Contents

A Identify Target Joint Distribution

We show how to derive the conditions of identifying the target joint distribution with the help of the proposed data generation process, which is shown in Equation (10).

$$\begin{aligned} p_{\mathbf{x},\mathbf{y}|\mathbf{u}_\mathcal{T}} &\overset{(1)}{=} \int_{\mathbf{z}_1}\int_{\mathbf{z}_2}\int_{\mathbf{z}_3}\int_{\mathbf{z}_4} p_{\mathbf{x},\mathbf{y},\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_4|\mathbf{u}_\mathcal{T}} d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4 \\ &\overset{(2)}{=} \int_{\mathbf{z}_1}\int_{\mathbf{z}_2}\int_{\mathbf{z}_3}\int_{\mathbf{z}_4} p_{\mathbf{x},\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_4|\mathbf{y},\mathbf{u}_\mathcal{T}} \cdot p_{\mathbf{y}|\mathbf{u}_\mathcal{T}} d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4 \\ &\overset{(3)}{=} \int_{\mathbf{z}_1}\int_{\mathbf{z}_2}\int_{\mathbf{z}_3}\int_{\mathbf{z}_4} p_{\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_4} \cdot p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_4|\mathbf{y},\mathbf{u}_\mathcal{T}} \cdot p_{\mathbf{y}|\mathbf{u}_\mathcal{T}} d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4. \end{aligned} \tag{10}$$

The derivation in Equation (10) can be separated into three steps. (1) We introduce the latent variables $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, and $\mathbf{z}_4$, which have mentioned in Section 2.1. (2) We factorize the joint distribution in (1) into $p_{\mathbf{x},\mathbf{z}_1,\mathbf{z}_2,,\mathbf{z}_3,,\mathbf{z}_4|\mathbf{y},\mathbf{u}_\mathcal{T}}$ and $p_{\mathbf{y}|\mathbf{u}_\mathcal{T}}$ with the help of Bayes Rule. (3), we further use Bayes Rule to factorize $p_{\mathbf{x},\mathbf{z}_1,\mathbf{z}_2,,\mathbf{z}_3,,\mathbf{z}_4|\mathbf{y},\mathbf{u}_\mathcal{T}}$. Since $\mathbf{x}$ is independent of $\mathbf{u}, \mathbf{y}$ given $\mathbf{z}_1, \mathbf{z}_2, , \mathbf{z}_3, , \mathbf{z}_4$, we can obtain $p_{\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2,,\mathbf{z}_3,,\mathbf{z}_4}$.

The aforementioned factorization tells us that we need to model three distributions to identify the target joint distribution. First, we need to model $p_{\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_4}$, implying that we need to model the conditional distribution of observed data give latent variables, which coincides with a generative model for observed data. Second, we need to estimate the label pseudo distribution of target domain $p_{\mathbf{y}|\mathbf{u}_\mathcal{T}}$. Third, we need to model $p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_4|\mathbf{y},\mathbf{u}_\mathcal{T}}$ meaning that the latent variables should be identified with theoretical guarantees. In the next section, we will introduce how to identify these latent variables with subspace identification block-wise identification results.

B Proof of the Identification of latent variables

B.1 Proof of Subspace Identification

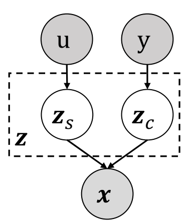


Figure 5: A simple data generalization process for introducing subspace identification.

In this subsection, we provide proof of the subspace identification based on the data generation process in Figure 5.

Theorem 3. *(**Subspace Identification of** $\mathbf{z}_s$.) We follow the data generation process in Figure 5 and make the following assumptions:*

- *A1 (Smooth and Positive Density): The probability density function of latent variables is smooth and positive, i.e., $p_{\mathbf{z}|\mathbf{u}} > 0$ over $\mathcal{Z}$ and $\mathcal{U}$.*
- *A2 (Conditional independent): Conditioned on $\mathbf{u}$, each z_i is independent of any other z_j for $i, j \in \{1, \cdots, n\}, i \neq j$, i.e. $\log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \sum_i^n q_i(z_i, \mathbf{u})$ where $q_i(z_i, \mathbf{u})$ is the log density of the conditional distribution, i.e., $q_i : \log p_{z_i|\mathbf{u}}$.*
- *A3 (Linear independence): For any $\mathbf{z}_s \in \mathcal{Z}_s \subseteq \mathbb{R}^{n_s}$, there exist $n_s + 1$ values of $\mathbf{u}$, i.e., $\mathbf{u}_j$ with $j = 0, 1, \cdots, n_s$, such that these n_s vectors $\mathbf{w}(\mathbf{z}, \mathbf{u}_j) - \mathbf{w}(\mathbf{z}, \mathbf{u}_0)$ with $j = 1, \cdots, n_s$ are linearly*

independent, where vector $\mathbf{w}(\mathbf{z}, \mathbf{u}_j)$ *is defined as follows:*

$$\mathbf{w}(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial q_1(z_1, \mathbf{u})}{\partial z_1}, \cdots, \frac{\partial q_i(z_i, \mathbf{u})}{\partial z_i}, \cdots \frac{\partial q_{n_s}(z_{n_s}, \mathbf{u})}{\partial z_{n_s}} \right), \tag{11}$$

By modeling the aforementioned data generation process, $\mathbf{z}_s$ *is subspace identifiable.*

Proof. We begin with the matched marginal distribution $p_{\mathbf{x}|\mathbf{u}}$ to bridge the relation between $\mathbf{z}$ and $\hat{\mathbf{z}}$. Suppose that $\hat{g} : \mathcal{Z} \to \mathcal{X}$ is a invertible estimated generating function, we have Equation (12).

$$\forall \mathbf{u} \in \mathcal{U}, \quad p_{\hat{\mathbf{x}}|\mathbf{u}} = p_{\mathbf{x}|\mathbf{u}} \Longleftrightarrow p_{\hat{g}(\hat{\mathbf{z}})|\mathbf{u}} = p_{g(\mathbf{z})|\mathbf{u}}. \tag{12}$$

Sequentially, by using the change of variables formula, we can further obtain Equation (13)

$$p_{\hat{g}(\hat{\mathbf{z}}|\mathbf{u})} = p_{g(\mathbf{z}|\mathbf{u})} \Longleftrightarrow p_{g^{-1}\circ g(\hat{\mathbf{z}})|\mathbf{u}} |\boldsymbol{J}_{g^{-1}}| = p_{\mathbf{z}|\mathbf{u}} |\boldsymbol{J}_{g^{-1}}| \Longleftrightarrow p_{h(\hat{\mathbf{z}})|\mathbf{u}} = p_{\mathbf{z}|\mathbf{u}}, \tag{13}$$

where $h := g^{-1} \circ g$ is the transformation between the ground-true and the estimated latent variables, respectively. $\boldsymbol{J}_{g^{-1}}$ denotes the absolute value of Jacobian matrix determinant of g^{-1}. Since we assume that g and $\hat{g}$ are invertible, $|\boldsymbol{J}_{g^{-1}}| \neq 0$ and h is also invertible.

According to A2 (conditional independent assumption), we can have Equation (14).

$$p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^{n} p_{z_i|\mathbf{u}}(z_i|\mathbf{u}); \quad p_{\hat{\mathbf{z}}|\mathbf{u}}(\hat{\mathbf{z}}|\mathbf{u}) = \prod_{i=1}^{n} p_{\hat{z}_i|\mathbf{u}}(\hat{z}_i|\mathbf{u}). \tag{14}$$

For convenience, we take logarithm on both sides of Equation (14) and further let $q_i := \log p_{z_i|\mathbf{u}}, \hat{q}_i := \log p_{\hat{z}_i|\mathbf{u}}$. Hence we have:

$$\log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \sum_{i=1}^{n} q_i(z_i, \mathbf{u}); \quad \log p_{\hat{\mathbf{z}}|\mathbf{u}} = \sum_{i=1}^{n} \hat{q}_i(\hat{z}_i, \mathbf{u}). \tag{15}$$

By combining Equation (15) and Equation (13), we have:

$$p_{\mathbf{z}|\mathbf{u}} = p_{h(\hat{\mathbf{z}}|\mathbf{u})} \Longleftrightarrow p_{\hat{\mathbf{z}}|\mathbf{u}} = p_{\mathbf{z}|\mathbf{u}} |\boldsymbol{J}_{h^{-1}}| \Longleftrightarrow \sum_{i=1}^{n} q_i(z_i, \mathbf{u}) + \log |\boldsymbol{J}_{h^{-1}}| = \sum_{i=1}^{n} \hat{q}_i(\hat{z}_i, \mathbf{u}), \tag{16}$$

where $\boldsymbol{J}_{h^{-1}}$ are the Jacobian matrix of h^{-1}.

