Rethinking Domain Invariant Representations: Domain-Disentangled Invariant Representation Learning

Anonymous Author(s)

ABSTRACT

Domain generalization (DG) aims to address the challenge of generalizing across known multiple-source domains and improving model performance in unknown target domains. While many DG methods, such as domain-invariant representation (DIR) learning, excel in handling significant distribution shifts during testing but exhibit a notable trade-off, sacrificing performance on in-distribution data. This trade-off is crucial to address in real-world applications where distribution shifts are uncertain, encompassing both out-of-distribution (OOD) and nearly independent and identically distributed (IID) scenarios. In this paper, we conduct an analysis to identify the limitations of DIR and pinpoint the failure of DIR learning resulting from the neglect of non-domain-invariant and task-relevant information, termed Domain-Orthogonal Invariant (DOI) information. To address this issue, we propose **D**omain-**D**isentangled **I**nvariant **R**epresentation (**DDIR**) learning for domain generalization, aiming to retain distinctive DOI information while ensuring their utilization does not introduce unnecessary redundancy. Therefore, the key issue is how to separate DOIs and select beneficial ones. Our approach introduces information disentanglement loss, which distinguishes domain invariant information and DOI information extraction functionalities within a single backbone network. Moreover, during inference, we propose optimal DOI selection approaches for individual target samples to avoid utilizing redundant DOI information. Experimental results demonstrate the effectiveness of our approach in enhancing generalization performance in both unseen IID and OOD scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**; • **Mathematics of computing** → **Information theory**.

KEYWORDS

Domain generalization, Representation learning, Domain-invariant representation, Domain-disentangled invariant representation.

ACM Reference Format:

Anonymous Author(s). 2024. Rethinking Domain Invariant Representations: Domain-Disentangled Invariant Representation Learning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Deep neural networks are well-suited for handling independent and identically distributed (IID) data, where the training and test data share the same distribution. However, when confronted with out-of-distribution (OOD) data from unseen domains, these methods often struggle due to the shift in data distribution, leading to a significant performance drop in models. Domain generalization (DG) [42] is proposed to solve this problem. It seeks to minimize the risk of generalization against potential data transfers so that the model can achieve the best performance on unseen data.

Extracting domain-invariant representations (DIR) [1, 5, 7, 12, 16, 23, 27, 43, 45, 50] constitutes a core method within DG methods, aiming to find the invariant representations across domains that remain consistent despite domain discrepancy variations. From the perspective of information theory [38], these methods aim to identify the minimum sufficient classification information across the known multi-source domains.

Note that the DIRs do not encompass all the classification task-relevant information originating from source domains. It places more emphasis on domain invariant attributes. As illustrated in Fig.1, the shared information (blue part) among various domains within the same category is highlighted, whereas information pertinent to task but specific to the domain (pink part) is attenuated during training. We term this attenuated information as Domain-Orthogonal Invariant (DOI) information.

If the extractor preserves its ability to capture this kind of information, the evaluation of test instances without task-related DOI information results in a representation devoid of meaning (namely redundant information). However, for other instances that contain this part of the DOI information, the information obtained by the DIR extractor is insufficient. For example, giraffes manifest a yellowish-brown hue with darker patches in photographs, with color information being largely irrelevant in sketches but holding significance in paintings or cartoons. From the training perspective of DIR learning, when the source domains for extracting DIR include sketch, the color becomes non-communal information and is susceptible to being disregarded. For unseen data with color, the trained DIR extractor lacks task-relevant information. Drawing on the InfoMin principle [37], which advocates for extractors that maximize classification information while minimizing redundancy, we discern the importance of incorporating DOI information. Our analysis suggests that DOI information could prove advantageous in both OOD and IID scenarios.

To effectively utilize DOI information across all potential target domains, we propose ***D**omain-**D**isentangled **I**nvariant **R**epresentation* (DDIR) learning framework for domain generalization that partitions disjoint DOI information of each source domain and conventional domain invariant (DI) information. Utilizing partitioned storage guarantees the preservation of DOI information, facilitating

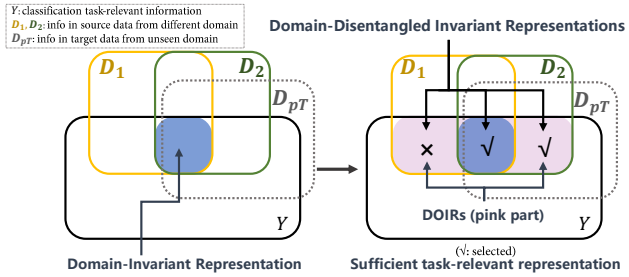


Figure 1: The DDIR learning aims to segregate and retain different DOI extraction capabilities, ensuring their utilization avoids redundancy in both IID and OOD scenarios. Compared with DIR, the DDIR extraction framework can retain sufficient task-relevant insights for unseen target instances.

inference based on the combined DI information and DOI information from the optimal source domain during testing. Achieving a complete separation between DOI information and DI information will diminish the impact of the DI information proportion on the selection of the most deterministic DOI representations. Therefore, we further design the DOI extraction loss with the objective of utilizing domain-invariant information as the repulsive element and DOI information as the attractive element.

Our contributions are summarized as follows:

- We provide theoretical insights to analyze the limitations of the current DIR methods in IID and OOD scenarios and propose Domain-Disentangled Invariant Representation (DDIR) learning.
- We present the first single-backbone framework implementation for DDIR learning, designed to separate and preserve different DOI information while ensuring its utilization does not introduce unnecessary redundancy in both IID and OOD scenarios.
- The experiments on five public datasets and analytical studies demonstrate the effectiveness of the proposed DDIR learning method and substantiate that the influence of DOI information might surpass expectations.

2 RELATED WORK

2.1 Information Bottleneck Theory

The Information Bottleneck theory [8, 38] was proposed to determine the optimal trade-off between information extraction and compression. "InfoMin principle [37]" posit that maximizing information is beneficial only when it is task-relevant. It complements the widely accepted "InfoMax principle [21]", which emphasizes capturing maximum information about the stimulus. In an ideal situation for classification, the extractor should retain all information relevant to the target classification task while minimizing the irrelevant noise. The objective of our method is to ensure the sufficiency of the representation containing relevant information for classification task when applied to classify unseen domains.

2.2 Domain Generalization

Domain-Invariant Representation (DIR) Learning, which is the prevailing approach employed in DG [42], encompasses various techniques, including adversarial training [12, 27], invariant risk minimization [1, 17] and contrastive-based methods [16, 23, 43, 45], all aimed at acquiring domain invariant representations across all existing domains and may also extend to newly generated domains.

Furthermore, we categorize prevalent DIR methods into two categories. The first category focuses on Marginal Distribution Alignment [12], aiming to achieve alignment between the representation distribution $p(z)$ across domains. The second category, Joint Distribution Alignment [17, 23, 27], aims to attain similarity across domains in the representation distributions of each class within the joint distribution $p(z, y)$. Although many DIR methods get good results in OOD tests, the process of minimizing representational discrepancies often overlooks DOI information. As a result, these DG methods show diminished effectiveness with in-distribution data, giving rise to the recognized challenge termed the IID-OOD dilemma [47]. Current research prioritizes [28, 47, 49] assessing performance on both IID and OOD data. Our focus lies in addressing the root causes of this phenomenon and proposing a solution for utilizing DOI information.

Test-time Adaptation [40] generally refers to the process of adapting a model to an unseen domain after initial generalization. During this stage, the source domain is unavailable, but the target distribution can be obtained. Our method, unlike test-time adaptation, is a typical DG method. It eliminates the need for familiarity with the target domain distribution, enabling direct result retrieval for any given target input without gradient backpropagation.

Domain-Specific Representations emerged in early domain adaptation methods [24, 51]. Sometimes, domain-specific representation includes DIR information [49]. To distinguish from these methods, we define task-relevant representation that exclude DIR information as DOIR. mDSDI [3] shares the most similar motivation with ours, however, it employs dual encoder, potentially leading to parameter complexity nearly twice that of a single backbone. Furthermore, mDSDI and its related studies based on same ideas [22], suggests a classification strategy that concatenates DIR and all domain-specific representations, introducing redundant information from representations that is irrelevant to the targets. Our method effectively addresses both of these concerns.

