Appendix of Multi-domain Discriminant Analysis

Shoubo Hu[*], **Kun Zhang**[†], **Zhitang Chen**[‡], **Laiwan Chan**[*]
[*]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[†]Department of Philosophy, Carnegie Mellon University
[‡]Huawei Noah's Ark Lab

Overview

A Quantities' Property Illustration

The illustrations comparing average domain discrepancy with multi-domain within-class scatter, and average class discrepancy with multi-domain between-class scatter are given in Figure 1 and 2.

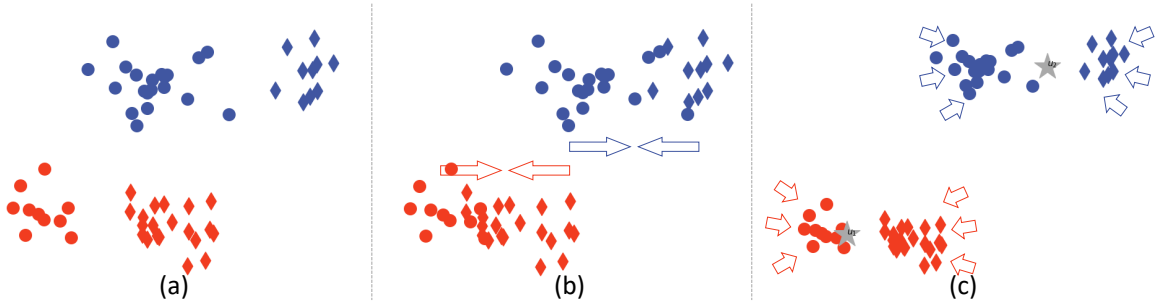


Figure 1: Comparison Between Average Domain Discrepancy and Multi-domain Within-class Scatter. Colors denote classes and markers denote domains. (a) The distribution of data in the subspace $\mathbb{R}^q$ transformed from RKHS $\mathcal{H}$ using $\mathbf{W}^0$. (b) By minimizing average domain discrepancy, the resulting transformation $\mathbf{W}^{add}$ makes the means within each class closer. (c) By minimizing multi-domain within-class scatter, the resulting transformation $\mathbf{W}^{mws}$ makes distribution of each class more compact towards the corresponding mean representation.

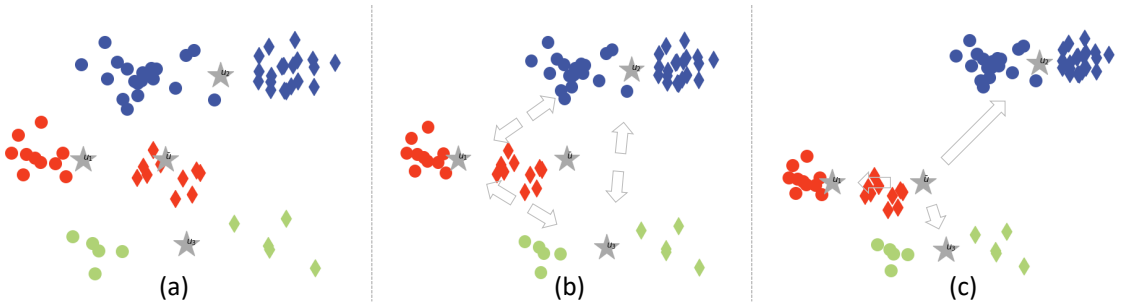


Figure 2: Comparison Between Average Class Discrepancy and Multi-domain Between-class Scatter. Colors denote classes and markers denote domains. (a) The distribution of data in the subspace $\mathbb{R}^q$ transformed from RKHS $\mathcal{H}$ using $\mathbf{W}^0$. (b) By maximizing average class discrepancy, the resulting transformation $\mathbf{W}^{acd}$ treats the distances between each pair of mean representations equally and maximizes them; (c) By maximizing multi-domain between-class scatter, the resulting transformation $\mathbf{W}^{mbs}$ maximizes the average distance between the overall mean and the mean representation of different classes. However, each distance is added a weight, which is proportional to the number of instances in the corresponding class. As a result, it is approximate equivalent to the scheme where one pools data of different domains of the same class together and trains classifier.

B Derivation of the Lagrangian

Since the objective

$$\arg\max_{\mathbf{B}} = \frac{\operatorname{tr}\left(\mathbf{B}^T\left(\beta\mathbf{F}+(1-\beta)\mathbf{P}\right)\mathbf{B}\right)}{\operatorname{tr}\left(\mathbf{B}^T(\gamma\mathbf{G}+\alpha\mathbf{Q}+\mathbf{K})\mathbf{B}\right)} \tag{1}$$

is invariant to re-scaling $\mathbf{B} \to \delta\mathbf{B}$, we rewrite (1) as a constrained optimization problem:

$$\arg\max_{\mathbf{B}} \ \operatorname{tr}\left(\mathbf{B}^T\left(\beta\mathbf{F}+(1-\beta)\mathbf{P}\right)\mathbf{B}\right) \tag{2}$$

$$s.t. \ \operatorname{tr}\left(\mathbf{B}^T(\gamma\mathbf{G}+\alpha\mathbf{Q}+\mathbf{K})\mathbf{B}\right) = 1, \tag{3}$$