Sequentially, we take the first-order derivative with $\hat{z}_j$ on Equation (16), where $j \in \{n_s + 1, \cdots, n\}$, and have

$$\sum_{i=1}^{n} \frac{\partial q_i(z_i, \mathbf{u})}{\partial z_i} \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \frac{\partial \log |\boldsymbol{J}_{h^{-1}}|}{\partial \hat{z}_j} = \frac{\partial q_j(\hat{z}_j, \mathbf{u})}{\partial \hat{z}_j}. \tag{17}$$

Suppose $\mathbf{u} = u_0, u_1, \cdots, u_{n_s}$, we subtract the Equation (17) corresponding to u_k with that corresponds to u_0, and we have:

$$\sum_{i=1}^{n} \left(\frac{\partial q_i(z_i, u_k)}{\partial z_i} - \frac{\partial q_i(z_i, u_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = \frac{\partial \hat{q}_j(\hat{z}_j, u_k)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0)}{\partial \hat{z}_j}. \tag{18}$$

Since the distribution of estimated $\hat{\mathbf{z}}_j$ does not change across different domains, $\frac{\partial \hat{q}_j(\hat{z}_j, u_k)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0)}{\partial \hat{z}_j} = 0$. Since $\frac{\partial q_i(z_i, u_k)}{\partial z_i}$ does not change across different domains, $\frac{\partial q_i(z_i, u_k)}{\partial z_i} = \frac{\partial q_i(z_i, u_0)}{\partial z_i}$ for $i \in \{n_s + 1, \cdots, n\}$. So we have

$$\sum_{i=1}^{n_s} \left(\frac{\partial q_i(z_i, u_k)}{\partial z_i} - \frac{\partial q_i(z_i, u_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = 0. \tag{19}$$

Based on the linear independence assumption (A3), the linear system is a $n_s \times n_s$ full-rank system. Therefore, the only solution is $\frac{\partial z_i}{\partial \hat{z}_j} = 0$ for $i \in \{1, \cdots, n_s\}$ and $j \in \{n_s + 1, \cdots, n\}$.

Since $h(\cdot)$ is smooth over $\mathcal{Z}$, its Jacobian can be formalized as follows

$$\boldsymbol{J}_h = \left[\begin{array}{c|c} \mathbf{A} := \frac{\partial \mathbf{z}_s}{\partial \hat{\mathbf{z}}_s} & \mathbf{B} := \frac{\partial \mathbf{z}_s}{\partial \hat{\mathbf{z}}_c} \\ \hline \mathbf{C} := \frac{\partial \mathbf{z}_c}{\partial \hat{\mathbf{z}}_s} & \mathbf{D} := \frac{\partial \mathbf{z}_c}{\partial \hat{\mathbf{z}}_c}. \end{array}\right] \tag{20}$$

Note that $\frac{\partial z_i}{\partial \hat{z}_j} = 0$ for $i \in \{1, \cdots, n_s\}$ and $j \in \{n_s + 1, \cdots, n\}$ means that $\mathbf{B} = 0$. Since $h(\cdot)$ is invertible, $\boldsymbol{J}_h$ is a full-rank matrix. Therefore, for each $z_{s,i}, i \in \{1, \cdots, n_s\}$, there exists a h_i such that $z_{s,i} = h_i(\hat{\mathbf{z}})$. $\square$

B.2 Proof of Corollary1.1

Corollary 3.1. *We follow the data generation in Section 3.1, and make the following assumptions which are similar to A1-A3:*

A4 (Smooth and Positive Density): The probability density function of latent variables is smooth and positive, i.e., $p_{\mathbf{z}|\mathbf{u},\mathbf{y}} > 0$ over $\mathcal{Z}, \mathcal{U}$, and $\mathcal{Y}$.

A5 (Conditional independent): Conditioned on $\mathbf{u}$ and $\mathbf{y}$, each z_i is independent of any other z_j for $i, j \in \{1, \cdots, n\}, i \neq j$, i.e. $\log p_{\mathbf{z}|\mathbf{u},\mathbf{y}}(\mathbf{z}|\mathbf{u},\mathbf{y}) = \sum_i^n q_i(z_i, \mathbf{u}, \mathbf{y})$ where $q_i(z_i, \mathbf{u}, \mathbf{y})$ is the log density of the conditional distribution, i.e., $q_i : \log p_{z_i|\mathbf{u},\mathbf{y}}$.

A6 (Linear independence): For any $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^n$, there exists $n_1 + n_2 + n_3 + 1$ combination of $(\mathbf{u}, \mathbf{y})$, i.e. $j = 1, \cdots, U$ and $c = 1, \cdots, C$ and $U \times C + 1 = n_1 + n_2 + n_3$, where U and C denote the number of source domains and the number of labels. such that these $n' = n_1 + n_2 + n_3$ vectors $\mathbf{w}(\mathbf{z}, \mathbf{u}_j, \mathbf{y}_c) - \mathbf{w}(\mathbf{z}, \mathbf{u}_0, \mathbf{y}_0)$ are linearly independent, where $\mathbf{w}(\mathbf{z}, \mathbf{u}_j, \mathbf{y}_c)$ is defined as follows:

$$\mathbf{w}(\mathbf{z}, \mathbf{u}_j, \mathbf{y}_c) = \left(\frac{\partial q_1(z_1, \mathbf{u}, \mathbf{y})}{\partial z_1}, \cdots, \frac{\partial q_i(z_i, \mathbf{u}, \mathbf{y})}{\partial z_i}, \cdots \frac{\partial q_{n'}(z_{n'}, \mathbf{u}, \mathbf{y})}{\partial z_{n'}}\right). \tag{21}$$

By modeling the aforementioned data generation process, $\mathbf{z}_2$ is subspace identifiable, and $\mathbf{z}_1, \mathbf{z}_3$ can be reconstructed from $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2$ and $\hat{\mathbf{z}}_2, \hat{\mathbf{z}}_3$, respectively.

Proof. We begin with the match marginal distribution $p_{\mathbf{x}|\mathbf{u},\mathbf{y}}$ to bridge the relation between $\mathbf{z}$ and $\hat{\mathbf{z}}$. Suppose that $\hat{g} : \mathcal{Z} \to \mathcal{X}$ is an invertible estimated generating function, we have Equation (22).

$$\forall \mathbf{u} \in \mathcal{U}, \mathbf{y} \in \mathcal{Y}, p_{\hat{\mathbf{x}}|\mathbf{u},\mathbf{y}} = p_{\mathbf{x}|\mathbf{u},\mathbf{y}} \Longleftrightarrow p_{\hat{g}(\hat{\mathbf{z}})|\mathbf{u},\mathbf{y}} = p_{g(\mathbf{z})|\mathbf{u},\mathbf{y}}. \tag{22}$$

Sequentially, by using the change of variables formula, we can further obtain Equation(23).

$$p_{\hat{g}(\hat{\mathbf{z}})|\mathbf{u},\mathbf{y}} = p_{g(\mathbf{z})|\mathbf{u},\mathbf{y}} \Longleftrightarrow p_{g^{-1}\circ g(\hat{\mathbf{z}})|\mathbf{u},\mathbf{y}}|\boldsymbol{J}_{g^{-1}}| = p_{\mathbf{z}|\mathbf{u},\mathbf{y}}|\boldsymbol{J}_{g^{-1}}| \Longleftrightarrow p_{h(\hat{\mathbf{z}})|\mathbf{u},\mathbf{y}} = p_{\mathbf{z}|\mathbf{u},\mathbf{y}}, \tag{23}$$

where $h := g^{-1} \circ g$ is the transformation between the ground-true and the estimated latent variables. $\boldsymbol{J}_{g^{-1}}$ denotes the absolute value of Jacobian matrix determinant of g^{-1}. Since we assume that g and $\hat{g}$ are invertible, $|\boldsymbol{J}_{g^{-1}}| \neq 0$ and h is also invertible.

According to A5 (conditional independent assumption), we can have Equation (24).

$$p_{\mathbf{z}|\mathbf{u},\mathbf{y}}(\mathbf{z}|\mathbf{u},\mathbf{y}) = \prod_{i=1}^{n} p_{z_i|\mathbf{u},\mathbf{y}}(z_i|\mathbf{u},\mathbf{y}); \quad p_{\hat{\mathbf{z}}|\mathbf{u},\mathbf{y}}(\hat{\mathbf{z}}|\mathbf{u},\mathbf{y}) = \prod_{i=1}^{n} p_{\hat{z}_i|\mathbf{u},\mathbf{y}}(\hat{z}_i|\mathbf{u},\mathbf{y}). \tag{24}$$

For convenience, we take logarithms on both sides of the Equation(24) and further let $q_i := \log p_{z_i|\mathbf{u},\mathbf{y}}, \hat{q}_i := \log p_{\hat{z}_i|\mathbf{u},\mathbf{y}}$. Hence we have:

$$\log p_{\mathbf{z}|\mathbf{u},\mathbf{y}}(\mathbf{z}|\mathbf{u},\mathbf{y}) = \sum_{i=1}^{n} q_i(z_i, \mathbf{u}, \mathbf{y}); \quad \log p_{\hat{\mathbf{z},\mathbf{y}}|\mathbf{u}} = \sum_{i=1}^{n} \hat{q}_i(\hat{z}_i, \mathbf{u}, \mathbf{y}). \tag{25}$$

By combining Equation (25) and Equation (23), we have:

$$p_{\mathbf{z}|\mathbf{u},\mathbf{y}} = p_{h(\hat{\mathbf{z}}|\mathbf{u},\mathbf{y})} \Longleftrightarrow p_{\hat{\mathbf{z}}|\mathbf{u},\mathbf{y}} = p_{\mathbf{z}|\mathbf{u},\mathbf{y}}|\boldsymbol{J}_{h^{-1}}| \Longleftrightarrow \sum_{i=1}^{n} q_i(z_i, \mathbf{u}, \mathbf{y}) + \log|\boldsymbol{J}_{h^{-1}}| = \sum_{i=1}^{n} \hat{q}_i(\hat{z}_i, \mathbf{u}, \mathbf{y}), \tag{26}$$

where $\boldsymbol{J}_{h^{-1}}$ are the Jacobian matrix of h^{-1}.