2.3 Contrastive Learning

In contrast to entropy minimization methods, contrastive learning methods [6, 37] embody a learning paradigm that utilizes supervised information from positive instance of an anchor instance [37]. Previous studies [41] have demonstrated that contrastive learning methods can lead to overfitting of the minimum sufficient information of positive view and there are instances where contrastive learning is employed to eliminate domain information [23]. In our work, we leverage this property to fit specific mutual information.

3 PRELIMINARY AND MOTIVATION

3.1 Problem Definition and Notations

We consider the classification problem of the source and unseen target domains with the same label set $\mathcal{Y} = \{i\}_{i=1}^{V}$, where V is the number of classes. Let there be K source domains $\mathcal{S} = \{S_i\}_{i=1}^{K}$ over the common data space $\mathcal{X}$. The k-th fully labeled source domain is denoted as $\mathcal{D}_{s_k} = \bigcup_{j=1}^{V} \mathcal{D}_{s_k}^{j} = \{(\mathbf{x}_i^k, y_i)\}_{i=1}^{M_{s_k}}$, where $y_i \in \mathcal{Y}$ stands

for the label, $\mathcal{D}_{s_k}^j$ represents the subdomain where all sample labels are j, and M_{s_k} stands for the size of the k-th source domain.

We define F as the multi-domain shared feature extraction function to obtain a high-dimensional common representation $z_c \in \mathcal{Z}$. H^0 and G^0 represent the domain-invariant represenations projection head and classifier, respectively, which are used to obtain domain-invariant representations $R_0 \in \mathcal{R}$ and generate classification results based on R_0. H^{s_k} and G^{s_k} with $s_k \in \mathcal{S}$ represent the unshared projection heads and classifiers, which are used to generate non-shared task-relevant representations $R_{\mathcal{D}_{S_k}} \in \mathcal{R}$ and perform classification based on these.

Each branch network is connected to a common feature extractor, forming a complete prediction chain $\mathcal{H} = \{h(\cdot) = G \circ H \circ F(\cdot) : \mathcal{X} \xrightarrow{F} \mathcal{Z} \xrightarrow{H} \mathcal{R} \xrightarrow{G} \triangle^{V\text{-}1}\}$, where $\circ$ denotes the composition operator and $\triangle^{V\text{-}1} = \{\boldsymbol{\pi} \in \mathbf{R}^V : \|\boldsymbol{\pi}\|_1 = 1 \wedge \boldsymbol{\pi} \geq 0\}$ denotes a (V-1)-simplex, $\wedge$ denotes the logic operation AND. We denote the information entropy contained in a vector A as $H(A)$. We use the semicolon $I(A\,;B)$ to denote the mutual information between A and B and use the comma $H(A\,,B)$ to denote the joint entropy.

3.2 Causes of DIR Learning Failures

Based on the causal generation assumption proposed by recent works [9, 25], we further abstract the causal mechanisms among the classification task Y and the information contained in data from different domains, which are denoted as $X_{\mathcal{D}_1}$ and $X_{\mathcal{D}_2}$. All classification task-relevant information contained in the multi-source domain is $I(Y; \mathcal{D}_1, \mathcal{D}_2)$ (all the colored (pink and blue) parts on the right of Fig.1). We can further decouple the information contained in data into three components: domain-invariant, domain-orthogonal invariant and task-irrelevant information.

- **Domain-invariant information** $H(R_0)$ is shared on all source domains (the blue part in Fig.1). First, let us explicitly clarify the two prevailing definitions within popular DIR learning methods. Leveraging this particular information is highly desirable, aligning with the objective of the DIR learning methods. We give a definition of the ideal DIR encoder.

Proposition 1. *(Optimal Domain-Invariant Representation Encoder). Let f_{dir} denote an optimal encoder for generating DIR (R_0). For a training set involving two domains, f_{dir} possesses the capability to exclusively capture and retain all domain-invariant information within the generated representations.*

$$H(f_{dir}(X)) = H(R_0) = I(Y; \mathcal{D}_1; \mathcal{D}_2). \tag{1}$$

- **Domain-orthogonal invariant** [1] **(DOI) information.** $H(R_{\mathcal{D}_1})$, $H(R_{\mathcal{D}_2})$ is the task-relevant information contained in $X_{\mathcal{D}_1}, X_{\mathcal{D}_2}$ except for the domain invariant information. This specific information varies for each domain, and ideally, the two sets of DOI information are disjoint. The description of the DOI information in the $\mathcal{D}_1$ (vice versa) is :

$$H(R_{\mathcal{D}_1}) = H(\mathcal{D}_1; Y|\mathcal{D}_2) = H(\mathcal{D}_1; Y|R_0). \tag{2}$$

[1]We employ the general "orthogonal" term because, ideally, the DOIs have no intersection. This term signifies a vector decomposition of task-relevant information, excluding domain-invariant information, establishing an orthogonal relationship. "Invariant" is context-dependent: in DIR, it denotes cross-domain invariance, while in DOIR, it refers to the invariance of all instances within the specific domain.

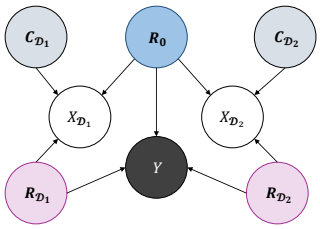


Figure 2: The causal relationship of the information in different source data $X_{\mathcal{D}_1}, X_{\mathcal{D}_2}$, DIR R_0, task-irrelevant element C, non-shared DOIR $R_{\mathcal{D}_1}, R_{\mathcal{D}_2}$ and classification task Y.

- **Task-irrelevant information** $H(C_{\mathcal{D}_k}) = H(\mathcal{D}_k|Y)$ is definitely not useful for classification task. The objective for the extractor is to minimize attention to $H(C)$, thereby reducing redundancy in representations.

Proposition 2. *(Optimal Representation Extractor for Classification Task Y under the InfoMin [37] principle and assumption 1 (Appendix A.1)). Let f_Y^* denote an optimal encoder for task Y.* ***(In IID scenarios)*** *For an instance $X_{\mathcal{D}_1}$ drawn from the source support $\mathcal{D}_1$ (and vice versa), we have*

$$I(X_{\mathcal{D}_1}; Y) = I(f_Y^*(X_{\mathcal{D}_1}); Y) = I(R_0; Y) + I(\mathcal{D}_1; Y|R_0) + I(X_{\mathcal{D}_1}; Y|\mathcal{D}_1), \tag{3}$$

(In OOD scenarios) *For an unseen instance $X_{pT} \in \mathcal{D}_{pT}$ outside the source support, we have*

$$\begin{aligned} I(X_{pT}; Y) &= I(f_Y^*(X_{pT}); Y) = I(R_0'; Y) + I(\mathcal{D}_{pT}; Y|\mathcal{D}_s) \\ &+ I(\mathcal{D}_{pT}; Y; \mathcal{D}_s|R_0) + I(X_{pT}; Y|\mathcal{D}_s, \mathcal{D}_{pT}), \end{aligned} \tag{4}$$

where $H(\mathcal{D}_s) = H(\mathcal{D}_1, \mathcal{D}_2)$ and $H(R_0') = I(Y; \mathcal{D}_1; \mathcal{D}_2; \mathcal{D}_{pT})$ represents the DIR containing the new domain. Considering the approximated nature of classification features within the same domain, the last terms in Eq.(3) and (4) tend to approach 0. In subsequent discussions, we will exclude analysis in this regard (Appendix A.1).

By comparing Proposition 1 and 2, we analyze two factors contributing to the failure of DIR learning in attaining optimal representations within the target domain:

Excess noise:

1. $H(R_0) = I(Y; \mathcal{D}_1; \mathcal{D}_2) > I(R_0; \mathcal{D}_{pT}) = H(R_0')$: In the first case, the target instances are devoid of some domain-invariant information shared between the source. These partial representations captured by the DIR extractor lacks semantic meaning and become noise.