which yields the Lagrangian

$$\begin{aligned}\mathcal{L} = &\operatorname{tr}\left(\mathbf{B}^T\left(\beta\mathbf{F}+(1-\beta)\mathbf{P}\right)\mathbf{B}\right)\\ &-\operatorname{tr}\left(\left(\mathbf{B}^T(\gamma\mathbf{G}+\alpha\mathbf{Q}+\mathbf{K})\mathbf{B}-\mathbf{I}_q\right)\mathbf{\Gamma}\right),\end{aligned} \tag{4}$$

where $\mathbf{\Gamma}$ is a diagonal matrix containing the Lagrange multipliers and $\mathbf{I}_q$ denotes the identity matrix of dimension q. Setting the derivative with respect to $\mathbf{B}$ in the Lagrangian (4) to zero yields the following generalized eigenvalue problem:

$$\left(\beta\mathbf{F}+(1-\beta)\mathbf{P}\right)\mathbf{B} = \left(\gamma\mathbf{G}+\alpha\mathbf{Q}+\mathbf{K}\right)\mathbf{B}\mathbf{\Gamma}. \tag{5}$$

C Proof of Theorem 2

Theorem 1. *Under assumptions 2 – 4, and further assuming that $\|\hat{f}\|_{\mathcal{H}_{\tilde{k}}} \leq 1$ and $\|f^*\|_{\mathcal{H}_{\tilde{k}}} \leq 1$, where $\hat{f}$ denotes the empirical risk minimizer, f^* denotes the expected risk minimizer, then with probability at least $1-\delta$ there is*

$$\begin{aligned}&\mathbb{E}[\ell(\hat{f}(\tilde{X}^t\mathbf{W}), Y^t)] - \mathbb{E}[\ell(f^*(\tilde{X}^t\mathbf{W}), Y^t)]\\ &\leq 4L_\ell L_{k_\gamma} U_{k'_x} U_{k_x}\sqrt{\frac{\operatorname{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})}{n}} + \sqrt{\frac{2\log 2\delta^{-1}}{n}},\end{aligned} \tag{6}$$

where the expectations are taken over the joint distribution of the test domain $\mathbb{P}^t(X^t, Y^t)$, n is the number of training samples, and $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$.

Proof. First, we use the following result.

Theorem 2 (Generalization bound based on Rademacher complexity)**.** *Define $\mathcal{A} = \{x \mapsto \ell(f(x), y) : f \in \mathcal{H}\}$ to be the loss class, the composition of the loss function with each of the hypotheses. With probability at least $1 - \delta$:*

$$L(\hat{f}) - L(f^*) \leq 4\mathcal{R}_n(\mathcal{A}) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}, \tag{7}$$

where $L(\hat{f})$ denotes the expected test risk of the empirical risk minimizer, $L(\hat{f})$ denotes the expected test risk of the expected risk minimizer, $\mathcal{R}_n(\mathcal{A})$ denotes the Rademacher complexity of loss class $\mathcal{A}$, and n denotes the number of training points.

By applying theorem 2, with probability at least $1 - \delta$ there is

$$\mathbb{E}_{\mathbb{P}_X^t}[\ell(\hat{f}(\tilde{X}^t\mathbf{W}), Y^t)] - \mathbb{E}_{\mathbb{P}_X^t}[\ell(f^*(\tilde{X}^t\mathbf{W}), Y^t)] \leq 4\mathcal{R}_n(\mathcal{A}) + \sqrt{\frac{2\log 2\delta^{-1}}{n}}, \tag{8}$$

where $\mathcal{A}$ denotes the loss class $\{x \mapsto \ell(f(\mathbb{P}, x), y) : \|f\|_{\mathcal{H}_{\bar{k}}} \leq 1\}$, $\mathcal{R}_n(\cdot)$ denotes the Rademacher complexity and n is the number of training points.

Since the loss function ℓ is L_ℓ-Lipschitz in its first variable, there is

$$\mathcal{R}_n(\mathcal{A}) = \mathcal{R}_n(\ell \circ f) \leq L_\ell \mathcal{R}_n(\mathcal{H}_{\bar{k}}). \tag{9}$$

To obtain the Rademacher complexity of $\mathcal{H}_{\bar{k}}$, i.e. $\mathcal{R}_n(\mathcal{H}_{\bar{k}})$, we adopt the following theorem.

Theorem 3 (Rademacher complexity of L_2 ball)**.** *Let $\mathcal{F} = \{z \mapsto \langle w, z\rangle : \|w\|_2 \leq B_2\}$ (bound on weight vectors). Assume $\mathbb{E}_{Z\sim p^*}\left[\|Z\|_2^2\right] \leq C_2^2$ (bound on spread of data points). Then*

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{B_2 C_2}{\sqrt{n}}, \tag{10}$$

where n denotes the number of training points.

According to the function class we restricted, B_2 in theorem 3 in our case is 1. For the bound of feature maps of data in $\mathcal{H}_{\bar{k}}$ (corresponds to C_2), there is

$$\begin{aligned}
&\left\|\bar{k}\left(\tilde{X}^t\mathbf{W}, \cdot\right)\right\| && (11)\\
&= \|\gamma_{k_\gamma}\left(\gamma(\mathbb{P}^t)\right) \otimes k_X(X^t, \cdot)\mathbf{W}\| && (12)\\
&\leq L_{k_\gamma}\|\gamma(\mathbb{P}^t)\|\|k_X(X^t, \cdot)\mathbf{W}\| && (13)\\
&\leq L_{k_\gamma} U_{k'} U_k \|\mathbf{W}\|_{HS}. && (14)
\end{aligned}$$

