Y}\|_{HS} = O_p(\varepsilon_n^{-3/2}n^{-1/2}).$$

Appendix A.3 Upper Bounds on the Target Error

Lemma A.1. For any hypothesis h, $h' \in \mathcal{H}$,

$$|\epsilon_S(h, h') - \epsilon_T(h, h')| \leqslant \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P_{X|Y}^S, P_{X|Y}^T)] + \|P_Y^S - P_Y^T\|_1.$$

Proof.

$$\begin{aligned}
|\epsilon_S(h, h') - \epsilon_T(h, h')| &\leqslant \sup_{h,h'\in\mathcal{H}} |\epsilon_S(h, h') - \epsilon_T(h, h')| \\
&= \sup_{h,h'\in\mathcal{H}} |\int_{\mathcal{Y}}\int_{\mathcal{X}} \mathbf{1}_{h(x)\neq h'(x)} p_{X|Y}^S p_Y^S dxdy - \int_{\mathcal{Y}}\int_{\mathcal{X}} \mathbf{1}_{h(x)\neq h'(x)} p_{X|Y}^T p_Y^T dxdy| \\
&= \sup_{h,h'\in\mathcal{H}} \int_{\mathcal{Y}}\int_{\mathcal{X}} \mathbf{1}_{h(x)\neq h'(x)} |p_{X|Y}^S p_Y^S - p_{X|Y}^T p_Y^S + p_{X|Y}^T p_Y^S - p_{X|Y}^T p_Y^T| dxdy \\
&= \sup_{h,h'\in\mathcal{H}} \int_{\mathcal{Y}}\int_{\mathcal{X}} \mathbf{1}_{h(x)\neq h'(x)} |(p_{X|Y}^S - p_{X|Y}^T)p_Y^S + p_{X|Y}^T(p_Y^S - p_Y^T)| dxdy \\
&\leqslant \int_{\mathcal{Y}} \sup_{h,h'\in\mathcal{H}} \int_{\mathcal{X}} \mathbf{1}_{h(x)\neq h'(x)} |p_{X|Y}^S - p_{X|Y}^T| p_Y^S dxdy + \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X|Y}^T |p_Y^S - p_Y^T| dxdy \\
&\leqslant \int_{\mathcal{Y}} \sup_{h,h'\in\mathcal{H}} \int_{\mathcal{X}} \mathbf{1}_{h(x)\neq h'(x)} |p_{X|Y}^S - p_{X|Y}^T| p_Y^S dxdy + \int_{\mathcal{Y}} |p_Y^S - p_Y^T| dxdy \\
&= \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P_{X|Y}^S, P_{X|Y}^T)] + \|P_Y^S - P_Y^T\|_1
\end{aligned}$$

Lemma A.2 (Lemma 6, [2])**.** For the domain $\mathcal{D}$, let S be a labeled sample set of size m drawn from $\mathcal{D}$ and labeling them according to the gourndtruth labeling function f. Let $\hat{\epsilon}(h)$ be the empirical error of some fixed hypothesis h on this sample set, and let $\epsilon(h)$ be the true error, then

$$P[|\hat{\epsilon}(h) - \epsilon(h)| \geqslant \varepsilon_n] \leqslant 2\exp(-2m{\varepsilon_n}^2)$$

Theorem 3. Let $\mathcal{H}$ be a hypothesis space of VC dimension d, m be the sample size of source domain and f_s be the ground truth labeling function for the source domain. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_S(h)$ and $h_T^* = \operatorname*{argmin}_{h\in\mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\epsilon_T(\hat{h}) \leqslant \epsilon_T(h^*) + 2(\lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P_{X|Y}^S, P_{X|Y}^T)] + \|P_Y^S - P_Y^T\|_1) + 2\eta_{d,m,\delta}, \tag{A10}$$

where $\eta_{d,m,\delta} = \sqrt{\frac{1}{2m}(\log\frac{d}{\delta})}$, $\lambda = \min_{h\in\mathcal{H}}\{\epsilon_S(h) + \epsilon_T(h)\}$.