Sequentially, we take the first-order derivative with $\hat{z}_j$ on Equation (26), where $j \in \{n_1 + n_2 + n_3 + 1, \cdots, n\}$, and have

$$\sum_{i=1}^{n} \frac{\partial q_i(z_i, \mathbf{u}, \mathbf{y})}{\partial z_i} \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \frac{\partial \log |\boldsymbol{J}_{h^{-1}}|}{\partial \hat{z}_j} = \frac{\partial q_j(\hat{z}_j, \mathbf{u}, \mathbf{y})}{\partial \hat{z}_j}. \tag{27}$$

According to A6, there exist $n_1 + n_2 + n_3 + 1$ conbinations of $(\mathbf{u}, \mathbf{y})$, so we subtract the Equation (27) to $\mathbf{u}_k, \mathbf{y}_l$ with that corresponds to $\mathbf{u}_0, \mathbf{y}_0$, and we have:

$$\sum_{i=1}^{n} \left(\frac{\partial q_i(z_i, u_k, \mathbf{y}_l)}{\partial z_i} - \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = \frac{\partial \hat{q}_j(\hat{z}_j, u_k, \mathbf{y}_l)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0, \mathbf{y}_0)}{\partial \hat{z}_j}. \tag{28}$$

Since the distribution of estimated $\hat{\mathbf{z}}_j$ does not change across different domains and labels, $\frac{\partial \hat{q}_j(\hat{z}_j, u_k, \mathbf{y}_l)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0, \mathbf{y}_0)}{\partial \hat{z}_j} = 0$. Since $\frac{\partial q_i(z_i, u_k, \mathbf{y}_l)}{\partial z_i}$ does not change across different domains, $\frac{\partial q_i(z_i, u_k, \mathbf{y}_l)}{\partial z_i} = \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i}$ for $i \in \{1, \cdots, n_1 + n_2 + n_3\}$. So we have:

$$\sum_{i=1}^{n_1+n_2+n_3} \left(\frac{\partial q_i(z_i, u_k, \mathbf{y}_l)}{\partial z_i} - \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = 0. \tag{29}$$

Based on the linear independence assumption (A3), the linear system is a $n \times n$ full-rank system. Therefore, the only solution is $\frac{z_i}{\hat{z}_j} = 0$ for $i \in \{1, \cdots, n_1 + n_2 + n_3\}$ and $j \in \{n_1 + n_2 + n_3 + 1, \cdots, n\}$.

Since $h(\cdot)$ is smooth over $\mathcal{Z}$, its Jacobian can be formalized as follows

$$\boldsymbol{J}_h = \left[\begin{array}{c|c|c|c} \boldsymbol{J}_h^{1,1} & \boldsymbol{J}_h^{1,2} & \boldsymbol{J}_h^{1,3} & \boldsymbol{J}_h^{1,4} \\ \hline \boldsymbol{J}_h^{2,1} & \boldsymbol{J}_h^{2,2} & \boldsymbol{J}_h^{2,3} & \boldsymbol{J}_h^{2,4} \\ \hline \boldsymbol{J}_h^{3,1} & \boldsymbol{J}_h^{3,2} & \boldsymbol{J}_h^{3,3} & \boldsymbol{J}_h^{3,4} \\ \hline \boldsymbol{J}_h^{4,1} & \boldsymbol{J}_h^{4,2} & \boldsymbol{J}_h^{4,3} & \boldsymbol{J}_h^{4,4} \end{array} \right] \tag{30}$$

where $\boldsymbol{J}^{ij} := \frac{\partial \mathbf{z}_i}{\partial \hat{\mathbf{z}}_j}$ and $i, j \in \{1, 2, 3, 4\}$.

Since $\frac{z_i}{\hat{z}_j} = 0$ for $i \in \{1, \cdots, n_1 + n_2 + n_3\}$ and $j \in \{n_1 + n_2 + n_3 + 1, \cdots, n\}$, $\boldsymbol{J}_h^{3,4} = 0, \boldsymbol{J}_h^{2,4} = 0, \boldsymbol{J}_h^{1,4} = 0$.

we take the first-order derivative with $\hat{z}_j$ on Equation (26), where $j \in \{n_1 + n_2 + 1, \cdots, n\}$, and have

$$\sum_{i=1}^{n} \frac{\partial q_i(z_i, \mathbf{u}, \mathbf{y})}{\partial z_i} \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \frac{\partial \log |\boldsymbol{J}_{h^{-1}}|}{\partial \hat{z}_j} = \frac{\partial q_j(\hat{z}_j, \mathbf{u}, \mathbf{y})}{\partial \hat{z}_j}. \tag{31}$$

Then we fix the value of $\mathbf{y}$ be $\mathbf{y}_0$, so there exist U combinations of $(\mathbf{u}, \mathbf{y}_0)$. We subtract the Equation (31) corresponds to $(\mathbf{u}_k, \mathbf{y}_0)$ with that corresponds to $(\mathbf{u}_0, \mathbf{y}_0)$ and have:

$$\sum_{i=1}^{n} \left(\frac{\partial q_i(z_i, u_k, \mathbf{y}_0)}{\partial z_i} - \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = \frac{\partial \hat{q}_j(\hat{z}_j, u_k, \mathbf{y}_0)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0, \mathbf{y}_0)}{\partial \hat{z}_j}. \tag{32}$$

Since the distribution of estimated $\hat{\mathbf{z}}_j$ does not change across different domains, $\frac{\partial \hat{q}_j(\hat{z}_j, u_k, \mathbf{y}_0)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0, \mathbf{y}_0)}{\partial \hat{z}_j} = 0$. Since $\frac{\partial q_i(z_i, u_k, \mathbf{y}_0)}{\partial z_i}$ does not change across different domains, $\frac{\partial q_i(z_i, u_k, \mathbf{y}_0)}{\partial z_i} = \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i}$ for $i \in \{1, \cdots, n_1 + n_2\}$. So we have:

$$\sum_{i=1}^{n_1+n_2} \left(\frac{\partial q_i(z_i, u_k, \mathbf{y}_0)}{\partial z_i} - \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = 0. \tag{33}$$

Based on the linear independence assumption (A3), the linear system is a $n \times n$ full-rank system. Therefore, the only solution is $\frac{z_i}{\hat{z}_j} = 0$ for $i \in \{1, \cdots, n_1 + n_2\}$ and $j \in \{n_1 + n_2 + 1, \cdots, n\}$. Combining Equation (30), we can find that $\boldsymbol{J}_h^{1,3} = 0, \boldsymbol{J}_h^{1,4} = 0, \boldsymbol{J}_h^{2,3} = 0$, and $\boldsymbol{J}_h^{2,4} = 0$.

Similarly, we let $j \in \{1, \cdots, n_1\} \bigcup \{n_1 + n_2 + n_3 + 1, \cdots, n\}$ and have:

$$\sum_{i=1}^{n} \frac{\partial q_i(z_i, \mathbf{u}, \mathbf{y})}{\partial z_i} \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \frac{\partial \log |\boldsymbol{J}_{h^{-1}}|}{\partial \hat{z}_j} = \frac{\partial q_j(\hat{z}_j, \mathbf{u}, \mathbf{y})}{\partial \hat{z}_j}. \tag{34}$$

Then fix the value of $\mathbf{u}$ be $\mathbf{u}_0$, so there exist C combinations of $(\mathbf{u}_0, \mathbf{y}_l)$. We subtract the Equation (34) corresponds to $(\mathbf{u}_0, \mathbf{y}_l)$ with that corresponds to $(\mathbf{u}_0, \mathbf{y}_0)$ and have:

$$\sum_{i=n_1+1}^{n_1+n_2+n_3} \left(\frac{\partial q_i(z_i, u_0, \mathbf{y}_l)}{\partial z_i} - \frac{\partial q_i(z_i, u_0, \mathbf{y}_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = \frac{\partial \hat{q}_j(\hat{z}_j, u_0, \mathbf{y}_l)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_j(\hat{z}_j, u_0, \mathbf{y}_0)}{\partial \hat{z}_j}. \tag{35}$$

Based on the linear independence assumption (A3), the linear system is a $n \times n$ full-rank system. Therefore, the only solution is $\frac{z_i}{\hat{z}_j} = 0$ for $i \in \{n_1+1, \cdots, n_1+n_2+n_3\}$ and $j \in \{1, \cdots, n_1\} \bigcup \{n_1 + n_2 + n_3 + 1, \cdots, n\}$. Combining Equation (30), we can find that $\boldsymbol{J}_h^{2,1} = 0, \boldsymbol{J}_h^{2,4} = 0, \boldsymbol{J}_h^{3,1} = 0$, and $\boldsymbol{J}_h^{3,4} = 0$.