Lack information:

2. $I(\mathcal{D}_{pT}; Y|\mathcal{D}_s) > 0$: In the second case, task-relevant information within the target data does not include in the source domain.

3. $I(\mathcal{D}_{pT}; Y; \mathcal{D}_s|R_0) > 0$: In the third case, task-relevant information within the target domain is present in the source domains. However, due to its lack of domain invariance, the process of DIR learning results in the neglect of this type of information.

However, these three cases are not all solvable. Concerning the first case, the lack of knowledge about the target domain prevents us from identifying which components of DIR may constitute noise in the target domain. Similarly, in relation to the second case, the invisibility of the target domain during the training phase inhibits the

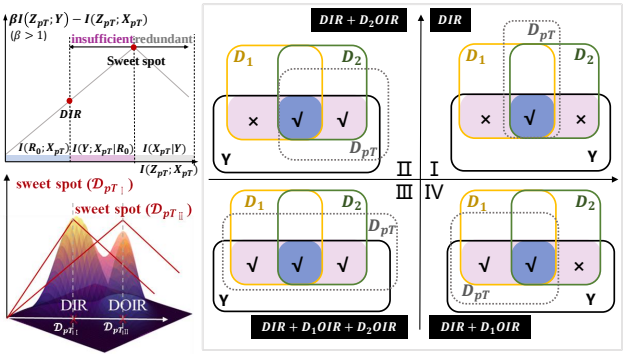


Figure 3: Four sweet spot scenarios, determined by $\mathcal{D}_T$.

acquisition of representations that remain concealed. Consequently, we propose Domain-Disentangled Invariant Representation (DDIR) learning framework primarily addresses the third case.

3.3 Sweet Spot for Different $\mathcal{D}_{pT}$

We define the ***sweet spot*** in representation extraction for multi-source domain generalization setting (top-left corner of Fig.3), where the extracted information includes only the representations relevant to the target domain and task, without any unrelated noise.

Specifically, for a classification task of a certain target sample, identifying the *sweet spot* depends on the target domain $\mathcal{D}_{pT}$ itself (bottom-left corner of Fig.3), and there are four scenarios, as depicted in the four quadrants of the right of Fig.3:

I. The first quadrant may occur due to the target domain $\mathcal{D}_{pT}$ being distant from multi-source support. From a probabilistic perspective, the likelihood of DI information existing is relatively high. When the target domain is relatively far from the source domains, DI information is likely to be task-relative information. Additionally, there is an unavoidable aspect: our knowledge encompasses every DI information is beneficial for classification tasks in source domains. We lack prior knowledge concerning the specific exclusion of internal components within the scope of DI information.

II. The second quadrant may occur when $\mathcal{D}_{pT}$ is close to $\mathcal{D}_{S_2}$.

III. The third quadrant situation, $\mathcal{D}_{pT}$ comprises a set of classification information from known source domains. This may result from factors like inherent limitations in the source domains or the emergence of complementary information, which can coexist.

IV. The fourth quadrant may occur when $\mathcal{D}_{pT}$ is close to $\mathcal{D}_{S_1}$.

Among them, scenarios II and IV correspond to IID scenarios, whereas I and III pertain to OOD. The traditional DIR learning methods perform best only in the scenario I. Scenarios II, III and IV are faced with how to exploit the $I(\mathcal{D}_{pT}; Y; \mathcal{D}_s|R_0)$, namely DOI information.

4 DOMAIN-DISENTANGLED INVARIANT REPRESENTATION LEARNING

The proposed DDIR learning framework aims to endow the model with the capability to extract DOIR while ensuring that this information does not degrade into noise for target data. Sec.4.1 addresses the rationale behind domain disentanglement. Sec.4.2 discusses how our method achieves domain disentanglement. Sec.4.3 discusses how our method accomplishes optimal domain selection, and Sec.4.4 summarizes the entire training and inference process.

4.1 Significance of Disentanglement

Therefore, the proposed solution supports the domain-disentangled information concept, namely $I(R_0; R_{\mathcal{D}_k})$=0. Besides the mentioned point that DOIR contains task-relevant information, which can better leverage potentially effective classification details, there are two additional significant reasons for using decoupled information:

Firstly, the inclusion of DIR addresses the specific scenarios in the first quadrant of the *sweet spot*.

Secondly, the omission of DI information in DOIR is essential. DIR impacts optimal domain discrimination, as it is likely a necessary piece of information. Variations in DIR content within domain-specific classifiers can lead to the model incorrectly identifying the nearest domain, resulting in a wastage of valuable information.

In appendix A.2, we compare with domain-specific representation learning [49] that lacks disentanglement.

4.2 Disentangle Strategy via Single Backbone

This subsection elucidates the process of disentangling diverse information $\{DI, \mathcal{D}_1OI, ..., \mathcal{D}_KOI\}$ into distinct branch classifiers connected by a single backbone extraction function.

First, we randomly sample N/K instances from each of the K source domains and concatenate them to form the anchor set $\{(\mathbf{x}_i^{k_i}, y_i)\}_{i=1}^N$. We then select and concatenate the corresponding positive instances batch

$$\{\{(\mathbf{x}_i^{k_i}, y_i)\}, \{(\mathbf{x}_{i+N}^{k_{i+N}}, y_{i+N})\}, \{(\mathbf{x}_{i+2N}^{k_{i+2N}}, y_{i+2N})\}\}_{i=1}^N,$$

which satisfies $y_i = y_{i+N} = y_{i+2N}$ and $k_i = k_{i+2N} \neq k_{i+N}$.

DI Branch and DI Extraction Loss. We use contrastive learning to distill domain-invariant information. The DI extraction loss function $\ell_{DI}(i, j)$ for a positive pair (i, j) is defined as

$$\ell_{DI}(i,j) = -\log \frac{\exp(\text{sim}(\mathbf{z}_i^{k_i}, \mathbf{z}_j^{k_j})/\tau)}{\sum_{l=1}^{3N} \mathbb{1}_{\mathbf{z}_l^{k_l} \in \mathbf{Z}_i^+} \exp(\text{sim}(\mathbf{z}_i^{k_i}, \mathbf{z}_l^{k_l})/\tau)}. \quad (5)$$

where $\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$ denotes normalized cosine similarity, τ is a temperature parameter, $\mathbb{1} \in \{0, 1\}$ is an indicator function evaluating to 1 if $\{\mathbf{z}_l^{k_l} \in \mathbf{Z}_i^+ | y_i \neq y_l\}$. The positive instances are chosen only from distinct domains of the same class. This configuration is influenced by the observation that contrastive learning leads to overfitting on mutual information from positive pairs [23, 41]. And negative instances are all instances of different classes with the anchor in the batch. According to the batch design, the $\mathcal{L}_{DI}$ extracts the DIR for one batch can be formulated as:

$$\mathcal{L}_{DI}(\theta_F; \theta_{G_0}; \theta_{H_0}) = \frac{1}{4N} \Sigma_{i=1}^{2N} [\ell_{DI}(i, i+N) + \ell_{DI}(i+N, i)]. \quad (6)$$

Simultaneously, in order to segregate redundant information, we compute the cross-entropy $\mathcal{L}_{CE}$ across all samples. mDSDI [3] employs a dual-backbone network, wherein one backbone extracts DIR information, and another backbone, in conjunction with domain classifiers, captures task-relevant information for each domain. However, achieving this with a single-backbone network proves challenging due to the influence of domain-specific information