Proof. Let $h^* = \operatorname{argmin}_{h\in\mathcal{H}} \epsilon_T(h) + \epsilon_S(h)$. Then,

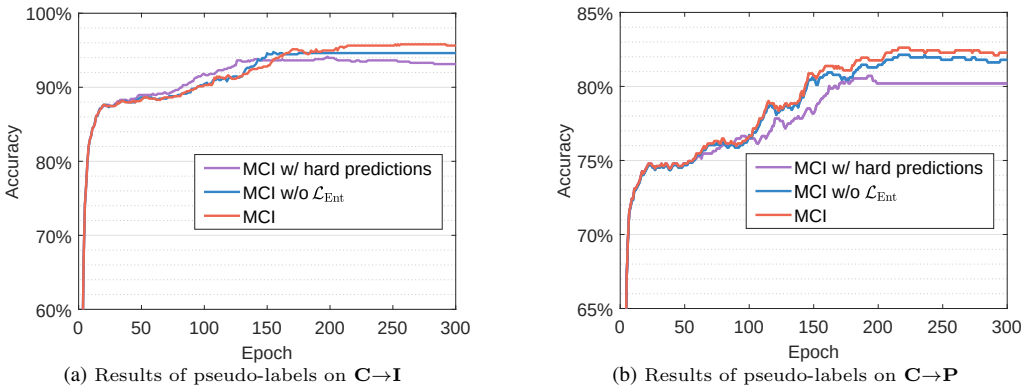
$$\begin{aligned}
|\epsilon_S(h) - \epsilon_T(h)| &= |\epsilon_S(h, f_s) - \epsilon_T(h, f_t)| \\
&= |\epsilon_S(h, f_s) - \epsilon_S(h, h^*) + \epsilon_S(h, h^*) - \epsilon_T(h, h^*) + \epsilon_T(h, h^*) - \epsilon_T(h, f_t)| \\
&\leqslant |\epsilon_S(h, f_s) - \epsilon_S(h, h^*)| + |\epsilon_S(h, h^*) - \epsilon_T(h, h^*)| + |\epsilon_T(h, h^*) - \epsilon_T(h, f_t)| \\
&\leqslant \epsilon_S(h^*) + \epsilon_T(h^*) + |\epsilon_S(h, h^*) - \epsilon_T(h, h^*)| \\
&\leqslant \lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P^S_{X|Y}, P^T_{X|Y})] + \|P^S_Y - P^T_Y\|_1. \qquad \text{(A11)}
\end{aligned}$$

The second inequality follows from the triangle inequality [40]. The last inequality follows from the definition of λ and Lemma A.1. Now applying Lemma A.2 on (A11), we have for any $\delta \in (0, 1)$, with probability $1 - \delta$,

$$\begin{aligned}
\epsilon_T(\hat{h}) &\leqslant \epsilon_S(\hat{h}) + \lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P^S_{X|Y}, P^T_{X|Y})] + \|P^S_Y - P^T_Y\|_1 \\
&\leqslant \hat{\epsilon}_S(\hat{h}) + \eta(d, m, \delta) + \lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P^S_{X|Y}, P^T_{X|Y})] + \|P^S_Y - P^T_Y\|_1 \\
&\leqslant \hat{\epsilon}_S(h^*_T) + \eta(d, m, \delta) + \lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P^S_{X|Y}, P^T_{X|Y})] + \|P^S_Y - P^T_Y\|_1 \\
&\leqslant \epsilon_S(h^*_T) + 2\eta(d, m, \delta) + \lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P^S_{X|Y}, P^T_{X|Y})] + \|P^S_Y - P^T_Y\|_1 \\
&\leqslant \epsilon_T(h^*) + 2\eta(d, m, \delta) + 2(\lambda + \frac{1}{2}\mathbb{E}_Y[d_{\mathcal{H}\Delta\mathcal{H}}(P^S_{X|Y}, P^T_{X|Y})] + \|P^S_Y - P^T_Y\|_1).
\end{aligned}$$

The first and the last inequalities follows from (A11). The second and the fourth inequalities follows from Lemma A.2. The third inequality follows the definition of $\hat{h}$.

Appendix B Experimental Results



(a) Results of pseudo-labels on **C→I** (b) Results of pseudo-labels on **C→P**

Figure B1 (Color online) The accuracy of pseudo-labels in the training process.