In summary, Equation (30) can be written as follows

$$\boldsymbol{J}_h = \left[\begin{array}{c|c|c|c} \boldsymbol{J}_h^{1,1} & \boldsymbol{J}_h^{1,2} & \boldsymbol{J}_h^{1,3} = 0 & \boldsymbol{J}_h^{1,4} = 0 \\ \hline \boldsymbol{J}_h^{2,1} = 0 & \boldsymbol{J}_h^{2,2} & \boldsymbol{J}_h^{2,3} = 0 & \boldsymbol{J}_h^{2,4} = 0 \\ \hline \boldsymbol{J}_h^{3,1} = 0 & \boldsymbol{J}_h^{3,2} & \boldsymbol{J}_h^{3,3} & \boldsymbol{J}_h^{3,4} = 0 \\ \hline \boldsymbol{J}_h^{4,1} & \boldsymbol{J}_h^{4,2} & \boldsymbol{J}_h^{4,3} & \boldsymbol{J}_h^{4,4} \end{array} \right]. \tag{36}$$

Since $h(\cdot)$ is invertible, $\boldsymbol{J}_h$ is a full-rank matrix. Therefore, for each $z_{2,i}, i \in \{n_1 + 1, \cdots, n_1 + n_2\}$, there exists a $h_{2,i}$ such that $z_{2,i} = h_i(\hat{\mathbf{z}}_2)$. Moreover, for each $z_{1,i}, i \in \{1, \cdots, n_1 + 1\}$, there exists a $h_{1,i}$ such that $z_{1,i} = h_{1,i}(\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2)$. And for each $z_{3,i}, i \in \{n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3\}$, there exists a $h_{3,i}$ such that $z_{3,i} = h_{3,i}(\hat{\mathbf{z}}_2, \hat{\mathbf{z}}_3)$. □

B.3 Proof of Blockwise Identification

Lemma 4. *[30] Following the data generation process in Section 2.1 and the assumptions A4-A6 in Theorem 3, we further make the following assumption:*

- *A7 (Domain Variability: For any set $A_{\mathbf{z}} \subseteq \mathcal{Z}$) with the following two properties: 1) $A_{\mathbf{z}}$ has nonzero probability measure, i.e. $\mathbb{P}[\mathbf{z} \in A_{\mathbf{z}} | \{\mathbf{u} = \mathbf{u}', \mathbf{y} = \mathbf{y}'\}] > 0$ for any $\mathbf{u}' \in \mathcal{U}$ and $\mathbf{y}' \in \mathcal{Y}$. 2) $A_{\mathbf{z}}$ cannot be expressed as $B_{\mathbf{z}_4} \times \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$ for any $B_{\mathbf{z}_4} \subset \mathcal{Z}_4$.*

$\exists \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, such that $\int_{\mathbf{z} \in \mathcal{A}_{\mathbf{z}}} p_{\mathbf{z}|\mathbf{u}_1,\mathbf{y}_1} d\mathbf{z} \neq \int_{\mathbf{z} \in \mathcal{A}_{\mathbf{z}}} p_{\mathbf{z}|\mathbf{u}_2,\mathbf{y}_2} d\mathbf{z}$. By modeling the data generation process in Section 2.1, the $\mathbf{z}_4$ is block-wise identifiable.

Proof. We divide the proof into four steps for better understanding.

In Step 1, we leverage the properties of the data generation process and the marginal distribution matching condition to express the marginal invariance with the indeterminacy transformation $\overline{h} : \mathcal{Z} \to \mathcal{Z}$ between the estimated and the ground-truth latent variables. The introduction of $\overline{h}(\cdot)$ allows us to formalize the block-identifiability condition.

In Step 2 and Step 3, we show that the estimated $\hat{\mathbf{z}}_4$ does not depend on the ground-truth changing variables, i.e., $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, that is, $\overline{h}_4(\mathbf{z})$ does not depend on the input $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$. To this end, in Step 2, we derive its equivalent statements which can ease the rest of the proof and avert technical issues (e.g. sets of zero probability measures). In Step 3, we prove the equivalent statement by contradiction. Specifically, we show that if $\hat{\mathbf{z}}_4$ depends of $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, the invariance derived in Step 1 would break.

In Step 4, we use the conclusion in Step 3, the smooth and bijective properties of $h(\cdot)$, and the conclusion in Corollary 1.1, to show the invertibility of the indeterminacy function between the ground-truth $\mathbf{z}_4$ and estimated $\hat{\mathbf{z}}_4$, i.e. the mapping $\hat{\mathbf{z}}_4 = \overline{h}_4(\mathbf{z}_4)$ being invertible.

Step 1. As the data generation process in Section 2.1 establishes the independence between the generation process $\hat{\mathbf{z}}_4 \sim p_{\hat{\mathbf{z}}_4}$ and $\mathbf{u}$ it follows that for any $A_{\mathbf{z}_4} \subseteq \mathcal{Z}_4$, we let $n_s = n_1 + n_2 + n_3$, then we have:

$$\begin{aligned}
&\forall \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y} \\
&\mathbb{P}\left[\{\hat{g}^{-1}_{n_s:n}(\hat{\mathbf{x}}) \in A_{\mathbf{z}_4}\}|\{\mathbf{u}=\mathbf{u}_1, \mathbf{y}=\mathbf{y}_1\}\right] = \mathbb{P}\left[\{\hat{g}^{-1}_{n_s:n}(\hat{\mathbf{x}}) \in A_{\mathbf{z}_4}\}|\{\mathbf{u}=\mathbf{u}_2, \mathbf{y}=\mathbf{y}_2\}\right] \\
&\Longleftrightarrow \\
&\forall \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y} \\
&\mathbb{P}\left[\hat{x} \in (\hat{g}^{-1}_{n_s:n})^{-1}(A_{\mathbf{z}_4})|\{\mathbf{u}=\mathbf{u}_1, \mathbf{y}=\mathbf{y}_1\}\right] = \mathbb{P}\left[\hat{x} \in (\hat{g}^{-1}_{n_s:n})^{-1}(A_{\mathbf{z}_4})|\{\mathbf{u}=\mathbf{u}_2, \mathbf{y}=\mathbf{y}_2\}\right],
\end{aligned} \tag{37}$$

where $\hat{g}^{-1}_{n_s:n} : \mathcal{X} \rightarrow \mathcal{Z}_4$ denotes the estimated transformation from the observation to the $\mathbf{z}_4$ latent variables; and $(\hat{g}^{-1}_{n_s:n})^{-1}(A_{\mathbf{z}_4}) \subseteq \mathcal{X}$ is the pre-image set of $A_{\mathbf{z}_4}$, that is , the set of estimated observations $\hat{\mathbf{x}}$ originating from $\mathbf{z}_4$ in $A_{\mathbf{z}_4}$.

Because of the matching observation distributions between the estimated model and the true model, the relation in the Equation (37) can be extended to observation $\mathbf{x}$ from the true generating process, i.e.,

$$\begin{aligned}
&\mathbb{P}\left[\{\mathbf{x} \in (\hat{g}^{-1}_{n_s:n})^{-1}(A_{\mathbf{z}_4})\}|\{\mathbf{u}=\mathbf{u}_1, \mathbf{y}=\mathbf{y}_1\}\right] = \mathbb{P}\left[\{\mathbf{x} \in (\hat{g}^{-1}_{n_s:n})^{-1}(A_{\mathbf{z}_4})\}|\{\mathbf{u}=\mathbf{u}_2, \mathbf{y}=\mathbf{y}_2\}\right] \\
&\Longleftrightarrow \\
&\mathbb{P}\left[\{\hat{g}^{-1}_{n_s:n}(\mathbf{x}) \in A_{\mathbf{z}_4}\}|\mathbf{u}=\mathbf{u}_1, \mathbf{y}=\mathbf{y}_1\right] = \mathbb{P}\left[\{\hat{g}^{-1}_{n_s:n}(\mathbf{x}) \in A_{\mathbf{z}_4}\}|\mathbf{u}=\mathbf{u}_2, \mathbf{y}=\mathbf{y}_2\right].
\end{aligned} \tag{38}$$

Since g and $\hat{g}$ are smooth and injective, there exists a smooth and injective $\overline{h} = \hat{g}^{-1} \circ g : \mathcal{Z} \rightarrow \mathcal{Z}$. We note that by definition $\overline{h} = h$ where h is introduced in the proof of Theorem 3. Expressing $\hat{g}^{-1} = \overline{h} \circ g^{-1}$ and $\overline{h}_4(\cdot) := \overline{h}_{n_s:n}(\cdot) : \mathcal{Z} \rightarrow \mathcal{Z}_4$ in Equation (38) yields

$$\begin{aligned}
&\mathbb{P}\left[\{\overline{h}_4(\mathbf{z}) \in A_{\mathbf{z}_4}\}|\{\mathbf{u}=\mathbf{u}_1, \mathbf{y}=\mathbf{y}_1\}\right] = \mathbb{P}\left[\{\overline{h}_4(\mathbf{z}) \in A_{\mathbf{z}_4}\}|\{\mathbf{u}=\mathbf{u}_2, \mathbf{y}=\mathbf{y}_2\}\right] \\
&\Longleftrightarrow \\
&\mathbb{P}\left[\{\mathbf{z} \in \overline{h}_4^{-1}(A_{\mathbf{z}_4})\}|\{\mathbf{u}=\mathbf{u}_1, \mathbf{y}=\mathbf{y}_1\}\right] = \mathbb{P}\left[\{\mathbf{z} \in \overline{h}_4^{-1}(A_{\mathbf{z}_4})\}|\{\mathbf{u}=\mathbf{u}_2, \mathbf{y}=\mathbf{y}_2\}\right] \\
&\Longleftrightarrow \\
&\int_{\mathbf{z} \in \overline{h}_4^{-1}(A_{\mathbf{z}_4})} p_{\mathbf{z}|\mathbf{u},\mathbf{y}}(\mathbf{z}|\mathbf{u}_1, \mathbf{y}_1) d\mathbf{z} = \int_{\mathbf{z} \in \overline{h}_4^{-1}(A_{\mathbf{z}_4})} p_{\mathbf{z}|\mathbf{u},\mathbf{y}}(\mathbf{z}|\mathbf{u}_2, \mathbf{y}_2) d\mathbf{z},
\end{aligned} \tag{39}$$

where $\overline{h}_4^{-1}(A_{\mathbf{z}_4}) = \{\mathbf{z} \in \mathcal{Z} : \overline{h}_4(\mathbf{z}) \in A_{\mathbf{z}_4}\}$ is the pre-image of $A_{\mathbf{z}_4}$, i.e., those latent variables containing $\mathbf{z}_4$ in $A_{\mathbf{z}_4}$ after the indeterminacy transformation h.