Note that $\mathbf{W} = \mathbf{\Phi}^T\mathbf{B}$ and $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$ is invertible. It follows that $\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})$ defines a norm consistent with the Hilbert-Schmidt norm $\|\mathbf{W}\|_{HS}$. Therefore, by applying theorem 3, there is

$$\mathcal{R}_n(\mathcal{A}) \leq L_\ell L_{k_\gamma} U_{k'} U_k \sqrt{\frac{\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})}{n}}. \tag{15}$$

Combining it with (8) gives the results. □

D Proof of Theorem 3

Theorem 4. *Under assumptions 2 – 4, and assuming that all source sample sets are of the same size, i.e.* $n^s = \bar{n}$ *for* $s = 1, \ldots, m$*, then with probability at least* $1 - \delta$ *there is*

$$\begin{aligned}
&\sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1} \left| \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \ell \left(f(\hat{\tilde{X}}_i^s \mathbf{W}), y_i^s \right) - \mathcal{E}(f, \infty) \right| \\
\leq & U_\ell \left(\left(\frac{\log 2\delta^{-1}}{2m\bar{n}} \right)^{\frac{1}{2}} + \left(\frac{\log \delta^{-1}}{2m} \right)^{\frac{1}{2}} \right) + \sqrt{\mathrm{tr}(\mathbf{B}^T \mathbf{K} \mathbf{B})} \left(c_1 \left(\frac{\log 2\delta^{-1} m}{\bar{n}} \right)^{\frac{1}{2}} + c_2 \left(\left(\frac{1}{m\bar{n}} \right)^{\frac{1}{2}} + \left(\frac{1}{m} \right)^{\frac{1}{2}} \right) \right)
\end{aligned} \tag{16}$$

where $c_1 = 2\sqrt{2} L_\ell U_{k_x} L_{k_\gamma} U_{k'_x}$, $c_2 = 2 L_\ell U_{k_x} U_{k_\gamma}$.

Proof. Follow the idea in Blanchard et al. [2011], the supremum of the generalization error bound can be decomposed as

$$\begin{aligned}
& \sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1} \left| \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \ell \left(f(\hat{\tilde{X}}_i^s \mathbf{W}), y_i^s \right) - \mathcal{E}(f, \infty) \right| \\
\leq & \sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1} \left| \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \left(\ell \left(f(\hat{\tilde{X}}_i^s \mathbf{W}), y_i^s \right) - \ell \left(f(\tilde{X}_i^s \mathbf{W}), y_i^s \right) \right) \right| && (17) \\
& + \sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1} \left| \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \ell \left(f(\tilde{X}_i^s \mathbf{W}), y_i^s \right) - \mathcal{E}(f, \infty) \right| && (18) \\
:= & (I) + (II), && (19)
\end{aligned}$$

where $\hat{\tilde{X}}_i^s = (\hat{\mathbb{P}}^s, x_i^s)$, $\tilde{X}_i^s = (\mathbb{P}^s, x_i^s)$.

Bound of term (I)

According to the assumption that the loss ℓ is L_ℓ -Lipschitz in its first variable, we have

$$\begin{aligned}
(I) \leq & L_\ell \sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1} \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \left| f(\hat{\tilde{X}}_i^s \mathbf{W}) - f(\tilde{X}_i^s \mathbf{W}) \right| && (20) \\
\leq & L_\ell \sup_{\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1} \frac{1}{m} \sum_{s=1}^{m} \left\| f \left((\hat{\mathbb{P}}^s, \cdot) \mathbf{W} \right) - f \left((\mathbb{P}^s, \cdot) \mathbf{W} \right) \right\|_\infty && (21)
\end{aligned}$$

For any $x \in \mathcal{X}$ and $\|f\|_{\mathcal{H}_{\bar{k}}} \leq 1$, using the reproducing property of the kernel $\bar{k}$ and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left| f \left((\hat{\mathbb{P}}^s, x) \mathbf{W} \right) - f \left((\mathbb{P}^s, x) \mathbf{W} \right) \right| &= \left| \left\langle \bar{k} \left((\hat{\mathbb{P}}^s, x) \mathbf{W}, \cdot \right) - \bar{k} \left((\mathbb{P}^s, x) \mathbf{W}, \cdot \right), f \right\rangle \right| && (22) \\
&\leq \|f\| \left\| \bar{k} \left((\hat{\mathbb{P}}^s, x) \mathbf{W}, \cdot \right) - \bar{k} \left((\mathbb{P}^s, x) \mathbf{W}, \cdot \right) \right\| && (23)
\end{aligned}$$

According to the assumption, there is $\|f\| \leq 1$. For the second term in (23) we have

$$\left\|\bar{k}\left((\hat{\mathbb{P}}^s, x)\mathbf{W}, \cdot\right) - \bar{k}\left((\mathbb{P}^s, x)\mathbf{W}, \cdot\right)\right\| \tag{24}$$

$$= \left\|\gamma_{k_\gamma}\left(\gamma(\hat{\mathbb{P}}^s)\right) \otimes k_X(x, \cdot)\mathbf{W} - \gamma_{k_\gamma}\left(\gamma(\mathbb{P}^s)\right) \otimes k_X(x, \cdot)\mathbf{W}\right\| \tag{25}$$

$$\leq \left\|\gamma_{k_\gamma}\left(\gamma(\hat{\mathbb{P}}^s)\right) \otimes k_X(x, \cdot) - \gamma_{k_\gamma}\left(\gamma(\mathbb{P}^s)\right) \otimes k_X(x, \cdot)\right\| \|\mathbf{W}\|_{HS} \tag{26}$$