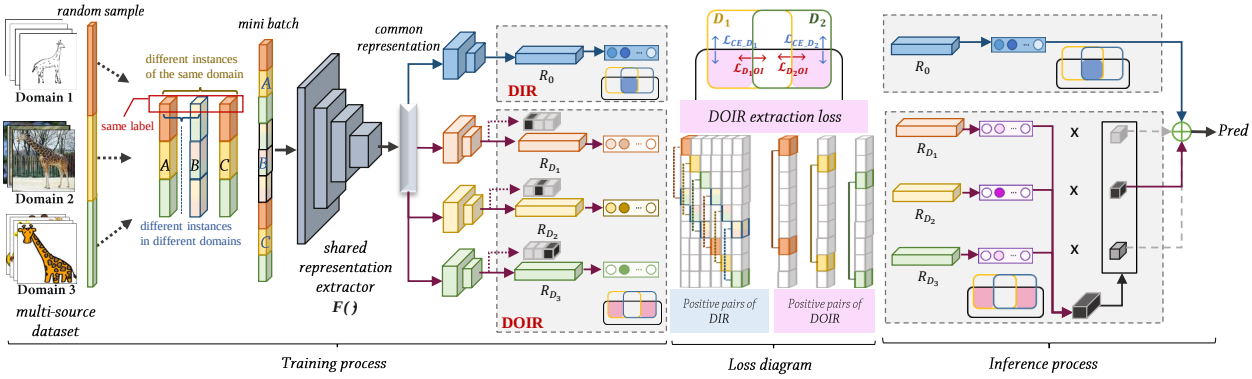


Figure 4: An overall pipeline of the proposed Domain-Disentangled Invariant Representation learning framework.

on the extraction of DI information. Fitting mutual information on representations extracted by branches can separate information from another perspective, creating repulsion between branches.

DOI Branches and DOI Extraction Loss. For DOI branches, we employ a DOI extraction loss function $\ell_{\mathcal{D}_k OI}$ different from ℓ_{DI} :

$$\ell_{\mathcal{D}_k OI}(i,j) = -\log \frac{\exp(\text{sim}(\mathbf{z}_i^{k_i}, \mathbf{z}_j^{k_j})/\tau)}{\sum_{l=1}^{3N} \mathbb{1}_{\mathbf{z}_l^{k_l} \in \mathbf{Z}_i^{++}} \exp(\text{sim}(\mathbf{z}_i^{k_i}, \mathbf{z}_l^{k_l})/\tau)}. \quad (7)$$

where positive examples are exclusively chosen from the same domain as the anchor, focusing on instances of the same class. For negative examples $\{\mathbf{z}_l^{k_l} \in \mathbf{Z}_i^{++} | y_i \neq y_l \text{ or } k_i \neq k_l\}$, we encompass not only instances of different classes for the anchor (as traditional contrastive learning) but also instances of the same class from different domains. The inclusion of this additional component aims to further eliminate DI information within the branches. The function $\mathcal{L}_{\mathcal{D}_k OI}$ extracts the DOIR for one batch can be formulated as:

$$\mathcal{L}_{\mathcal{D}_k OI}(\theta_F; \theta_{G_k}; \theta_{H_k}) = \frac{K}{2N} \Sigma_{i=1+N(k-1)/K}^{Nk/K} [\ell_{\mathcal{D}_k OI}(i, i+2N) + \ell_{\mathcal{D}_k OI}(i+2N, i)]. \quad (8)$$

The loss of separating redundant information from DOIR, denoted as $\mathcal{L}_{CE\_\mathcal{D}_k}$, is based on the original cross-entropy but specifically perform cross-entropy computation and back-propagation only with respect to instances from the k-th source domain.

4.3 Strategies for DOI Selection

During the inference stage, determining the proximity of a single sample to a source domain without relying on the entire target domain distribution poses a significant challenge. Addressing this issue, we propose two approaches: ODS and IDS.

Guided by the assumption posited in [49] that samples closer to a specific source domain should yield more confident predictions on the respective classifier, we introduce the Optimal Prediction Entropy (OPE) strategy

$$\arg\max_{\{i \in \mathcal{Y}, s_k \in \mathcal{S}\}} \triangle_{s_k}^{V\text{-}1}[i], \quad (9)$$

where $\triangle_{s_k}^{V\text{-}1}[i]$ denotes the probability of predicting the i-th class. However, we observe that selecting the DOI classifier based on the most confident prediction introduces significant randomness. Moreover, it becomes challenging to discern whether the scenario corresponds to the first or third quadrant in Fig.3.

Therefore, we propose to further integrate a domain selection strategy, domain classifiers (DC), based on relative relationships among source domains for the domain selection. Note that domain classifiers $\{D_k(\cdot): \mathcal{R}_{D_k} \xrightarrow{D_k} \triangle_{D_k}^1\}$ do not transmit information back to the shared extractor. This deliberate omission stems from our goal to solely obtain evaluations on whether the representation belongs to the k-th source domain. Transmitting such information to the extractor is unnecessary and disrupt the model's ability to extract domain-invariant representations. The training function for DC is

$$\mathcal{L}_{d\_k}(\theta_{D_k}; \mathbf{x}_i^k) = \mathcal{L}_{CE}(\mathbf{x}_i^k, y_d), \qquad y_d = \begin{cases} 1 & \mathbf{x}_i^k \in \mathcal{D}_{s_k}, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

where y_d labels are set to 1 for instances from the corresponding domain, and set to 0 for instances from other domains.

In light of these two fundamental strategies and the consideration of the inclusion or exclusion of the third quadrant scenarios, we classify DDIR into two inference approaches:

Optimal DOI Selection (ODS) identifies and singles out the most optimal domain by evaluating predefined criteria. This approach exclusively opts for domains that meet the predetermined criteria for optimality, discarding any domains that fall short of these conditions. For the ODS approach, there's a risk of choosing the suboptimal domain or selecting DOI information when it's clearly in the first quadrant of Fig.3. To address this, we propose filtering out the first quadrant scenarios and eliminating suboptimal domains using DC strategy and then select the most confident prediction according to the OPE strategy.

Inclusive DOI Selection (IDS) is designed to encompass a variable and potentially unlimited number of domains that satisfy specified criteria. Unlike ODS, IDS does not impose a limit on the number of selected DOI information, embracing all domains that meet the predetermined conditions. IDS considers the third quadrant scenarios of Fig.3, but it tends to choose more suboptimal domains. To overcome this, we combine DC and OPE strategies: first, we eliminate opinions deemed non-local domains by DC. Then, we sum the category confidence and domain confidence products of the remaining domains.

To be specific, for a target data, the inference function of ODS as

$$\arg\max_{y_i \in \mathcal{Y}} \left[(1\text{-}\alpha_1)\, \triangle_0^{V\text{-}1} + \alpha_1 \max_k \eta_H(\triangle_{D_k}^1, \frac{1}{2}, O)\, \triangle_{s_k}^{V\text{-}1} \right], \quad (11)$$

The inference function of IDS as

$$\arg\max_{y_i \in \mathcal{Y}} \left[(1\text{-}K\alpha_2)\ \triangle_0^{V\text{-}1} + \alpha_2 \sum_{k=1}^{K} \eta_H(\triangle_{D_k}^1, \tfrac{1}{2}, I)\ \triangle_{s_k}^{V\text{-}1} \right], \quad (12)$$

where, $\eta_H(\triangle_{D_k}^1, \lambda, opt) = \begin{cases} 1 & \triangle_{D_k}^1[1] \geq \lambda \wedge opt{=}O, \\ \triangle_{D_k}^1[1] & \triangle_{D_k}^1[1] \geq \lambda \wedge opt{=}I, \\ 0 & \text{otherwise.} \end{cases}$ and $\triangle_{D_k}^1[1]$ represents the softmax value indicating the probability of the target instance belonging to domain D_k and $\wedge$ denotes the logical AND.

4.4 Overall Training

In the training process, the initial focus is on training the entire network to enable the enrichment of diverse information within distinct classifier branches.

$$\mathcal{L}_{tr}(\theta_F; \{\theta_{G_r,H_r}\}_{r=0}^K) = (1\text{-}K\alpha)(\mathcal{L}_{DI} + \mathcal{L}_{CE}) + \alpha \sum_{k=1}^{K} (\mathcal{L}_{\mathcal{D}_k OI} + \mathcal{L}_{CE\_\mathcal{D}_k}), \quad (13)$$

where α is an adjustable hyper-parameter that regulates the proportion of the DI and DOI information in the common representation z_c extracted by the backbone network.