Table B1 Accuracies (%) on Office-31. CDTrans* and SHOT* uses DeiT-Base backbone. TVT° uses ViT-Base backbone. "-B" indicates ViT-Base backbone.

Backbone	Methods	A→W	D→W	W→D	A→D	D→A	W→A	Mean
AlexNet	Source-Only [48]	61.6	95.4	99.0	63.8	51.1	49.8	70.1
	DAN [4]	68.5	96.0	99.0	67.0	54.0	53.1	72.9
	DANN [7]	73.0	96.4	99.2	72.3	53.4	51.2	74.3
	CDAN+E [9]	78.3	97.2	**100.0**	76.3	57.3	57.3	77.7
	ATM [14]	**80.2**	97.9	**100.0**	**77.5**	62.1	61.3	**79.6**
	MCI	78.5	**98.6**	99.6	75.9	**62.5**	**62.2**	**79.6**
ViT	ViT-B [46]	91.2	99.2	**100.0**	90.4	81.1	80.6	90.4
	CDTrans* [26]	96.7	99.0	**100.0**	97.0	81.1	81.9	92.6
	SHOT* [24]	94.3	99.0	**100.0**	95.3	79.4	80.2	91.4
	TVT° [27]	96.4	**99.4**	**100.0**	96.4	84.9	**86.1**	93.8
	SSRT-B [18]	**97.7**	99.2	**100.0**	**98.6**	83.5	82.2	93.5
	MCI	94.7	**99.4**	**100.0**	97.9	**86.5**	85.6	**94.0**

Table B2 Accuracies (%) on Office-Home (ResNet-50).

Office-Home	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Mean
Source-Only [47]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [4]	43.6	5.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [7]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
KGOT [29]	36.2	59.4	65.0	48.6	56.5	60.2	52.1	37.8	67.1	59.0	41.9	72.0	54.7
ETIC [31]	49.3	71.3	76.3	58.6	70.5	71.0	59.7	47.4	77.3	66.0	50.4	79.6	64.8
CDAN+E [9]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
DSAN [13]	54.4	70.8	75.4	60.4	67.8	68.0	62.6	55.9	78.5	73.8	60.6	83.1	67.6
DMP [34]	52.3	73.0	77.3	64.3	72.0	71.8	63.6	52.7	78.5	72.0	57.7	81.6	68.1
ATM [14]	52.4	72.6	80.2	61.1	72.0	72.6	59.5	52.0	79.1	73.3	58.9	83.4	67.9
BuresNet [35]	54.7	74.4	77.1	63.7	72.2	71.8	64.1	51.7	78.4	73.1	58.0	82.4	68.5
WCEL [36]	56.3	75.1	77.6	60.5	68.6	70.2	61.2	49.6	76.6	68.7	59.8	83.6	67.3
SymmNets-V2 [22]	48.1	74.3	78.7	64.6	71.8	74.1	64.4	50.0	80.2	74.3	53.1	83.2	68.1
SHOT [24]	57.1	**78.1**	81.5	68.0	**78.2**	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
DCAN+SCDA [17]	**60.7**	76.4	**82.8**	**69.8**	77.5	**78.4**	**68.9**	**59.0**	**82.7**	**74.9**	**61.8**	**84.5**	**73.1**
MCI	51.7	76.3	80.1	60.6	75.2	76.3	64.8	51.4	81.7	69.3	54.8	83.3	68.8

Table B3 Accuracies (%) on Office-Home (ViT). CDTrans* and SHOT* uses DeiT-Base backbone. TVT° uses ViT-Base backbone. "-B" indicates ViT-Base backbone.

Office-Home	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Mean
ViT-B [46]	54.7	83.0	87.2	77.3	83.4	85.5	74.4	50.9	87.2	79.6	53.8	88.8	75.5
CDtrans* [26]	68.8	85.0	86.9	81.5	87.1	87.3	79.6	63.3	88.2	82.0	66.0	90.6	80.5
TVT° [27]	74.89	86.82	89.47	82.78	87.95	88.27	79.81	71.94	90.13	85.46	74.62	90.56	83.56
SSRT-B [18]	**75.2**	89.0	**91.1**	**85.1**	88.3	**90.0**	**85.0**	**74.2**	**91.3**	**85.7**	**78.6**	**91.8**	**85.4**
SHOT* [24]	67.1	83.5	85.5	76.6	83.4	83.7	76.3	65.3	85.3	80.4	66.7	83.4	78.1
MCI	71.2	**89.8**	89.5	82.4	**89.4**	89.2	81.0	68.0	90.3	83.2	75.6	90.9	83.4

Table B4 Accuracies (%) on VisDA-2017 (ResNet-101).