$$= \|\mathbf{W}\|_{HS} \left(\left\langle \bar{k}((\hat{\mathbb{P}}^s, x), \cdot) - \bar{k}((\mathbb{P}^s, x), \cdot), \bar{k}((\hat{\mathbb{P}}^s, x), \cdot) - \bar{k}((\mathbb{P}^s, x), \cdot)\right\rangle\right)^{\frac{1}{2}} \tag{27}$$

$$\leq \|\mathbf{W}\|_{HS}\, k(x, x)^{\frac{1}{2}} \left(k_\gamma(\gamma(\mathbb{P}^s), \gamma(\mathbb{P}^s)) + k_\gamma(\gamma(\hat{\mathbb{P}}^s), \gamma(\hat{\mathbb{P}}^s)) - 2k_\gamma(\gamma(\mathbb{P}^s), \gamma(\hat{\mathbb{P}}^s))\right)^{\frac{1}{2}} \tag{28}$$

$$\leq U_k \|\mathbf{W}\|_{HS} \left\|\gamma_{k_\gamma}\left(\gamma(\mathbb{P}^s)\right) - \gamma_{k_\gamma}\left(\gamma(\hat{\mathbb{P}}^s)\right)\right\| \tag{29}$$

$$\leq U_k L_{k_\gamma} \|\mathbf{W}\|_{HS} \left\|\gamma(\hat{\mathbb{P}}^s) - \gamma(\mathbb{P}^s)\right\|. \tag{30}$$

Combining (23), (30) and $\|f\| \leq 1$, there is

$$\left|f\left((\hat{\mathbb{P}}^s, x)\mathbf{W}\right) - f\left((\mathbb{P}^s, x)\mathbf{W}\right)\right| \leq U_k L_{k_\gamma} \|\mathbf{W}\|_{HS} \left\|\gamma(\hat{\mathbb{P}}^s) - \gamma(\mathbb{P}^s)\right\|. \tag{31}$$

Now we derive the bound on $\left\|\gamma(\hat{\mathbb{P}}^s) - \gamma(\mathbb{P}^s)\right\|$. For independent real zero-mean random variables $x_1, \ldots, x_n$ such that $|x_i| \leq C$ for $i = 1, \ldots, n$, Hoeffding's inequality [Hoeffding, 1963] states that $\forall \epsilon > 0$:

$$\mathrm{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| > \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2C^2}\right). \tag{32}$$

Set the $\delta = 2\exp\left(-\frac{n\epsilon^2}{2C^2}\right)$, then with probability at least $1 - \delta$:

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| < \sqrt{2}C\sqrt{\frac{\log 2\delta^{-1}}{n}}. \tag{33}$$

Similar result holds for zero-mean independent random variables $\phi(x_1), \ldots, \phi(x_n)$ with values in a separable complex Hilbert space and such that $\|\phi(x_i)\| \leq C$, for $i = 1, \ldots, n$ [Rosasco et al., 2010]:

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\right\| < \sqrt{2}C\sqrt{\frac{\log 2\delta^{-1}}{n}}. \tag{34}$$

For independent uncentered variables $\phi'(x_i)$ with mean M, bounded by C. Let $\phi(x_i) = \phi'(x_i) - M$ denote the re-centered variables, now bounded at worst by $2C$ by the triangle inequality. Set $\delta = 2\exp\left(-\frac{n\epsilon^2}{8C^2}\right)$, we obtain with probability at least $1 - \delta$ that:

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \phi'(x_i) - M\right\| < 2\sqrt{2}C\sqrt{\frac{\log 2\delta^{-1}}{n}} \tag{35}$$

Based on the result of (35), we have

$$\left\|\gamma(\hat{\mathbb{P}}^s) - \gamma(\mathbb{P}^s)\right\| = \left\|\frac{1}{n^s}\sum_{i=1}^{n} \phi'(x_i^s) - \mathbb{E}_{X\sim\mathbb{P}^s}[\phi'(X)]\right\| \leq 3U_{k'}\sqrt{\frac{\log 2\delta^{-1}}{\bar{n}}} \tag{36}$$

Combining (31) and (36) we have

$$\sup_{\|f\|_{\mathcal{H}_{\bar{k}}}\leq 1} \left\| f\left((\hat{\mathbb{P}}^s,\cdot)\mathbf{W}\right) - f\left((\mathbb{P}^s,\cdot)\mathbf{W}\right)\right\|_\infty \leq 2\sqrt{2}U_k L_{k_\gamma} U_{k'} \|\mathbf{W}\|_{HS} \sqrt{\frac{\log 2\delta^{-1}}{\bar{n}}} \tag{37}$$

Conditionally to the draw of $\{\mathbb{P}^s\}_{1\leq s\leq m}$, we can apply (37) to each $(\mathbb{P}^s, \hat{\mathbb{P}}^s)$ the the union bound over $s = 1, \ldots, m$ to get that with probability at least $1-\delta$:

$$(I) \leq 2\sqrt{2}L_\ell U_k L_{k_\gamma} U_{k'} \|\mathbf{W}\|_{HS} \sqrt{\frac{\log 2\delta^{-1} + \log m}{\bar{n}}} \tag{38}$$

Bound of term (II)

This section follows the idea of Blanchard et al. [2011] so steps of proof that are largely unchanged are omitted. First, we define the conditional (idealized) test error for a given test distribution $\mathbb{P}^t_{XY}$ as

$$\mathcal{E}(f,\infty|\mathbb{P}^t_{XY}) := \mathbb{E}_{(X^t,Y^t)\sim\mathbb{P}^t_{XY}}\left[\ell\left(f(\tilde{X}^t\mathbf{W}), Y^t\right)\right], \tag{39}$$

where $\tilde{X}^t = (P^t_X, X^t)$.