Subsequently, dedicated training is applied to the domain discriminative classifiers.

$$\mathcal{L}_d(\theta_{D_1}, ..., \theta_{D_K}) = \sum_{k=1}^{K} \mathcal{L}_{d\_k}. \quad (14)$$

For inference, we determine the predicted label for each individual data using ODS approach Eq.(11) or IDS approach Eq.(12).

5 EXPERIMENTAL EVALUATION

5.1 Experiment Settings

5.1.1 Datasets. We conduct comprehensive experiments on public datasets, including RotatedMNIST [13], PACS[19], OfficeHome [32], VLCS [10] and DomainNet [30]. See Appendix C for details.

5.1.2 Configuration. For a fair comparison with the other arts, we employ the MNIST-CNN [14] for RotatedMNIST and ImageNet-pretrained ResNet-50 [15] for others. Our implementation is based on DomainBed [14], incorporating consistent practices in data splits and evaluation strategies. The design of the projection heads and classifiers follows that of [45], albeit with varying quantities (detailed specifications in Sec.4). The size of high-dimensional representation z_c is 2048, the size of DIR or DOIR is 512. We resize all the instances to 224×224. Following [6, 44], we use the temperature τ of 0.07. For hyperparameter α, the proportion of DOIR in the final decision should be smaller than that of DIR; thus, α_1 is set to 0.2 and $\alpha = \alpha_2$ is set to {0.16, 0.08} for {3, 5} source domains, respectively.

5.1.3 Baselines. The majority of comparative baselines utilize the DIR learning or are derived from the same underlying idea. (Appendix B) Empirical Risk Minimization (**ERM**) [39], Domain-adversarial neural networks (**DANN**) [12], Deep Correlation Alignment (**CORAL**) [36], Distributionally Robust Optimization (**GroupDRO /GDRO**) [33], Invariant Risk Minimization (**IRM**) [1], Risk Extrapolation (**VREx**) [17], Adaptive Risk Minimization (**ARM**) [46], Style Agnostic Networks (**SagNet**) [26], **AND-mask** [29] Smoothed-AND masking (**SAND-mask**) [34] **Fish** regularization [35] **Fishr** regularization [31] Cross contrasting feature perturbation (**CCFP**) [18], Direct-Effect Risk Minimization (**DERM**) [20], Domain-Specific Risk Minimization (**DRM**) [49], meta-Domain Specific-Domain Invariant (**mDSDI**) [3], and Rationale invariance (**RIDG**) [5].

5.2 Results and Discussion

Leave-one-domain-out (LODO) results (OOD scenarios). We strictly follow the LODO evaluation with the codebase of Domainbed: select one domain from the dataset as the unseen target domain, and set the remaining domains as the source domains. We show the LODO results on 5 datasets with ResNet-50 in Tab.1. The results represent the average accuracy across all domains used as target domain. The proposed method with IDS and ODS approaches achieves optimal and suboptimal performance in all datasets.

In-distribution results (IID scenarios). We show the In-distribution and HM [47] (*i.e.* harmonic mean of LODO (ODS approach) and ID results, $HM = \frac{2x_1x_2}{x_1+x_2}$) results on PACS, Office-Home and VLCS in Tab.2. We also present covariate shift metrics $\mathcal{M}_{\text{cov}}$ for different datasets, as provided by [48]. Our method outperforms approaches based on different strategies, including gradient operation-based Fish, distributionally robust optimization-based GroupDRO, feature disentanglement-based SagNet, meta-based ARM and DIR-based DANN, CORAL, IRM, VRex. DDIR outperforms the majority of DIR methods, which often fall short of the performance achieved by ERM in IID scenarios. This observation underscores the negative impact of lacking DOI information on prediction tasks.

DDIR exhibits superior performance across all datasets in terms of LODO, ID and HM metrics, competing well against other DG methods in OOD scenarios and exhibiting obvious improvements in IID scenarios. In datasets like Office-Home, where $\mathcal{M}_{\text{cov}}$ is lower and the target domain is closer to the source, DOI information proves beneficial. This leads to notable improvements, which is intuitive. Moreover, as the number of domains increases, the probability of source domains providing valuable information rises, leading to more significantly improved performance.

5.3 Analytical Experiments

Proof-of-Concept Experiment First, we perform a basic verification of the DDIR learning idea through a proof-of-concept experiment. We use the PACS dataset with two separate ResNet-18 backbone. The two non-shared backbones are trained in an ideal decoupling case where one backbone is connected to a domain-invariant predictor and the other backbone is connected to the domain-orthogonal invariant predictors. Since features acquired via direct contrastive learning are not decoupled, we merge them directly in the final output using expert opinions. By adjusting the weights of the DIR expert (λ) and the DOIR expert ($1-\lambda$), we assess the significance of DOI information. We calculate average scores based on rankings, allocating 10 points to the top position and decreasing progressively to 0 points for the last position. The average score of the first 1000 iterations over 3 trials is shown in Fig.5(a). The best result is around $\lambda = 0.2$, *i.e.* weight(DIR: DOIR) $\approx$ 4:1. This goes far beyond the assumption that DOI information can be ignored. This experiment used two backbones, so it is not fair to compare this with other DG methods. DDIR find a way to decouple DI and DOI information using single backbone, mitigating this concern.

Quantitative Ablation Results. We conducted ablation studies or replaced loss designs on both training loss and inference approaches on Office-Home with ResNet-50. The results are presented in Tab.3. **For training**, replacing $\mathcal{L}_{CE_{\mathcal{D}}k}$ with $\mathcal{L}_{CE}$ enhances performance in ERM (T6,T7) but degrades it in DDIR (T1,T4). This suggests

Table 1: Leave-one-domain-out results on Domainbed with training-domain model selection. The best performing in bold.

Algorithm	RotatedMNIST	PACS	OfficeHome	VLCS	DomainNet	Avg.
ERM [39]	98.0 ± 0.0	85.5 ± 0.2	66.5 ± 0.3	77.5 ± 0.4	40.9 ± 0.1	73.7
DANN [12]	97.8 ± 0.1	83.6 ± 0.4	65.9 ± 0.6	78.6 ± 0.4	38.3 ± 0.1	72.8
CORAL [36]	98.0 ± 0.1	86.2 ± 0.3	68.7 ± 0.3	78.8 ± 0.6	41.5 ± 0.1	74.6
GroupDRO [33]	98.0 ± 0.0	84.4 ± 0.8	66.0 ± 0.7	76.7 ± 0.6	33.3 ± 0.2	71.7
IRM [1]	97.7 ± 0.1	83.5 ± 0.8	64.3 ± 2.2	78.5 ± 0.5	33.9 ± 2.8	71.6
VREx [17]	97.9 ± 0.1	84.9 ± 0.6	66.4 ± 0.6	78.3 ± 0.2	33.6 ± 2.9	72.2
ARM [46]	98.2 ± 0.1	85.1 ± 0.4	64.8 ± 0.3	77.6 ± 0.3	35.5 ± 0.2	72.2
SagNet [26]	98.0 ± 0.0	86.3 ± 0.2	68.1 ± 0.1	77.8 ± 0.5	40.3 ± 0.1	74.1
AND-mask [29]	97.6 ± 0.1	84.4 ± 1.3	65.6 ± 0.3	78.1 ± 1.3	37.2 ± 0.9	72.6
SAND-mask [34]	97.4 ± 0.1	84.6 ± 1.2	65.8 ± 0.5	77.4 ± 0.7	32.1 ± 0.7	71.5
Fish [35]	97.9 ± 0.1	85.5 ± 0.1	68.6 ± 0.3	77.8 ± 0.3	42.7 ± 0.4	74.5
Fishr [31]	97.8 ± 0.0	85.5 ± 0.1	67.8 ± 0.4	77.8 ± 0.4	41.7 ± 0.2	74.1
CCFP [18]	97.8 ± 0.1	**86.6 ± 0.2**	68.9 ± 0.1	78.9 ± 0.3	41.2 ± 0.0	74.7
DERM [20]	97.6 ± 0.1	84.8 ± 0.5	65.7 ± 0.6	77.9 ± 0.5	41.0 ± 0.2	73.4
RIDG† [5]	97.9 ± 0.1	84.7 ± 0.2	68.6 ± 0.2	77.8 ± 0.4	41.9 ± 0.3	74.2
DDIR-IDS (ours)	**98.6 ± 0.1**	86.4 ± 0.3	**69.7 ± 0.2**	**79.3 ± 0.7**	42.7 ± 0.0	75.3
DDIR-ODS (ours)	98.5 ± 0.0	**86.6 ± 0.6**	**69.7 ± 0.2**	79.2 ± 1.1	**43.0 ± 0.1**	**75.4**

Table 2: In-distribution (ID) and harmonic mean (HM) [47] results on PACS, OfficeHome and VLCS datasets.