VisDA-2017	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Mean
Source-Only [47]	72.3	6.1	63.4	**91.7**	52.7	7.9	80.1	5.6	90.1	18.5	78.1	25.9	49.4
DAN [4]	68.1	15.4	76.5	87.0	71.1	48.9	82.3	51.5	88.7	33.2	88.9	42.2	62.8
DANN [7]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
ETIC [31]	49.3	71.3	76.3	58.6	70.5	71.0	59.7	47.4	77.3	66.0	50.4	79.6	64.8
CDAN [9]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
DSAN [13]	90.9	66.9	75.7	62.4	88.9	77.0	**93.7**	75.1	92.8	67.6	89.1	39.4	75.1
CTSN [21]	92.3	65.1	84.2	68.4	90.4	61.2	92.3	74.1	88.7	66.9	82.3	19.4	75.4
DMP [34]	92.1	75.0	78.9	75.5	91.2	81.9	89.0	77.2	93.3	77.4	84.8	35.1	79.3
ATM [14]	93.9	65.1	83.7	72.1	92.2	**92.5**	92.4	79.7	86.1	47.8	86.1	22.0	76.1
WCEL [36]	**95.2**	**82.5**	81.2	73.5	93.4	82.6	89.8	80.2	86.3	76.4	85.4	45.6	78.2
SUDA [19]	88.3	79.3	66.2	64.7	87.4	80.1	85.9	78.3	86.3	87.5	78.8	74.5	79.8
SymmNets-V2 [22]	87.3	62.2	79.1	66.7	80.3	79.7	87.8	75.6	88.9	31.4	**90.7**	25.8	71.3
SHOT [24]	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	**58.2**	**82.9**
MCI	94.1	66.4	**85.7**	69.2	93.2	**91.8**	93.6	78.9	90.1	**85.4**	87.7	26.8	80.3

Table B5 Accuracies (%) on DomainNet (ResNet-101). In each sub-table, the column-wise means source domain and the row-wise means target domain.

Source-Only [47]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	19.3	37.5	11.1	52.2	41.0	32.2
inf	30.2	-	31.2	3.6	44.0	27.9	27.4
pnt	39.6	18.7	-	4.9	54.5	36.3	30.8
qdr	7.0	0.9	1.4	-	4.1	8.3	4.3
rel	48.4	22.2	49.4	6.4	-	38.8	33.0
skt	46.9	15.4	37.0	10.9	47.0	-	31.4
Avg.	34.4	15.3	31.3	7.4	40.4	30.5	26.6

DANN [7]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	15.5	34.8	9.5	50.8	41.4	30.4
inf	31.8	-	30.2	3.8	44.8	25.7	27.3
pnt	39.6	15.1	-	5.5	54.6	35.1	30.0
qdr	11.8	2.0	4.4	-	9.8	8.4	7.3
rel	47.5	17.9	47.0	6.3	-	37.3	31.2
skt	47.9	13.9	34.5	10.4	46.8	-	30.7
Avg.	35.7	12.9	30.2	7.1	41.4	29.6	26.1

CDAN [9]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	20.4	36.6	9.0	50.7	42.3	31.8
inf	27.5	-	25.7	1.8	34.7	20.1	22.0
pnt	42.6	20.0	-	2.5	55.6	38.5	31.8
qdr	21.0	4.5	8.1	-	14.3	15.7	12.7
rel	51.9	23.3	50.4	5.4	-	41.4	34.5
skt	50.8	20.3	43.0	2.9	50.8	-	33.6
Avg.	38.8	17.7	32.8	4.3	41.2	31.6	27.7