Then (II) is further decomposed as

$$\begin{aligned}(II) \leq & \frac{1}{m}\sum_{s=1}^m \frac{1}{n^s}\sum_{i=1}^{n^s}\left(\ell\left(f(\tilde{X}^s_i\mathbf{W}), y^s_i\right) - \mathcal{E}(f,\infty|\mathbb{P}^s_{XY})\right) + \frac{1}{m}\sum_{s=1}^m\left(\mathcal{E}(f,\infty|\mathbb{P}^s_{XY}) - \mathcal{E}(f,\infty)\right) && (40)\\ := & (IIa) + (IIb) && (41)\end{aligned}$$

Bound of term (IIa)

In the case where conditioning on $\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}$, the observations in $\mathcal{D} = \{(x^s_i, y^s_i)\}^{m,n^s}_{s=1,i=1}$ are now independent (but not identically distributed) for this conditional distribution. We can thus apply the McDiarmid inequality [McDiarmid, 1989] to the function

$$\zeta(\mathcal{D}) := \sup_{\|f\|_{\mathcal{H}_{\bar{k}}}\leq 1} \frac{1}{m}\sum_{s=1}^m \frac{1}{n^s}\sum_{i=1}^{n^s}\left(\ell\left(f(\tilde{X}^s_i\mathbf{W}), y^s_i\right) - \mathcal{E}(f,\infty|\mathbb{P}^s_{XY})\right). \tag{42}$$

When $n^s = n^{s'} = \bar{n}$ for all s, s', that with probability $1-\delta$ over the draw of $\mathcal{D}$, it holds

$$|\zeta - \mathbb{E}\left[\zeta|\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}\right]| \leq U_l\sqrt{\frac{\log 2\delta^{-1}}{2m\bar{n}}}. \tag{43}$$

Then by the standard symmetrization technique, $\mathbb{E}\left[\zeta|\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}\right]$ can be bounded via Rademacher complexity as:

$$\begin{aligned}\mathbb{E}\left[\zeta|\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}\right] \leq & \frac{2}{m}\mathbb{E}_{(x^s_i,y^s_i)}\mathbb{E}_{(\epsilon^s_i)}\left[\sup_{\|f\|_{\mathcal{H}_{\bar{k}}}\leq 1}\sum_{s=1}^m\frac{1}{n^s}\sum_{i=1}^{n^s}\epsilon^s_i\left(\ell\left(f(\tilde{X}^s_i\mathbf{W}), y^s_i\right)\right)|\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}\right] && (44)\\ \leq & 2L_\ell U_k U_{k_\gamma}\|\mathbf{W}\|_{HS}\sqrt{\frac{1}{m\bar{n}}}, && (45)\end{aligned}$$

where the last inequality is from the bound of the Rademacher complexity of the loss class $\ell \circ f$.

Bound of term (IIb)

Since the $\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}$ are i.i.d., the McDiarmid inequality can be applied to the function

$$\xi\left(\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}\right) := \sup_{\|f\|_{\mathcal{H}_{\bar{k}}}\leq 1}\frac{1}{m}\sum_{s=1}^m\left(\mathcal{E}(f,\infty|\mathbb{P}^s_{XY}) - \mathcal{E}(f,\infty)\right), \tag{46}$$

then one obtains that with probability $1-\delta$ over the draw of $\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}$, it holds

$$|\xi - \mathbb{E}[\xi]| \leq U_\ell \sqrt{\frac{\log \delta^{-1}}{2m}}. \tag{47}$$

Similarly, by the standard symmetrization technique, $\mathbb{E}[\xi]$ is bounded as

$$\mathbb{E}[\xi] \leq \frac{2}{m} \mathbb{E}_{\{\mathbb{P}^s_{XY}\}_{1\leq s\leq m}} \mathbb{E}_{(X^s,Y^s)_{1\leq s\leq m}} \mathbb{E}_{(\epsilon^s)_{1\leq s\leq m}} \left[\sup_{\|f\|_{\mathcal{H}_{\tilde{k}}}\leq 1} \sum_{s=1}^{m} \epsilon^s \ell\left(f(\tilde{X}^s\mathbf{W}), Y^s\right)\right] \tag{48}$$

$$\leq 2L_\ell U_k U_{k_\gamma} \|\mathbf{W}\|_{HS} \sqrt{\frac{1}{m}}, \tag{49}$$

where the last inequality is again from the bound of the Rademacher complexity of the loss class $\ell \circ f$.

Finally, $\mathbf{W} = \mathbf{\Phi}^T\mathbf{B}$ and $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$ is invertible. It follows that $\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})$ defines a norm consistent with the Hilbert-Schmidt norm $\|\mathbf{W}\|_{HS}$. By combining the above results we obtain the announced result. □

E Experimental Configurations

Due to the difference in techniques adopted in different methods, there is/are different hyper-parameter(s) in each method require tuning in the experiments.