	PACS ($\mathcal{M}_{\text{cov}} = 0.33$)						Office-Home ($\mathcal{M}_{\text{cov}} = 0.24$)						VLCS ($\mathcal{M}_{\text{cov}} = 0.26$)					
$\mathcal{D}_T$	A	C	P	S	**ID**	**HM**	Ar	Cl	Pr	Rw	**ID**	**HM**	C	L	S	V	**ID**	**HM**
ERM	$97.6^{\pm 0.3}$	$96.4^{\pm 1.5}$	$95.3^{\pm 1.2}$	$96.3^{\pm 0.1}$	96.2	90.5	$82.9^{\pm 0.3}$	$82.8^{\pm 0.7}$	$78.4^{\pm 0.7}$	$80.0^{\pm 0.8}$	81.0	73.0	$78.2^{\pm 3.3}$	$87.8^{\pm 9.0}$	$86.3^{\pm 10.}$	$83.3^{\pm 11.}$	83.9	80.6
DANN	$89.6^{\pm 2.3}$	$92.2^{\pm 0.4}$	$93.0^{\pm 1.4}$	$91.8^{\pm 3.5}$	91.6	87.4	$80.2^{\pm 0.6}$	$79.8^{\pm 1.0}$	$75.9^{\pm 0.8}$	$76.0^{\pm 0.6}$	78.0	71.4	$80.3^{\pm 0.3}$	$86.3^{\pm 1.4}$	$84.4^{\pm 1.6}$	$82.9^{\pm 0.8}$	83.5	81.0
CORAL	$96.7^{\pm 0.7}$	$95.5^{\pm 0.9}$	$95.7^{\pm 0.8}$	$96.4^{\pm 0.8}$	96.1	90.9	$84.1^{\pm 0.7}$	$83.8^{\pm 1.9}$	$78.6^{\pm 3.4}$	$81.5^{\pm 0.2}$	82.0	74.8	$80.4^{\pm 1.7}$	$88.8^{\pm 0.3}$	$86.1^{\pm 1.9}$	$85.5^{\pm 1.6}$	85.2	81.9
IRM	$95.9^{\pm 1.6}$	$94.2^{\pm 2.5}$	$94.3^{\pm 1.0}$	$94.5^{\pm 1.8}$	94.7	88.7	$77.3^{\pm 0.8}$	$77.5^{\pm 0.9}$	$72.1^{\pm 1.0}$	$73.0^{\pm 2.1}$	75.0	69.2	$76.9^{\pm 2.9}$	$88.2^{\pm 8.9}$	$85.3^{\pm 9.8}$	$77.3^{\pm 1.0}$	81.9	80.2
VREx	$96.3^{\pm 1.0}$	$96.2^{\pm 0.6}$	$95.8^{\pm 0.9}$	$96.7^{\pm 1.0}$	96.2	90.2	$82.8^{\pm 1.3}$	$83.9^{\pm 1.0}$	$78.5^{\pm 1.3}$	$79.2^{\pm 1.1}$	81.0	73.0	$79.9^{\pm 0.6}$	$88.7^{\pm 1.2}$	$84.0^{\pm 2.0}$	$84.1^{\pm 0.6}$	84.2	81.1
GDRO	$97.0^{\pm 0.7}$	$96.0^{\pm 0.7}$	$96.1^{\pm 0.3}$	$96.0^{\pm 1.1}$	96.3	90.0	$82.2^{\pm 0.5}$	$81.9^{\pm 1.5}$	$77.6^{\pm 0.2}$	$80.4^{\pm 0.1}$	80.5	72.5	$78.6^{\pm 1.9}$	$88.3^{\pm 0.9}$	$85.9^{\pm 1.8}$	$84.8^{\pm 1.6}$	84.4	80.4
SagNet	$97.7^{\pm 0.3}$	$96.2^{\pm 1.6}$	$95.5^{\pm 1.1}$	$96.9^{\pm 0.4}$	96.4	91.1	$81.9^{\pm 1.1}$	$81.9^{\pm 0.7}$	$76.9^{\pm 1.1}$	$78.8^{\pm 0.9}$	79.9	73.5	$79.7^{\pm 1.5}$	$88.1^{\pm 0.9}$	$86.0^{\pm 1.1}$	$84.7^{\pm 1.1}$	84.6	81.1
ARM	$96.8^{\pm 0.1}$	$95.9^{\pm 0.6}$	$95.7^{\pm 0.2}$	$96.2^{\pm 0.5}$	96.1	90.3	$80.1^{\pm 0.5}$	$79.2^{\pm 0.7}$	$75.3^{\pm 0.7}$	$77.4^{\pm 0.5}$	78.0	70.8	$79.6^{\pm 1.4}$	$85.4^{\pm 4.4}$	$85.4^{\pm 1.2}$	$85.2^{\pm 0.3}$	83.9	80.6
Fish	$97.1^{\pm 0.8}$	$96.7^{\pm 0.9}$	$96.4^{\pm 0.6}$	$97.3^{\pm 0.5}$	96.9	90.8	$83.4^{\pm 0.7}$	$82.2^{\pm 1.8}$	$78.2^{\pm 1.0}$	$79.4^{\pm 0.5}$	80.8	74.2	**$81.1^{\pm 0.8}$**	$88.8^{\pm 1.1}$	$86.2^{\pm 0.9}$	$85.2^{\pm 1.5}$	85.3	81.4
DDIR	**$97.7^{\pm 0.1}$**	**$97.0^{\pm 0.1}$**	**$96.8^{\pm 0.1}$**	**$97.7^{\pm 0.3}$**	**97.3**	**91.6**	**$86.3^{\pm 0.4}$**	**$85.7^{\pm 0.6}$**	**$82.3^{\pm 1.0}$**	**$84.0^{\pm 0.7}$**	**84.6**	**76.5**	$80.9^{\pm 0.5}$	**$89.5^{\pm 0.5}$**	**$87.2^{\pm 0.4}$**	**$86.3^{\pm 0.4}$**	**86.0**	**82.5**

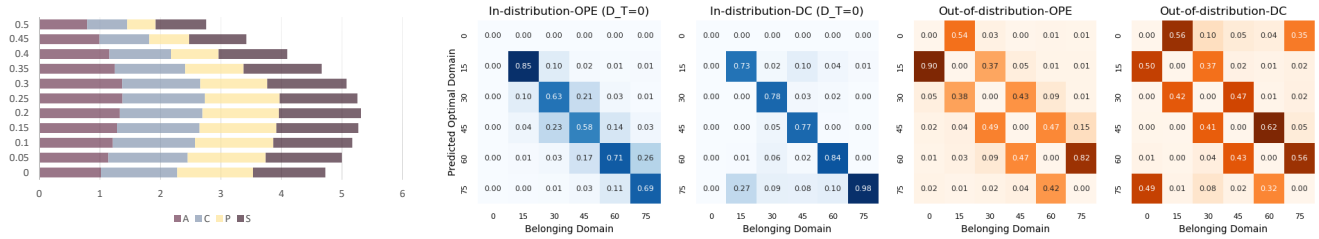
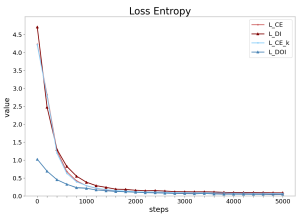


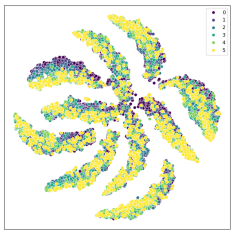
Figure 5: (Left) Proof-of-concept experiment for dual-encoder. (Right) Optimal domain selection for different strategies.

that for DOI branches, non-native domain representations in latent space can be beneficial amidst chaos. Adding same-domain, same-class samples to the positive set of $\mathcal{L}_{DI}$ (T2) or removing same-domain samples from the negative set of $\mathcal{L}_{\mathcal{D}_k OI}$ (T3) both lead to performance decline. **For the inference process**, we opted for 5 inference strategies. (I5) We concatenated all DOIRs and DIR and fed them into the classifier for training and prediction as [3, 22]. We find that this strategy underperforms all approaches based on DOIR selection. The reason lies in the introduction of redundant information from irrelevant source domains. Additionally, we compare the efficacy of individual OPE (I3) or DC (I4) strategies in selecting DOIR. Both strategies used in isolation perform lower than the ODS (I1) and IDS (I2) approaches proposed in DDIR.