DCAN [25]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	18.5	43.6	17.1	60.3	45.8	37.1
inf	39.7	-	38.4	5.9	54.6	28.5	33.4
pnt	48.6	19.7	-	9.9	61.7	41.2	36.2
qdr	33.2	5.6	16.1	-	18.4	16.2	17.9
rel	53.7	18.5	50.5	4.0	-	33.4	32.0
skt	57.6	17.3	47.3	10.1	55.3	-	37.5
Avg.	46.6	15.9	39.2	9.4	50.1	33.0	32.4

MDD+SCDA [17]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	20.4	43.3	15.2	59.3	46.5	36.9
inf	32.7	-	34.5	6.3	47.6	29.2	30.1
pnt	46.4	19.9	-	8.1	58.8	42.9	35.2
qdr	31.1	6.6	18.0	-	28.8	22.0	21.3
rel	55.5	23.7	52.9	9.5	-	45.2	37.4
skt	55.8	20.1	46.5	15.0	56.7	-	38.8
Avg.	44.3	18.1	39.0	10.8	50.2	37.2	33.3

MCI	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	19.2	43.9	18.9	63.9	47.2	38.6
inf	56.4	-	43.0	10.7	62.3	41.8	42.8
pnt	50.9	24.3	-	14.4	65.1	47.4	40.4
qdr	27.6	2.5	7.9	-	15.2	18.7	14.4
rel	57.5	26.4	55.8	22.2	-	47.8	41.9
skt	59.3	22.1	51.8	27.4	56.7	-	43.4
Avg.	50.3	18.9	40.5	18.7	52.7	40.6	**36.9**

Table B6 Computational time (s) and accuracy (%) of HSIC, ETIC, and MCI.

Method	Image-CELF task **C→I**			Office-31 task **W→A**		
	HSIC	ETIC	MCI	HSIC	ETIC	MCI
Time	0.0333	0.0027	0.0556	0.3333	0.0157	0.7000
Accuracy (%)	95.1	90.9	95.5	72.7	69.0	83.4