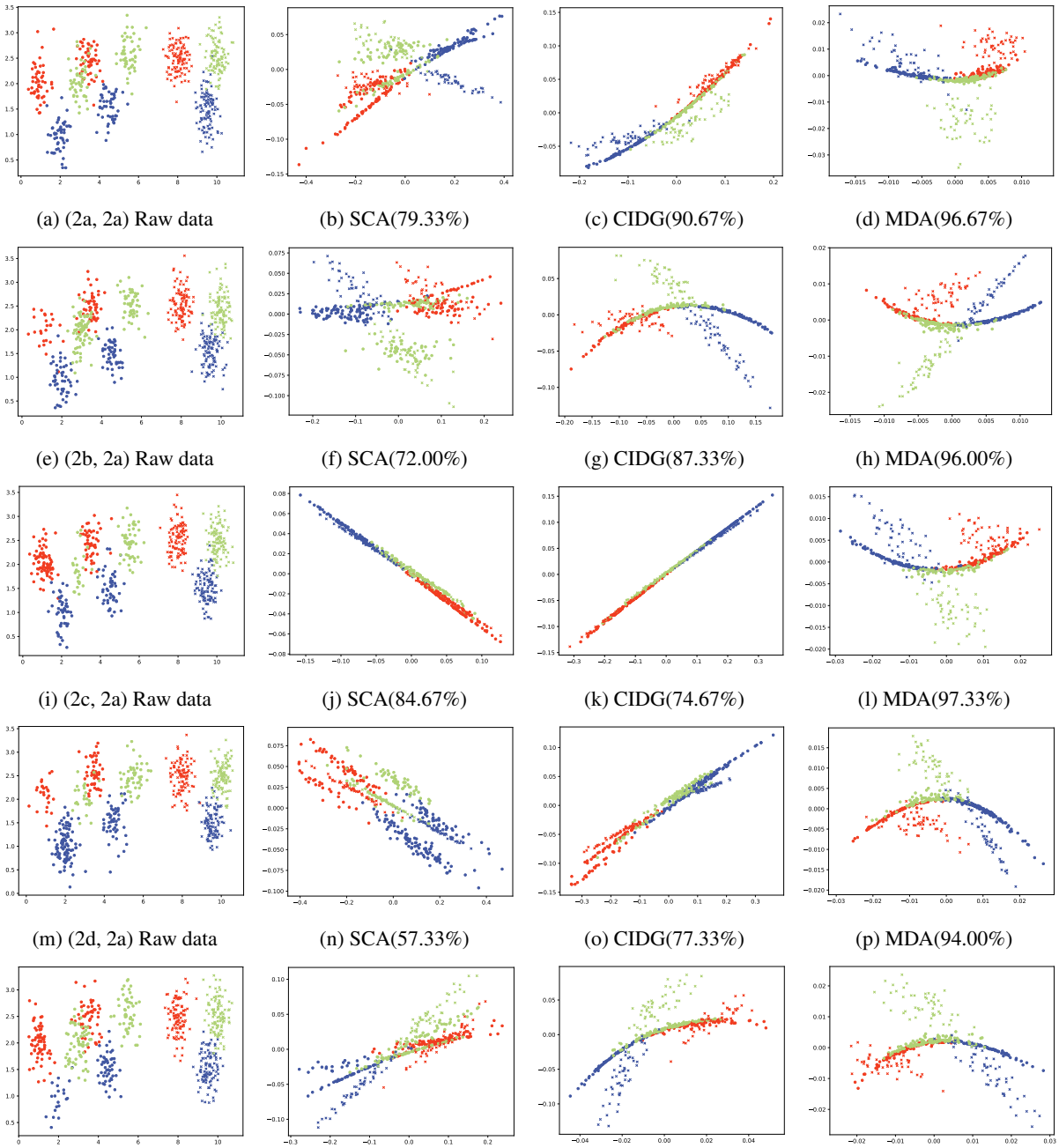
- 1NN: since there is no hyper-parameter to be determined in 1NN, instances in source domains are directly combined for training. Then we apply the trained model on target domains and report the test accuracy.
- SVM: the regularization coefficient C requires tuning in SVM. $C \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$ are validated in the experiments.
- KPCA and KFD: the kernel width σ_k requires tuning. $\sigma_k \in \{0.1d_M, 0.2d_M, 0.5d_M, d_M, 2d_M, 5d_M\}$, where $d_M = \text{median}\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right), \forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$, are validated.
- E-SVM: four hyper-parameters $(\lambda_1, \lambda_2, C_1, C_2)$ require tuning. $\lambda_1 \in \{0.1, 1, 10\}$, $\lambda_2 \in \{0.5\lambda_1, 1\lambda_1, 2\lambda_1\}$, and $C_1, C_2 \in \{0.1, 1, 10\}$ are validated.
- CCSA: two hyper-parameters (lr, α) require tuning. learning rate $lr \in \{0.5, 1.0, 1.5\}$ and $\alpha \in \{0.1, 0.25, 0.4\}$ are validated.
- DICA: Two parameters (λ, ϵ) require tuning. $\lambda \in \{1\text{e}{-3}, 1\text{e}{-2}, 1\text{e}{-1}, 1.0, 1\text{e}1, 1\text{e}2, 1\text{e}3\}$ and $\epsilon \in \{1\text{e}{-3}, 1\text{e}{-2}, 1\text{e}{-1}, 1.0, 1\text{e}1, 1\text{e}2, 1\text{e}3\}$ were validated.
- SCA: Two parameters (β, δ) require tuning. $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\delta \in \{1\text{e}{-3}, 1\text{e}{-2}, 1\text{e}{-1}, 1.0, 1\text{e}1, 1\text{e}2, 1\text{e}3, 1\text{e}4, 1\text{e}5, 1\text{e}6\}$ were validated.
- CIDG: Three hyper-parameters (β, α, γ) require tuning. $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\gamma \in \{1\text{e}{-3}, 1\text{e}{-2}, 1\text{e}{-1}, 1, 1\text{e}1, 1\text{e}2, 1\text{e}3, 1\text{e}4, 1\text{e}5, 1\text{e}6\}$, and $\alpha \in \{1, 1\text{e}1, 1\text{e}2, 1\text{e}3, 1\text{e}4, 1\text{e}5, 1\text{e}6, 1\text{e}7, 1\text{e}8, 1\text{e}9\}$, were validated.
- MDA: Three hyper-parameters (β, α, γ) require tuning. $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\gamma \in \{1\text{e}{-3}, 1\text{e}{-2}, 1\text{e}{-1}, 1.0, 1\text{e}1, 1\text{e}2, 1\text{e}3, 1\text{e}4, 1\text{e}5, 1\text{e}6\}$, and $\alpha \in \{1, 1\text{e}1, 1\text{e}2, 1\text{e}3, 1\text{e}4, 1\text{e}5, 1\text{e}6, 1\text{e}7, 1\text{e}8, 1\text{e}9\}$, were validated.