Based on the proposed generation process in Section 2.1, we rewrite Equation (39) as follows:

$$\begin{aligned}
&\forall A_{\mathbf{z}_4} \subseteq \mathcal{Z}_4, \\
&\int_{[\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top, \mathbf{z}_4^\top]^\top \in \overline{h}_4^{-1}(A_{\mathbf{z}_4})} p_{\mathbf{z}_4}(\mathbf{z}_4)(p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{u}_1, \mathbf{y}_1) \\
&- p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{u}_2, \mathbf{y}_2)) d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4 = 0
\end{aligned} \tag{40}$$

Step 2. In order to show the block-identifiability of $\mathbf{z}_4$, we would like to prove that $\mathbf{z}_c := \overline{h}([\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top, \mathbf{z}_4^\top]^\top)$ does not depend on $\mathbf{z}_{1:n_s}$. To this end, we first develop one equivalent statement (i.e., State 3 below) and prove it in a later step instead. By doing so, we are able to leverage the full-support density function assumption to avert technical issues.

- Statement 1: $\overline{h}_4([\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top, \mathbf{z}_4^\top]^\top)$ does not depend on $\mathbf{z}_{1:n_s}$
- Statement 2: $\forall \mathbf{z}_4 \in \mathcal{Z}_4$, it follows that $\overline{h}_4^{-1} = B_{\mathbf{z}_4} \times \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$ where $B_{\mathbf{z}_4} \neq \emptyset$ and $B_{\mathbf{z}_4} \subseteq \mathcal{Z}_4$.
- Statement 3: $\forall \mathbf{z}_4 \in \mathcal{Z}_4, r \in \mathbb{R}^+$, it follows that $\overline{h}_4^{-1}(\mathcal{B}_r(\mathbf{z}_4)) = B^+_{\mathbf{z}_4} \times \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$ where $\mathcal{B}_r(\mathbf{z}_4) := \{\mathbf{z}'_4 \in \mathcal{Z}_4 : ||\mathbf{z}'_4 - \mathbf{z}_4||^2 < r\}$, $B^+_{\mathbf{z}_4} \neq \emptyset$, and $B^+_{\mathbf{z}_4} \subseteq \mathcal{Z}_4$.

Statement 2 is a mathematical formulation of Statement 1. Statement 3 generalizes singletons $\mathbf{z}_4$ in Statement 2 to open, non-empty balls $\mathcal{B}_r(\mathbf{z}_4)$. Later, we use Statement 3 in Step 3 to show the contraction to Equation (40).

Leveraging the continuity of $\overline{h}_4(\cdot)$, we can show the equivalence between Statement 2 and Statement 3 as follows. We first show that Statement 2 implies Statement 3. $\forall \mathbf{z}_4, r \in \mathbb{R}^+, \overline{h}_c^{-1}(\mathcal{B}(\mathbf{z}_4)) = \bigcup_{\mathbf{z}'_4 \in \mathcal{B}_r(\mathbf{z}_4)} h_4^{-1}(\mathbf{z}'_4)$. Statement 2 indicates that every participating sets in the union satisfies $h_4^{-1}(\mathbf{z}'_4) = B'_{\mathbf{z}_4} \times \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$, thus the union $\overline{h}_c^{-1}(\mathcal{B}_r(\mathbf{z}_4))$ also satisfies this property, which is Statement 3.

Then, we show that Statement 3 implies Statement 2 by contradiction. Suppose that Statement 2 is false, then $\exists \hat{\mathbf{z}}_4 \in \mathcal{Z}_4$ such that there exist $\hat{\mathbf{z}}_4^B \in \{\mathbf{z}_{n_s:n} : \mathbf{z} \in \overline{h}_4^{-1}(\hat{\mathbf{z}}_4)\}$ and $\hat{\mathbf{z}}_{n_s}^B \in \mathcal{Z}_{n_s}$ resulting in $\overline{h}_4(\hat{\mathbf{z}}^B) \neq \hat{\mathbf{z}}_4$ where $\hat{\mathbf{z}}^B = [(\hat{z}_4^B)^\top, (\hat{\mathbf{z}}_{n_s}^B)^\top]^\top$. As $\overline{h}_4(\cdot)$ is continuous, there exists $\hat{r} \in \mathbb{R}^+$ such that $\overline{h}_4(\hat{\mathbf{z}}^B) \notin \mathcal{B}_{\hat{r}}(\hat{\mathbf{z}}_4)$. That is, $\hat{\mathbf{z}}^B \notin h_4^{-1}(\mathcal{B}_{\hat{r}}(\hat{\mathbf{z}}_4))$. Also, Statement 4 suggests that $h_4^{-1}(\mathcal{B}_{\hat{r}}(\hat{\mathbf{z}}_c)) = \hat{B}_{\mathbf{z}_4} \times \mathcal{Z}_{n_s}$. By definition of $\hat{\mathbf{z}}^B$, it is clear that $\hat{\mathbf{z}}_{n_s:n}^B \in \hat{B}_{\mathbf{z}_4}$. The fact that $\hat{\mathbf{z}}^B \notin h_4^{-1}(\mathcal{B}_{\hat{r}}(\hat{\mathbf{z}}_4))$ contradicts Statement 3. Therefore, Statement 2 is true under the premise of Statement 3. We have shown that Statement 3 implies Statement 2. In summary, Statement 2 and Statement 3 are equivalent, and therefore proving Statement 3 suffices to show Statement 1.

Step 3. In this step, we prove State 3 by contradiction. Intuitively, we show that if $\overline{h}_4(\cdot)$ depended on $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \hat{\mathbf{z}}_3$, the preimage $\overline{h}_4^{-1}(\mathcal{B}_r(\mathbf{z}_4))$ could be partitioned into two parts (i.e. $B^*_{\mathbf{z}}$ and $\overline{h}_4^{-1}(A^*_{\mathbf{z}_4}) \backslash B^*_{\mathbf{z}}$ defined below). The dependency between $\overline{h}_4(\cdot)$ and $\hat{\mathbf{z}}_4$ is captured by $B^*_{\mathbf{z}}$, which would not emerge otherwise. In contrast, $\overline{h}_4^{-1} \backslash B^*_{\mathbf{z}}$ also exists when $\overline{h}_4(\cdot)$ does not depend on $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \hat{\mathbf{z}}_3$. We evaluate the invariance relation Equation (40) and show that the integral over $\overline{h}_4^{-1}(A^*_{\mathbf{z}_4}) \backslash B^*_{\mathbf{z}}$ is always 0, however, the integral over $B^*_{\mathbf{z}}$ is necessarily non-zero, which leads to the contraction with Equation (40) and thus show the $\overline{h}_4(\cdot)$ cannot depend on $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \hat{\mathbf{z}}_3$,

First, note that because $\mathcal{B}_r(\mathbf{z}_4)$ is open and $\overline{h}_4(\cdot)$ is continuous, the pre-image $\overline{h}_4^{-1}(\mathcal{B}_r(\mathbf{z}_4))$ is open. In addition, the continuity of $h(\cdot)$ and the matched observation distributions $\forall \mathbf{u}' \in \mathcal{U}, \mathbb{P}[\{\mathbf{x} \in A_{\mathbf{x}}\} | \{\mathbf{u} = \mathbf{u}', \mathbf{y} = \mathbf{y}'\}] = \mathbb{P}[\{\hat{\mathbf{x}} \in A_{\mathbf{x}}\} | \{\mathbf{u} = \mathbf{u}', \mathbf{y} = \mathbf{y}'\}]$ lead to $h(\cdot)$ being bijection as shown in [29], which implies that $\overline{h}_4^{-1}(\mathcal{B}_r(\mathbf{z}_4))$ is non-empty. Hence, $\overline{h}_4^{-1}(\mathcal{B}_r(\mathbf{z}_4))$ is both non-empty and open. Suppose that $\exists A^*_{\mathbf{z}_4} := \mathcal{B}_{r^*}(\mathbf{z}_4^*$ where $\mathbf{z}_4^* \in \mathcal{Z}_4, r^* \in \mathbb{R}^+$, such that $B^*_{\mathbf{z}} = \{\mathbf{z} \in \mathcal{Z} : \mathbf{z} \in \overline{h}_c^{-1}(A^*_{\mathbf{z}_4}), \{\mathbf{z}_{n_s:n}\} \times \mathcal{Z}_{n_s} \nsubseteq \overline{h}_4^{-1}(A^*_{\mathbf{z}_4})\} \neq \emptyset$. Intuitively, $B^*_{\mathbf{z}}$ contains the partition of the pre-image $\overline{h}_4^*(A^*_{\mathbf{z}})$ that the style part $\mathbf{z}_{1:n_s}$ can not take on any value in $\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3$. Only certain values of the style part were able to produce specific outputs of indeterminacy $\overline{h}_4(\cdot)$. Clearly, this would suggest that $\overline{h}_4(\cdot)$ depends on $\mathbf{z}_4$. To show contraction with Equation (40), we evaluate the LHS of Equation (40) with such a $A^*_{\mathbf{z}_4}$:

$$
\begin{aligned}
&\int_{[\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top, \mathbf{z}_4^\top]^\top \in \overline{h}_4^{-1}(A^*_{\mathbf{z}_4})} P_{\mathbf{z}_4}(\mathbf{z}_4) \left(p_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}, \mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}_1, \mathbf{y}_1) - p_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{z}, \mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}_2, \mathbf{y}_2) \right) d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4 \\
= &\underbrace{\int_{[\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top, \mathbf{z}_4^\top]^\top \in \overline{h}_4^{-1}(A^*_{\mathbf{z}_4}) \backslash B^*_{\mathbf{z}}} P_{\mathbf{z}_4}(\mathbf{z}_4) \left(p_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}, \mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}_1, \mathbf{y}_1) - p_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}, \mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}_2, \mathbf{y}_2) \right) d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4}_{T_1} \\
+ &\underbrace{\int_{[\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top, \mathbf{z}_4^\top]^\top \in B^*_{\mathbf{z}}} P_{\mathbf{z}_4}(\mathbf{z}_4) \left(p_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}, \mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}_1, \mathbf{y}_1) - p_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}, \mathbf{y}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 | \mathbf{u}_2, \mathbf{y}_2) \right) d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4}_{T_2}
\end{aligned}
\tag{41}
$$

We first look at the value of T_1. When $\overline{h}_4^{-1}(A^*_{\mathbf{z}_4}) \backslash B^*_{\mathbf{z}} = \emptyset$, T_1 evaluates to 0. Otherwise, by definition, we can rewrite $\overline{h}_4^{-1}(A^*_{\mathbf{z}_4}) \backslash B^*_{\mathbf{z}}$ as $C^*_{\mathbf{z}_4} \times \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$ where $C^*_{\mathbf{z}_4} \subset \mathcal{Z}_4$. With this expression, it follows that

$$
\begin{aligned}
&\int_{[\mathbf{z}_1^\top,\mathbf{z}_2^\top,\mathbf{z}_3^\top,\mathbf{z}_4^\top]^\top\in C^*_{C^*_{\mathbf{z}}}} P_{\mathbf{z}_4}(\mathbf{z}_4)\left(p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u}_1,\mathbf{y}_1)-p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u}_2,\mathbf{y}_2)\right)d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4\\
=&\int_{\mathbf{z}_4\in C^*_{\mathbf{z}_4}} p_{\mathbf{z}_4}(\mathbf{z}_4)\int_{\mathbf{z}_1\in\mathcal{Z}_1}\int_{\mathbf{z}_2\in\mathcal{Z}_2}\int_{\mathbf{z}_3\in\mathcal{Z}_3}(p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u}_1,\mathbf{y}_1)-p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u}_2,\mathbf{y}_2))d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4\\
=&\int_{\mathbf{z}_4\in C_{\mathbf{z}_4^*}} p_{\mathbf{z}_4}(\mathbf{z}_4)(1-1)d_{\mathbf{z}_4}=0.
\end{aligned}
\tag{42}
$$

Therefore, in both cases T_1 evaluates to 0 for $A^*_{\mathbf{z}_4}$.

Now, we address T_2. As discuss above, $\overline{h}_4^{-1}(A^*_{\mathbf{z}_4})$ is open and non-empty. Because of the continuity of $\overline{h}_4(\cdot)$, $\forall \mathbf{z}_B \in B^*_{\mathbf{z}}$, there exists $r(\mathbf{z}_B)\in\mathbb{R}^+$ such that $\mathcal{B}_{r(\mathbf{z}_B)}(\mathbf{z}_B)\subseteq B^*_{\mathbf{z}}$. As $p_{\mathbf{z}|\mathbf{u},\mathbf{y}}>0$ over $(\mathbf{u},\mathbf{z},\mathbf{y})$, we have $\mathbb{P}[\{\mathbf{z}\in B^*_{\mathbf{z}}\}|\{\mathbf{u}=\mathbf{u}',\mathbf{y}=\mathbf{y}'\}]\geq\mathbb{P}[\{\mathbf{z}\in B_{r(\mathbf{z}_B)}(\mathbf{z}_B)\}|\{\mathbf{u}=\mathbf{u}',\mathbf{y}=\mathbf{y}'\}]>0$ for any $\mathbf{z}'\in\mathcal{U},\mathbf{y}\in\mathcal{Y}$. Assumption A7 indicates that $\exists\mathbf{u}_1^*,\mathbf{u}_2^*$, such that

$$
\begin{aligned}
T_2 := \int_{[\mathbf{z}_1^\top,\mathbf{z}_2^\top,\mathbf{z}_3^\top,\mathbf{z}_4^\top]^\top\in B^*_{\mathbf{z}}} & P_{\mathbf{z}_4}(\mathbf{z}_4)\left(p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u}_1,\mathbf{y}_1)\right.\\
&\left.-\,p_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u},\mathbf{y}}(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{u}_2,\mathbf{y}_2)\right)d\mathbf{z}_1 d\mathbf{z}_2 d\mathbf{z}_3 d\mathbf{z}_4\neq 0.
\end{aligned}
\tag{43}
$$

Therefore, for such $A^*_{\mathbf{z}_4}$, we would have $T_1+T_2\neq 0$ which leads to contradiction with Equation (40). We have proved by contradiction that Statement 3 is true and hence Statement 1 holds, that is, $\overline{h}_4(\cdot)$ does not depend on the changing variables $\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3$.

Step 4. With the knowledge that $\overline{h}_4(\cdot)$ does not depend on the changing variables $\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3$, we now show that there exists an invertible mapping between the true $\mathbf{z}_4$ and the estimated $\mathbf{z}_4$.

As $\overline{h}(\cdot)$ is smooth over $\mathcal{Z}$, its Jacobian can written as:

$$
\boldsymbol{J}_{\overline{h}}=\left[\begin{array}{c|c|c|c}
\boldsymbol{J}_{\overline{h}}^{1,1} & \boldsymbol{J}_{\overline{h}}^{1,2} & \boldsymbol{J}_{\overline{h}}^{1,3} & \boldsymbol{J}_{\overline{h}}^{1,4}\\ \hline
\boldsymbol{J}_{\overline{h}}^{2,1} & \boldsymbol{J}_{\overline{h}}^{2,2} & \boldsymbol{J}_{\overline{h}}^{2,3} & \boldsymbol{J}_{\overline{h}}^{2,4}\\ \hline
\boldsymbol{J}_{\overline{h}}^{3,1} & \boldsymbol{J}_{\overline{h}}^{3,2} & \boldsymbol{J}_{\overline{h}}^{3,3} & \boldsymbol{J}_{\overline{h}}^{3,4}\\ \hline
\boldsymbol{J}_{\overline{h}}^{4,1} & \boldsymbol{J}_{\overline{h}}^{4,2} & \boldsymbol{J}_{\overline{h}}^{4,3} & \boldsymbol{J}_{\overline{h}}^{4,4}
\end{array}\right],
\tag{44}
$$

in which $\boldsymbol{J}_{\overline{h}}^{i,j}$ denotes $\frac{\partial\hat{\mathbf{z}}_i}{\partial\hat{\mathbf{z}}_j}, i,j\in\{1,2,3,4\}$; and we use notation $\hat{\mathbf{z}}_4=\overline{h}(\mathbf{z})_{n_s:n}$, $\hat{\mathbf{z}}_1=\overline{h}(\mathbf{z})_{1:n_1}$, $\hat{\mathbf{z}}_2=\overline{h}(\mathbf{z})_{n_1+1:n_2}$, $\hat{\mathbf{z}}_3=\overline{h}(\mathbf{z})_{n_1+n_2+1:n_3}$. As we have shown that $\hat{\mathbf{z}}_4$ does not depend on the changing variables $\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3$, if follows $\boldsymbol{J}_{\overline{h}}^{4,1}=0, \boldsymbol{J}_{\overline{h}}^{4,2}=0, \boldsymbol{J}_{\overline{h}}^{4,3}=0$. On the other hand, as $h(\cdot)$ is invertible over $\mathcal{Z}$, $\boldsymbol{J}_{\overline{h}}$ is non-singular. Therefore, $\boldsymbol{J}_{\overline{h}}^{4,4}$ must be non-singular. We note that $\boldsymbol{J}_{\overline{h}}^{4,4}$ is the Jacobian of the function $\overline{h}_4':=\overline{h}_c(\mathbf{z}):\mathcal{Z}_4\rightarrow\mathcal{Z}_4$, which takes only the $\mathbf{z}_4$ of the input $\mathbf{z}$ into $\overline{h}_4$. According to Corollary 1.1, we also find that $\boldsymbol{J}_{\overline{h}}^{1,4}=0, \boldsymbol{J}_{\overline{h}}^{2,4}=0, \boldsymbol{J}_{\overline{h}}^{3,4}=0$. Together with the invertibility of $\overline{h}$, we can conclude that $\overline{h}_4'$ is invertible. Therefore, there exists an invertible function $\overline{h}_4'$ between the estimated and the true variables such that $\hat{\mathbf{z}}_4=\overline{h}_4'(\mathbf{z}_4)$, which concludes the proof that $\mathbf{z}_4$ is block identifiable via $\hat{g}^{-1}(\cdot)$.