References

1 Pan S J, Yang Q. A Survey on Transfer Learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345-1359
2 Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains. Mach Learn, 2010, 79: 151-175
3 Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test. J Mach Learn Res, 2012, 13: 723-773.
4 Long M, Cao Y, Cao Z, et al. Transferable Representation Learning with Deep Adaptation Networks. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 3071-3085
5 Sun B, Feng J, Saeko K. Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2016. 2058-2065
6 Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, 2012. 2066-2073
7 Ganin Y, Ustinova E, Ajakan H, et al. Domain-Adversarial Training of Neural Networks. J Mach Learn Res, 2016, 17: 1-35
8 Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 7167-7176
9 Long M, Cao Z, Wang J, et al. Conditional adversarial domain adaptation. In: Proceedings of the International Conference on Neural Information Processing Systems, 2018. 1640-1650
10 Kang G, Jiang L, Wei Y, et al. Contrastive adaptation network for single-and multi-source domain adaptation. IEEE Trans Pattern Anal Mach Intell, 2020, 44: 1793-1804
11 Zhao H, Des Combes R T, Zhang K, et al. On Learning Invariant Representations for Domain Adaptation. In: Proceedings of the International Conference on Machine Learning, 2019. 7523-7532
12 Long M, Wang J, Ding G, et al. Transfer Feature Learning with Joint Distribution Adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, 2013. 2200-2207
13 Zhu Y, Zhuang F, Wang J, et al. Deep Subdomain Adaptation Network for Image Classification. IEEE Trans Neural Netw Learn Syst, 2021, 32: 1713-1722
14 Li J, Chen E, Ding Z, et al. Maximum Density Divergence for Domain Adaptation. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 3918–3930
15 Hu L, Kan M, Shan S, et al. Unsupervised Domain Adaptation With Hierarchical Gradient Synchronization. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 4043-4052
16 Le T, Nguyen T, Ho N, et al. LAMDA: Label Matching Deep Domain Adaptation. In: Proceedings of the International Conference on Machine Learning, 2021. 6043-6054
17 Li S, Xie M, Lv F, et al. Semantic concentration for domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, 2021, 9102-9111.
18 Sun T, Lu C, Zhang T, et al. Safe self-refinement for transformer-based domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 7191-7200.
19 Zhang J, Huang J, Tian Z, et al. Spectral unsupervised domain adaptation for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 9829-9840.
20 Saito K, Watanabe K, Ushiku Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3723-3732
21 Zuo L, Jing M, Li J, et al. Challenging tough samples in unsupervised domain adaptation. Pattern Recognit, 2021, 110: 107540
22 Zhang Y, Deng B, Tang H, et al. Unsupervised Multi-Class Domain Adaptation: Theory, Algorithms, and Practice. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 2775-2792
23 Zhang Y, Jing C, Lin H, et al. Hard class rectification for domain adaptation. Knowl Syst, 2021, 222: 107011
24 Liang J, Hu D, Feng J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: Proceedings of International Conference on Machine Learning, 2020. 6028-6039.
25 Li S, Xie B, Lin Q, et al. Generalized domain conditioned adaptation network. IEEE Trans Pattern Anal Mach Intell, 2021. 4093-4109.
26 Xu T, Chen W, Pichao W, et al. CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation. International Conference on Learning Representations. 2022.
27 Yang J, Liu J, Xu N, et al. Tvt: Transferable vision transformer for unsupervised domain adaptation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. 2023. 520-530.
28 Courty N, Flamary R, Tuia D. Optimal Transport for Domain Adaptation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1853-1865
29 Zhang Z, Wang M, Nehoral A. Optimal Transport in Reproducing Kernel Hilbert Spaces: Theory and Applications. IEEE Trans Pattern Anal Mach Intell, 2020, 42: 1741-1754
30 Yan K, Kou L, Zhang D. Learning Domain-Invariant Subspace Using Domain Features and Independence Maximization. IEEE Trans Cybern, 2018, 48: 288-299
31 Liu L, Pal S, Harchaoui Z. Entropy Regularized Optimal Transport Independence Criterion. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2022. 11247-11279
32 Deng Z, Luo Y, Zhu J. Cluster Alignment With a Teacher for Unsupervised Domain Adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 9944-9953
33 Chen C, Xie W, Huang W, et al. Progressive Feature Alignment for Unsupervised Domain Adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 627-636
34 Luo Y, Ren C, Dai D, et al. Unsupervised Domain Adaptation via Discriminative Manifold Propagation. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 1653-1669
35 Ren C, Luo Y, Dai D. BuresNet: Conditional Bures Metric for Transferable Representation Learning. IEEE Trans Pattern Anal Mach Intell, 2022.
36 Lu Y, Zhu Q, Zhang B, et al. Weighted Correlation Embedding Learning for Domain Adaptation. IEEE Transactions on Image Processing, 2022, 31: 5303–5316
37 Beker C R. Joint measures and cross-covariance operators. Trans Am Math Soc, 1973, 186: 273-289
38 Fukumizu K, Gretton A, Sun X, et al. Kernel Measures of Conditional Dependence. In: Proceedings of the International Conference on Neural Information Processing Systems, 2008, 489-496
39 Fine S, Scheinberg K. Efficient SVM training using low-rank kernel representations. J Mach Learn Res, 2001, 2: 243-264

40 Koby C, Michael K, Jennifer W. Learning from Multiple Sources. J Mach Learn Res, 2008, 9: 1757-1774
41 Caputo B, Müller H, Martinez-Gomez J, et al. ImageCLEF 2014: Overview and analysis of the results. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, 2014. 192-211
42 Saenko K, Kulis B, Fritz M, et al. Adapting visual category models to new domains. In: Proceedings of the European Conference on Computer Vision, 2010. 213-226
43 Venkateswara H, Eusebio J, Chakraborty S, et al. Deep Hashing Network for Unsupervised Domain Adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5018-5027
44 Peng X, Usman B, Kaushik N, et al. Visda: The visual domain adaptation challenge. 2017. ArXiv: 1710.06924
45 Peng X, Bai Q, Xia X, et al. Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, 2019, 1406C1415.
46 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020ArXiv: 2010.11929.
47 He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770-778
48 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the International Conference on Neural Information Processing Systems, 2012. 1097-1105
49 Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248-255.
50 Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008. 2579-2605
51 Bach F R, Jordan M I. Kernel independent component analysis. J Mach Learn Res, 2002, 3: 1-48