For feature extraction methods (i.e., KPCA and KFD) and kernel-based DG methods (i.e., DICA, SCA, CIDG, and MDA), in real data experiments, different number of leading eigenvectors (corresponds to the dimension of the transformed subspace) that contribute to certain proportions (i.e. $\{0.2, 0.4, 0.6, 0.8, 0.92, 0.94, 0.96, 0.98\}$) of the sum of all eigenvalues are tested and the highest accuracies are reported for each method.

F Synthetic Experimental Results Visualization

In this section, we show the data distribution of source domains in the transformed domain-invariant subspace $\mathbb{R}^q$ of the synthetic experiment for kernel-based DG methods: SCA, CIDG, MDA, which are proposed for classification problems. The results are given in Figure 3 and 4.

We observe from the results that: 1) the transformation learned from MDA performs the best in terms of the separation of different classes of target domains; 2) the overlapped region in source domains(green and red classes) is handled slightly better in MDA than in CIDG; 3) SCA has difficulty in separating instances of different classes in part of the cases.



(a) (2a, 2a) Raw data (b) SCA(79.33%) (c) CIDG(90.67%) (d) MDA(96.67%)

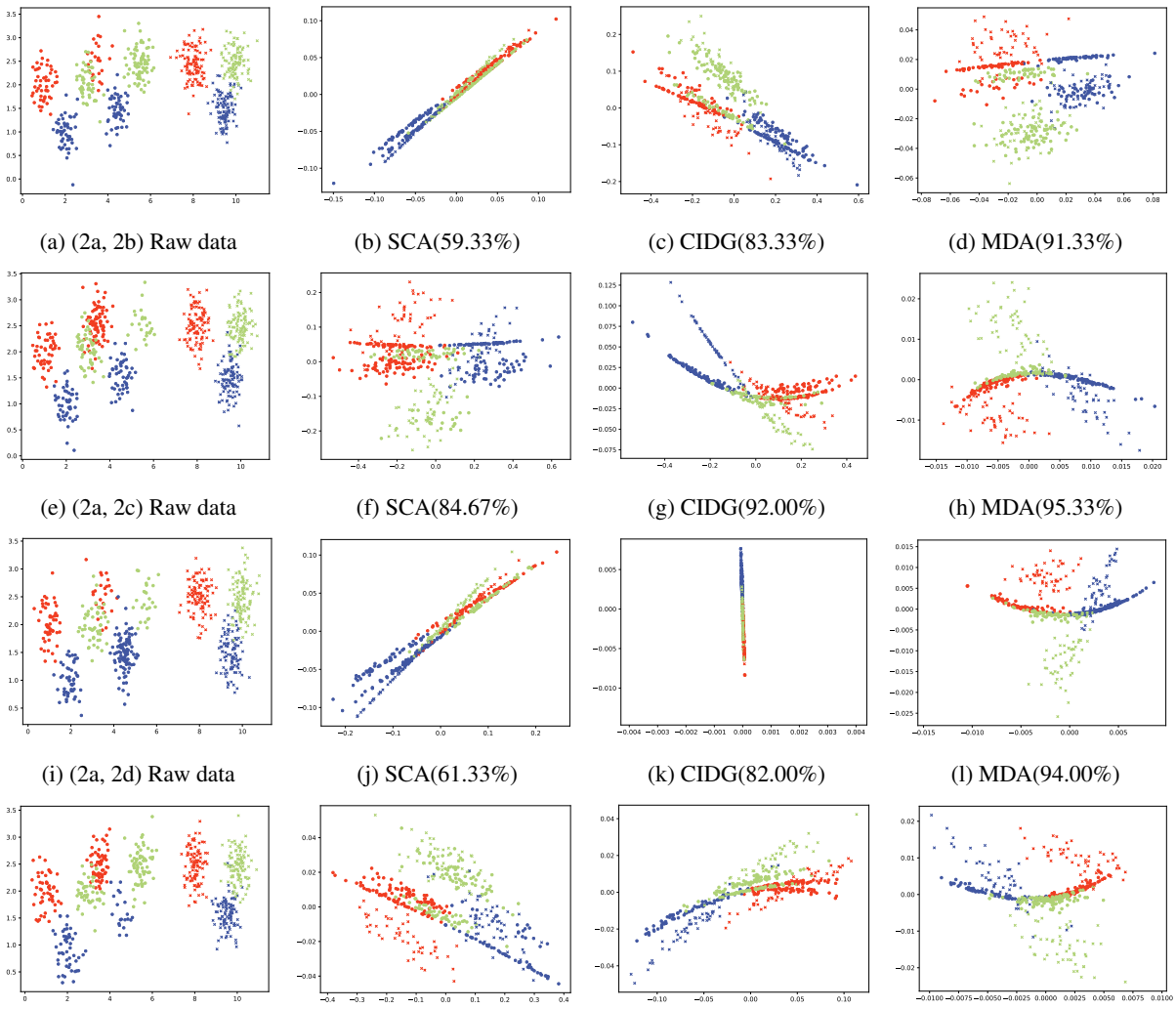
(e) (2b, 2a) Raw data (f) SCA(72.00%) (g) CIDG(87.33%) (h) MDA(96.00%)