□

C Implementation Details

The implementation details of the proposed SIG model are shown in Table 1. For Office-Home and ImageCLEF datasets, we employ the pre-trained ResNet50 as the backbone networks. For the PACS dataset, we use the pre-trained ResNet18 as the backbone network. It is noted that we employ a ResNet101-based cross-attention network (CAN) as the backbone network, which is shown in Figure 6. In CAN, we inject a cross-attention module into each block of the pre-trained ResNet. Technologically, we use the input feature (e.g. f_1 in Figure 6) and the domain index to calculate the weights $\mathbf{w}_c$. Sequentially, we take $\mathbf{w}_c\odot f_1$ as the input of the pre-trained ResNet Layers and obtain the output of each block.

Table 5: Implementation details of the SIG model in different datasets.

Datasets	Office-Home	ImageCLEF	PACS	DomainNet
Encoder	2-layers MLPs	2-layers MLPs	2-layers MLPs	1-layers MLPs
Decoder	2-layers MLPs	2-layers MLPs	2-layers MLPs	2-layers MLPs
learning rate	0.008	0.01	0.01	0.001
α	1.00E-05	1.00E-05	1.00E-05	1.00E-05
β	0.1	0.1	0.1	0.1
$\mathbf{z}_1$ dimension	2	4	2	2
$\mathbf{z}_2$ dimension	128	128	60	2048
$\mathbf{z}_3$ dimension	128	10	24	32
$\mathbf{z}_4$ dimension	10	4	2	2
Optimizer	SGD	SGD	SGD	SGD
Momentum	0.9	0.9	0.9	0.9
batch size	32	32	32	100
backbone	ResNet50	ResNet50	ResNet18	ResNet101-based CAN

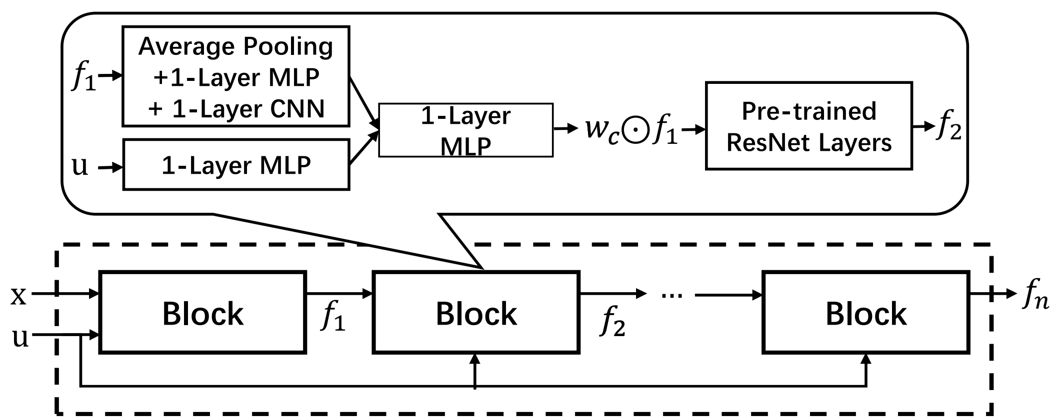


Figure 6: A illustrate framework of the ResNet101-based cross-attention networks (CAN). In each block of the ResNet101, we use the domain information and the inputs of each block to calculate the weights $\mathbf{w}_c$ of each dimension of the feature, which dynamically selects the most relevant features.

D Experiments

D.1 Simulation Data Experiments

We provide more details for the simulation experiments. First, we introduce the details of model architecture for simulation experiments. Second, we further provide the training hyper-parameters.

D.1.1 Model Architecture.

For the model architecture of our simulation experiments, the variational auto-encoder (VAE) encoder and decoder are 1-layer MLPs with a hidden dimension of 200, a ReLU activation function, a batch normalization layer, and a dropout layer.

D.1.2 Training Hyper-parameters.

We use an SGD optimizer with a momentum of 0.9 to train VAE models with 50 epochs. We also use a learning rate of 0.0035 with a batch size of 768. For the VAE training, we set the hyper-parameters of the KL loss to 1.

D.2 Real-world Data Experiments

We provide implementation details of real-world data experiments. First, we provide detailed descriptions of Office-Home, ImageCLEF, PACS, and DomainNet datasets. Second, we show more experiment results, including more baselines, the mean, and the standard deviation of the results.

D.2.1 Dataset Description

Office-Home is a benchmark dataset with 4 domains, where each domain contains 65 categories. These four domains are shown as follows: Art contains artistic images in the form of sketches, paintings, ornamentation, etc.; Clipart contains the collection of clipart images; Product contains images of objects without a background and Real-World contains images of objects captured with a regular camera. **ImageCLEF** is a standard domain adaptation benchmark dataset for image classification, consisting of three domains: Caltech-256(C), ImageNet ILSVRC(I), and Pascal VOC2012(P), consisting of 12 classes. **PACS** is a domain adaptation dataset with 9991 images from 4 domains of different styles: Photo, Artpainting, Cartoon, and Sketch. It is noted that these domains are shared with the same 7 categories. **DomainNet** is a challenging domain adaptation benchmark with 0.6 million images of 345 categories of 6 different styles: clipart, infograph, painting, quickdraw, real, and sketch.

D.2.2 More Experimental Results

To show the effectiveness of the proposed SIG model, we further consider more compared methods. Experiment results for Office-Home, ImageCLEF, PACS, and DomainNet are shown in Table 6, 7, 8, and 9, respectively. Note that We report the mean and the standard deviation of our method over 3 random seeds (i.e. 3,4,5).

Table 6: Classification results on the Office-home datasets. We employ ResNet50 as the backbone network. Baseline results are taken from ([30]).

Models	Art	Clipart	Product	RealWorld	Average
Source Only [16]	64.5 (0.68)	52.3 (0.63)	77.6 (0.23)	80.7 (0.81)	68.8
DANN [11]	64.2 (0.59)	58.0 (1.55)	76.4 (0.47)	78.8 (0.49)	69.3
DANN+BSP [7]	66.1 (0.27)	61.0 (0.39)	78.1 (0.31)	79.9 (0.13)	71.2
DAN [40]	68.2 (0.45)	57.9 (0.65)	78.4 (0.05)	81.9 (0.35)	71.6
MCD [50]	67.8 (0.38)	59.9 (0.55)	79.2 (0.61)	80.9 (0.18)	71.9
DCTN [69]	66.9 (0.60)	61.8 (0.46)	79.2 (0.58)	77.7 (0.59)	71.4
MIAN-γ [45]	69.8 (0.35)	**64.2** (0.68)	80.8 (0.37)	81.4 (0.24)	74.1
iMSDA [30]	75.4 (0.86)	61.4 (0.73)	83.5 (0.22)	84.4 (0.38)	76.1
SIG	**76.4** (0.37)	63.9 (0.34)	**85.4** (0.36)	**85.8** (0.22)	**77.8**

E Sensitive Analysis of Hyper-parameters

We also consider the sensitive analysis of α and β, which is shown in Figure 7(a) and 7(b). In detail, we consider different values of α ($\{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3\}$). According to the experiment results, we find that the model performance is stable with α. We also try different values of β ($\{1e-5, 3e-5, 5e-5, 7e-5, 9e-5, 1e-4, 5e-4, 1e-3\}$), we find that the model performance

Table 7: Classification results on the ImageCLEF datasets. We employ ResNet50 as the backbone network. Baseline results are taken from ([47]).

Mode	I,C→P	I,P→C	P,C→I	Average
Source Only [16]	77.2	92.3	88.1	85.8
DAN [40]	77.6	93.3	92.2	87.7
ADDA [61]	76.5	94.0	93.2	87.0
DANN [11]	77.9	93.7	91.8	87.8
D-CORAL [57]	77.1	93.6	91.7	87.5
DSBN [4]	77.7 (0.2)	94.1 (0.3)	91.9 (0.1)	87.9
DSAN [82]	77.6 (0.2)	95.1 (0.1)	91.4 (0.6)	88.1
MFSAN [81]	79.1	95.4	93.6	89.4
PTMDA [47]	79.1 (0.2)	97.3 (0.3)	94.1 (0.3)	90.1
SIG	**79.3** (0.57)	**97.3** (0.34)	**94.3** (0.07)	**90.3**

Table 8: Classification results on the PACS datasets. We employ ResNet18 as the backbone network. Baseline results are taken from ([30]).

Model	A	C	P	S	Average
Source Only [16]	74.9 (0.88)	72.1	94.5	64.7 (1.53)	76.7
DANN [11]	81.9 (1.13)	77.5 (1.26)	91.8 (1.21)	74.6 (1.03)	81.5
MDAN [79]	79.1 (0.36)	76.0 (0.73)	91.4 (0.85)	72.0 (0.80)	79.6
WBN [43]	89.9 (0.28)	89.7 (0.56)	97.4 (0.84)	58.0 (1.51)	83.8
MCD [50]	88.7 (1.01)	88.9 (1.53)	96.4 (0.42)	73.9 (3.94)	87
M3SDA [46]	89.3 (0.42)	89.9 (1.00)	97.3 (0.31)	76.7 (2.86)	88.3
CMSS [70]	88.6 (0.36)	90.4 (0.80)	96.9 (0.27)	82.0 (0.59)	89.5
LtC-MSDA [63]	90.1	90.4	97.2	81.5	89.8
T-SVDNet [33]	90.4	90.6	98.5	85.4	91.2
iMSDA [30]	93.7 (0.32)	92.4 (0.23)	98.4 (0.07)	89.2 (0.73)	93.4
SIG	**94.0** (0.07)	**93.6** (0.49)	**98.6** (0.06)	**89.5** (0.71)	**93.9**

is stable in the range of $1e-5 \sim 5e-4$, but it drop slightly when the value of β becomes too large, e.g. $1e-3$.