(i) (2c, 2a) Raw data (j) SCA(84.67%) (k) CIDG(74.67%) (l) MDA(97.33%)

(m) (2d, 2a) Raw data (n) SCA(57.33%) (o) CIDG(77.33%) (p) MDA(94.00%)

(q) (2e, 2a) Raw data (r) SCA(76.00%) (s) CIDG(86.67%) (t) MDA(94.00%)

Figure 3: Visualization of transformed data in $\mathbb{R}^q$ of cases (2a, 2a), (2b, 2a), (2c, 2a), (2d, 2a), (2e, 2a). Each row corresponds to a case of class-prior distributions. Each column corresponds to a DG methods. The first column shows the distribution of the raw data. Different colors denote different classes. Circle marker denotes the data of source domain and cross marker denotes the data of target domain.



(a) (2a, 2b) Raw data (b) SCA(59.33%) (c) CIDG(83.33%) (d) MDA(91.33%)

(e) (2a, 2c) Raw data (f) SCA(84.67%) (g) CIDG(92.00%) (h) MDA(95.33%)

(i) (2a, 2d) Raw data (j) SCA(61.33%) (k) CIDG(82.00%) (l) MDA(94.00%)

(m) (2a, 2e) Raw data (n) SCA(81.33%) (o) CIDG(86.00%) (p) MDA(94.00%)

Figure 4: Visualization of transformed data in $\mathbb{R}^q$ of cases (2a, 2b), (2a, 2c), (2a, 2d), (2a, 2e). Each row corresponds to a case of class-prior distributions. Each column corresponds to a DG methods. The first column shows the distribution of the raw data. Different colors denote different classes. Circle marker denotes the data of source domain and cross marker denotes the data of target domain.

G Related Work

Compared with domain adaptation, domain generalization is a younger line of research. Blanchard et al. [2011] are the first to formalize the domain generalization of classification tasks. Motivated by automatic gating of flow cytometry data, they adopted kernel-based methods and derived the dual of a kind of cost-sensitive SVM to solve for the optimal decision function. A feature projection-based method called Domain Invariant Component Analysis (DICA; [Muandet et al., 2013]) was then proposed in 2013. DICA was the first to bring the idea of learning a shared subspace into domain generalization. It finds a transformation to a subspace in which the differences between marginal distributions $\mathbb{P}(X)$ over domains are minimized while preserving the functional relationship between Y and X.

Along this line, subsequent feature projection-based methods have been proposed. Scatter Component Analysis (SCA; [Ghifary et al., 2017]) is the first unified framework for both domain adaptation and domain generalization. It combines domain scatter, kernel principal component analysis and kernel Fisher discriminant analysis into an objective and trades between them to learn the transformation. Unlike previous works, the authors of Conditional Invariant Domain Generalization (CIDG; [Li et al., 2018b]) are the first to analyze domain generalization of classification tasks from causal perspective and thus consider more general cases where both $\mathbb{P}(Y|X)$ and $\mathbb{P}(X)$ vary across domains. They combine total scatter of class-conditional distributions, scatter of class prior-normalized marginal distributions, and kernel Fisher discriminant analysis to achieve the goal of domain generalization.

Besides the aforementioned methods in general, domain generalization problem also attracted extensive attention of computer vision community. Khosla et al. [2012] proposed a max-margin framework (Undo-Bias) in which each domain is assumed to be controlled by the sum of the visual world and a bias. A modified SVM-based method is adopted for solving the weights and biases in the model. Unbiased Metric Learning (UML; [Fang et al., 2013]), which is based on a learning-to-rank framework, first learns a set of distance metrics and then validate to select the one with best generalization ability. Xu et al. [2014] adopted exemplar-SVM and introduced a nuclear norm based regularizer into the objective to learn a set of more robust examplar-SVMs for domain generalization purpose. Ghifary et al. [2015] introduced Multi-task Autoencoder (MTAE), a feature learning algorithm that uses a multi-task strategy to learn unbiased object features, where the task is the data reconstruction. More recently, domain generalization methods based on deep neural networks [Motiian et al., 2017, Li et al., 2017, 2018a,c] were proposed to cope with the problem induced by distribution shift.

References

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in neural information processing systems*, pages 2178–2186, 2011.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1657–1664. IEEE, 2013.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

D. Li, Y. Yang, Y. Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, Oct 2017. doi: 10.1109/ICCV.2017.591.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2018a.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representation. *arXiv preprint arXiv:1807.08479*, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *The European Conference on Computer Vision (ECCV)*, September 2018c.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.