F Visualization

To evaluate the effectiveness of the SIG model qualitatively, we also provide the visualization results in t-SNE as shown in Figure 8. According to the visualization, we can find that our SIG model can generate the features with a more clear class boundary.

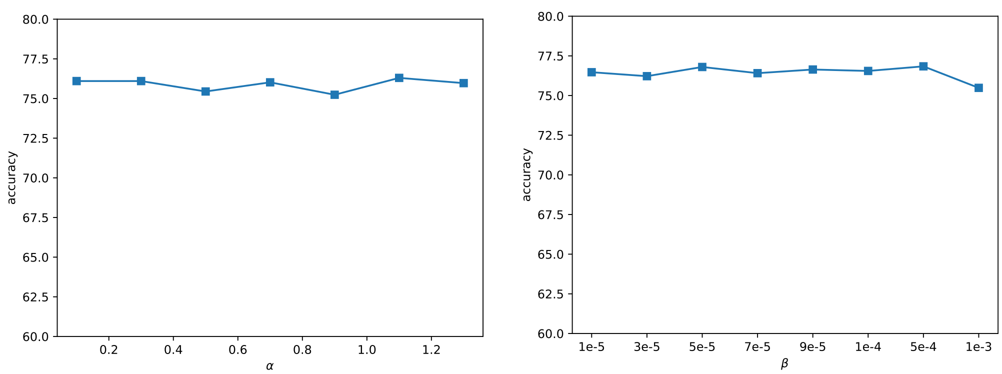
G Related Works

G.1 Domain Adaptation

Domain adaptation [3, 77, 36, 30, 75, 76, 65, 54, 49] leverages the knowledge from the labeled source data and unlabeled target data to build a model with ideal generalization. Several researchers solve the

Table 9: Classification results on the DomainNet datasets. We employ ResNet101 as the backbone network. Baseline results are taken from ([34] and [64]).

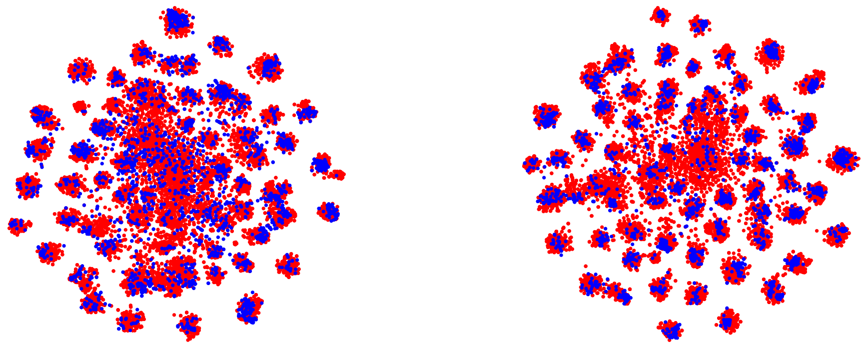
Model	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
Source Only [16]	52.1 (0.51)	23.4 (0.28)	47.6 (0.96)	13.0 (0.72)	60.7 (0.23)	46.5 (0.56)	40.6
ADDA [61]	47.5 (0.76)	11.4 (0.67)	36.7 (0.53)	14.7 (0.50)	49.1 (0.82)	33.5 (0.49)	32.2
MCD [50]	54.3 (0.64)	22.1 (0.70)	45.7 (0.63)	7.6 (0.49)	58.4 (0.65)	43.5 (0.57)	38.5
DANN [11]	60.6 (0.42)	25.8 (0.43)	50.4 (0.51)	7.70.68)	62.0 (0.66)	51.7 (0.19)	43.0
DCTN [69]	48.6 (0.73)	23.5 (0.59)	48.8 (0.63)	7.2 (0.46)	53.5 (0.56)	47.3 (0.47)	38.2
M^{3}SDA-β [46]	58.6 (0.53)	26.0 (0.89)	52.3 (0.55)	6.3 (0.58)	62.7 (0.51)	49.5 (0.76)	42.6
ML_MSDA [35]	61.4 (0.79)	26.2 (0.41)	51.9 (0.20)	19.1 (0.31)	57.0 (1.04)	50.3 (0.67)	44.3
meta-MCD [32]	62.8 (0.22)	21.4 (0.07)	50.5 (0.08)	15.5 (0.22)	64.6 (0.16)	50.4 (0.12)	44.2
LtC-MSDA [63]	63.1 (0.5)	28.7 (0.7)	56.1 (0.5)	16.3 (0.5)	66.1 (0.6)	53.8 (0.6)	47.4
CMSS [70]	64.2 (0.18)	28.0 (0.20)	53.6 (0.39)	16.9 (0.12)	63.4 (0.21)	53.8 (0.35)	46.5
DRT+ST [34]	71.0 (0.21)	31.6 (0.44)	61.0 (0.32)	12.3 (0.38)	71.4 (0.23)	60.7 (0.31)	51.3
SPS [64]	70.8	24.6	55.2	19.4	67.5	57.6	49.2
PFDA [10]	64.5	29.2	57.6	17.2	67.2	55.1	48.5
SIG	**72.7** (0.42)	**32.0** (0.71)	60.9 (0.87)	**20.5** (0.71)	**72.4** (0.14)	59.5 (0.70)	53.0



(a) Sensitive results of α (b) Sensitive results of β

Figure 7: Sensitive analysis of α and β on the $\rightarrow$ Task in Office-Home.

challenges of domain adaptation from different perspectives. One of the most conventional directions is to learn the domain-invariant representation [2], which is raised by [11]. Specifically, the key idea of these methods is to extract the domain-invariant representation by aligning the features from different domains. Some researchers [41] use maximum mean discrepancy (MMD) to realize the domain alignment. Tzeng et.al [62] extract the domain-invariant representation by using an adaptation layer and a domain confusion loss. Another type of idea assumes that the conditional distributions $P(\boldsymbol{z}|y)$ are stable across domains and extract the domain-invariant representation condition on each class [6, 5, 26]. Specifically, Xie et.al [68] minimize the domain discrepancy of inter-class features; Shu et.al [53] consider that the decision boundaries should not cross high-density data regions so they propose the virtual adversarial domain adaptation model. Target shift [77, 37, 66, 12, 48] is also common in domain adaptation, which assumes $p_{\mathbf{y}|\mathbf{u}}$ varies with different domains. Shui et.al [54] propose a unified framework to select relevant sources based on the similarity of the conditional distribution. And Remi et.al [58] analyze the generalized label shift and further provide theoretical guarantees on the transfer performance of any classifier. Recently, several researchers address the domain adaptation problem from the lens of causality [30, 42, 59, 8, 13, 55]. Zhang et.al [77] assume that $P(y)$ and $P(\boldsymbol{x}|y)$ change independently, and raise the target shift, conditional shift, and generalized target shift assumptions. Cai et.al [3] employ the causal generation process to



(a) iMSDA (b) SIG

Figure 8: The t-SNE visualizations of learned features on the → Art task in Office-Home. Red: source domains, Blue: target domain.

extract the disentangled semantic representation. Based on the causal analysis, Petar et.al [56] find that the domain-invariant should be extracted with the help of domain knowledge, so they propose domain-specific adversarial networks. Despite the outstanding performance of the aforementioned methods, these methods are built on the ad-hoc causal generation process and can not identify the latent variables. In the paper, the proposed **SIG** method is built on a more general causal generation process and identifies the latent variables with the help of the subspace identification guarantee.

G.2 Identification

To endow more explanation and generalization for the deep generative model, causal representation learning [51, 31, 38, 39, 80, 60], which captures the underlying factors and describe the latent generation process, is receiving more and more attention. One of the most classical approaches to learn the causal representation is the independent component analysis (ICA) [19, 18, 74, 73, 67, 9], in which the generation process is assumed to be a linear mixture function. However, the nonlinear ICA is a challenging task since the latent variables are not identifiable without any extra assumptions on the distribution of latent variables or the generation process [23, 80, 20, 28]. Recently, Aapo et.al [21, 22, 24, 27, 15, 14] provide the identification theories by introducing auxiliary variables, e.g. domain indexes, time indexes, and class label. These methods usually assume that the latent variables are conditionally independent and follow the exponential families. Recently, Zhang et.al [30, 67] break the restriction of exponential families assumption and propose the component-wise identification results for nonlinear ICA with a certain number of auxiliary variables. Following these theoretical results, Yao et.al [71, 72] recover time-delay latent causal variables and identify their relations from sequential data under the stationary environment and different distribution shifts. Xie et.al [67] employ the nonlinear ICA to reconstruct the joint distribution of images from different domains; and Kong et.al [30] use the component-wise identification results to solve the domain adaptation problem. However, existing identification results heavily rely on a sufficient number of domains and the too-strong monotonic transformation of latent variables, which is hard to satisfy in practice. In this paper, we propose the subspace identification results, which only rely on fewer auxiliary variables compared with component-wise identification and do not rely on any monotonic transformation assumptions.