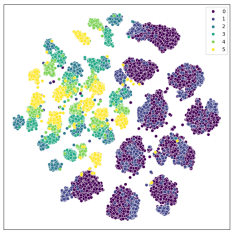
Accuracy of Optimal DOI Selection We further examine the characteristics of two fundamental domain selection strategies (OPE & DC). We conduct In-distribution and out-of-distribution tasks on RotatedMNIST and record the optimal domain selected for all



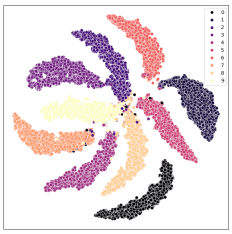
(a) Loss curves from the LODO experiment on Office-Home (unseen: Art).



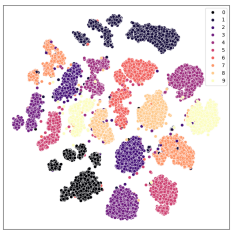
(b) DIR R_0 (color-coded according to domain).



(c) DOIR R_1 of $\texttt{Rm}_{15}$ (color-coded according to domain).



(d) DIR R_0 (color-coded according to label).



(e) DOIR R_1 of $\texttt{Rm}_{15}$ (color-coded according to label).

Figure 6: (a) Loss curves. (b-e) t-SNE visualization of representation in different branches on RMNIST (unseen domain: $\texttt{Rm}_0$).

Table 3: Ablation Results. (→: Replacement; −: Deletion)

	Training loss (inference: ODS)	Ar	Cl	Pr	Rw	**AVG.**
T1	training loss (default)	65.6	56.2	77.8	79.2	**69.7**
T2	default $\mathcal{L}_{DI} \rightarrow \mathcal{L}^*_{DI}$	63.0	55.1	75.9	77.9	**68.0**
T3	default $\mathcal{L}_{\mathcal{D}_k OI} \rightarrow \mathcal{L}^*_{\mathcal{D}_k OI}$	63.0	56.4	76.6	79.0	**68.7**
T4	default $\mathcal{L}_{CE\_\mathcal{D}_k} \rightarrow \mathcal{L}_{CE}$	65.2	54.7	76.5	80.1	**69.1**
T5	default $-\mathcal{L}_{\mathcal{D}_k OI} - \mathcal{L}_{CE\_\mathcal{D}_k}$	64.8	53.9	77.1	78.5	**68.6**
T6	default $-\mathcal{L}_{\mathcal{D}_k OI} - \mathcal{L}_{DI}$	64.1	53.2	75.9	76.2	**67.3**
T7	$\mathcal{L}_{CE}$ (multi-branch backbone)	62.6	54.8	75.8	78.6	**67.9**
T8	$\mathcal{L}_{CE}$ (unable to infer using ODS)	61.3	52.4	75.8	76.6	**66.5**
	Inference (training loss: default)	Ar	Cl	Pr	Rw	**AVG.**
I1	ODS	65.6	56.2	77.8	79.2	**69.7**
I2	IDS	64.7	57.9	77.0	79.2	**69.7**
I3	OPE only	63.0	55.7	78.3	78.8	**69.0**
I4	DC only	63.9	56.0	77.5	79.5	**69.2**
I5	concatenation	63.9	54.5	76.2	78.2	**68.2**

Table 4: Comparison between mDSDI and DDIR (w/ ODS).

	RMNIST	PACS	OfficeHome	VLCS	DNet
mDSDI(dual-encoder)	98.0	86.2	69.2	79.0	42.8
DDIR(**single**-encoder)	**98.5**	**86.6**	**69.7**	**79.2**	**43.0**

Table 5: Marginal Effect on Parameter Count.

# of Params (M)	ResNet-18	ResNet-50	ResNet-101
+ D_k	+ 5.12 E-04	+ 5.12 E-04	+ 5.12 E-04
+ R_k	+ 0.528	+ 1.315	+ 1.315
+ Encoder (mDSDI)	+ 11.18	+ 23.51	+ 42.50

targets, as shown in Fig. 5(b). For In-distribution tasks, both strategies exhibit similar and constructive outcomes in the vicinity of the optimal domain. For out-of-distribution tasks, DC demonstrates a noticeable degradation in performance, particularly in the edge unknown domain $\texttt{Rm}_0$, $\texttt{Rm}_{75}$. Visualizations show divergent error patterns between the two methods. OPE relies on predictive confidence, revealing a tendency for errors in domains closely resembling the optimal domain. Conversely, DC founded on the relative relationships across source domains, exhibits more errors at the extremes of the edge domain, namely $\texttt{Rm}_0$, $\texttt{Rm}_{75}$ domains, with fewer errors are observed in domains surrounding the optimal domain. This result also highlights the rationale behind the design of the two domain selection approaches. ODS and IDS both employ the DC strategy to eliminate suboptimal domains, and utilize the OPE strategy to identify the most confidently predicted expert opinions.

Loss curves. We plotted the loss curves for four entropy variations, as depicted in Fig.6(a). When testing on the 'Art' domain of Office-Home, all losses showed rapid and stable convergence.

Representation Space Visualization. From Fig.6, the DIR branch exhibits a high degree of domain-invariance in representations. In Fig.6(b), multi-source domain and unseen target domain $\texttt{Rm}_0$ are evenly mixed in each category and Fig.6(d) shows clear category distinctions, albeit with some regions in the center where multiple categories are concentrated. As for the DOIR branch of $\texttt{Rm}_{15}$ domain, as shown in Fig.6(c), it successfully distinguishes $\texttt{Rm}_{15}$ domain from $\texttt{Rm}_{30}$, $\texttt{Rm}_{45}$, $\texttt{Rm}_{60}$ and $\texttt{Rm}_{75}$ domains. Moreover, within $\texttt{Rm}_{15}$, the classification performance (Fig.6(e)) shows clearer segmentation for each category block.

Marginal Effects of DOI Branches. In Tab.4, the performance of DDIR with a single encoder exceeds that of mDSDI, which requires two encoders. As the complexity of the backbone increases, the model's parameter count may become twice that of single backbone network methods (Tab.5). This demonstrates DDIR's efficiency in resource utilization. With an increased number of source domains, the DDIR structure generates additional branches. We further present the impact of adding new branches on the total model parameter count across different backbones, as shown in Tab.5. As it stands, the scalability appears promising.

6 CONCLUSION

In this paper, we proposed domain-disentangled invariant representation (DDIR) learning for domain generalization. We theoretically and empirically verify that the DIR may not be the most sufficient representation, as its sufficiency depends on the similarity between the unseen target domain and the visible source domains. In pursuit of this objective, we propose to separate DOIR and design a new single-backbone network for partitioning storage. Subsequently, during testing, without the need for parameter backpropagation and target distribution information, we can invoke the optimal combination of DOI opinion and infer the label by combining domain-invariant representation information. The extensive experiments outperform state-of-the-art methods demonstrate the effectiveness of the DDIR learning method in both IID and OOD scenarios.

REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[2] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *European Conference on Computer Vision*. 456–473.

[3] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. 2021. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems* 34 (2021), 21189–21201.

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems* 34 (2021), 22405–22418.

[5] Liang Chen, Yong Zhang, Yibing Song, Anton Van Den Hengel, and Lingqiao Liu. 2023. Domain generalization via rationale invariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1751–1760.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[7] Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. 2020. Estimating generalization under distribution shifts via domain-invariant representations. *International conference on machine learning*.

[8] Felix Creutzig, Amir Globerson, and Naftali Tishby. 2009. Past-future information bottleneck in dynamical systems. *Physical Review E* (2009).

[9] Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, and Xinmei Tian. 2023. Moderately Distributional Exploration for Domain Generalization. *International Conference on Machine Learning* (2023).

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

[11] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. *Proceedings of the International Conference on Learning Representations* (2020).

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* (2016), 2096–2030.

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

[14] Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. *Proceedings of the International Conference on Learning Representations* (2021).

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9619–9628.

[17] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*. PMLR, 5815–5826.

[18] Chenming Li, Daoan Zhang, Wenjian Huang, and Jianguo Zhang. 2023. Cross contrasting feature perturbation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1327–1337.

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*. 5542–5550.

[20] Yuhui Li, Zejia Wu, Chao Zhang, and Hongyang Zhang. 2023. Direct-Effect Risk Minimization for Domain Generalization. *arXiv preprint arXiv:2211.14594* (2023).

[21] Ralph Linsker. 1988. Self-organization in a perceptual network. *Computer* 21, 3 (1988), 105–117.

[22] Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. 2022. Domain-invariant feature exploration for domain generalization. *TMLR* (2022).

[23] Yuan Ma, Yiqiang Chen, Han Yu, Yang Gu, Shijie Wen, and Shuai Guo. 2023. Letting Go of Self-Domain Awareness: Multi-Source Domain-Adversarial Generalization via Dynamic Domain-Weighted Contrastive Transfer Learning. In *ECAI 2023*. IOS Press, 1664–1671.

[24] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2008. Domain adaptation with multiple sources. *Advances in neural information processing systems* 21 (2008).

[25] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. 2021. Representation learning via invariant causal mechanisms. *Proceedings of the International Conference on Learning Representations* (2021).

[26] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8690–8699.

[27] A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. 2021. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems* 34 (2021), 5264–5275.

[28] Yulei Niu and Hanwang Zhang. 2021. Introspective distillation for robust question answering. *Advances in Neural Information Processing Systems* 34 (2021), 16292–16304.

[29] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. 2020. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329* (2020).

[30] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.

[31] Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2022. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*. PMLR, 18347–18377.

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *European conference on computer vision*. Springer, 213–226.

[33] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Proceedings of the International Conference on Learning Representations* (2020).

[34] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. 2021. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266* (2021).

[35] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. 2022. Gradient matching for domain generalization. *Proceedings of the International Conference on Learning Representations* (2022).

[36] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 443–450.

[37] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems* 33 (2020), 6827–6839.

[38] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *The 37th annual Allerton Conference on Communication, Control, and Computing* (2000).

[39] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* (1999), 988–999.

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. TENT: fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations*.

[41] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16041–16050.

[42] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[43] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*. Springer.

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3733–3742.

[45] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. 2022. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7097–7107.

[46] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. 2021. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems* 34 (2021), 23664–23678.

[47] Min Zhang, Junkun Yuan, Yue He, Wenbin Li, Zhengyu Chen, and Kun Kuang. 2023. MAP: Towards Balanced Generalization of IID and OOD through Model-Agnostic Adapters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11921–11931.

[48] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. 2023. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16036–16047.

[49] Yi-Fan Zhang, Jindong Wang, Jian Liang, Zhang Zhang, Baosheng Yu, Liang Wang, Dacheng Tao, and Xing Xie. 2023. Domain-Specific Risk Minimization for Domain Generalization. In *Proceedings of the 29th ACM SIGKDD Conference on*

Knowledge Discovery and Data Mining. 3409–3421.
[50] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13025–13032.
[51] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. 2020. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems* 32, 4 (2020), 1713–1722.

A APPENDICES

A.1 Additional details for Section 3.2 Analysis

Basic properties of mutual information, for any variables X, Y, Z:

P.0. $H(X) \geq I(X;Y) \geq 0$

P.1. $I(X,Y;Z) = I(X;Z) + I(Y;Z|X)$

P.2. $I(X;Y) = I(X;Y;Z) + I(X;Y|Z)$

P.3. $I(\mathcal{D};Y) = I(X_1;...;X_N;Y) \qquad \mathcal{D} = \{X_i\}_{i=1}^{N}$

In line with assumptions made in previous studies (e.g. [11]), when the multi-source domain assumption is built upon the hypothesis expressed as

$$I(X_{\mathcal{D}_1};X_{\mathcal{D}_2};Y) = I(X_{\mathcal{D}_1};Y) = I(X_{\mathcal{D}_2};Y) \tag{15}$$

DIR serves as the optimal solution for known supervised information. However, it is observed that this equation should be formulated as an inequality due to the domain-orthogonal information, represented as

$$\underbrace{I(X_{\mathcal{D}_1};X_{\mathcal{D}_2};Y)}_{\text{Generic label information}} \leq \underbrace{I(X_{\mathcal{D}_1};Y)}_{\text{Label information in } \mathcal{D}_1} \quad or \quad \underbrace{I(X_{\mathcal{D}_2};Y)}_{\text{Label information in } \mathcal{D}_2}$$

In a specific example, the left-hand side of the inequality may represent the generic characteristics of a giraffe, while the right-hand side corresponds to cartoon giraffes or hand-drawn illustrations of giraffes. It is easy to comprehend why we adhere to the inequality, as it inherently encompasses a broader range of information.

The terms "Domain" and "Class" embody relative concepts. When the desired label is, for instance, Giraffe, domain information transforms into non-semantic data. Conversely, when the desired label is Cartoon, entities like giraffes and elephants assume a specific domain relative to the Cartoon class.

We have supplemented this rigorous assumption Eq.15, specifically as follows:

Assumption 1. *Let $\mathcal{D}_a, \mathcal{D}_b \in \mathcal{D}$. For any $X^i_{\mathcal{D}_a}, X^j_{\mathcal{D}_a}$ belonging to $\mathcal{D}_a$, $X^i_{\mathcal{D}_b}$ belonging to $\mathcal{D}_b$.*

$$I(\mathcal{D}_a;Y)=I(X^i_{\mathcal{D}_a};Y)=I(X^j_{\mathcal{D}_a};Y) \neq I(X^k_{\mathcal{D}_b};Y)=I(\mathcal{D}_b;Y) \tag{16}$$

Compared to Eq.15, this assumption no longer requires the same task-relevant information for each domain.

Definition 1. *(Domain-invariant representation [27]).* Let $\mathcal{S}$ be a family of domains.
1.1. *(Marginal Distribution Alignment) The representation R_0 is said to satisfy the marginal distribution alignment condition if $p(R_0|S)$ is invariant w.r.t. domain $S \in \mathcal{S}$.*
1.2. *(Joint Distribution Alignment) The representation R_0 is said to satisfy the marginal and conditional distribution alignment condition if $p(R_0|Y, S)$ is invariant w.r.t. domain $S \in \mathcal{S}$.*

Note that marginal distribution alignment serves as a foundational prerequisite for DIR.

Details of Proposition 2. *(Optimal Representation Extractor for Classification Task Y under the InfoMin principle). Let f_Y^* denote an optimal encoder for task Y.*

According to the InfoMin principle, we can find the optimal representation of a certain task is:

Definition 2. *(Optimal Representation Z_{sr} of a Task) For a task whose goal is to predict a label Y from the input data X, the optimal representation Z encoded from X is the minimal sufficient statistic w.r.t Y.*

$$Z_{sr} = \arg\min_{Z} I(Z;X) - \beta I(Z;Y) \tag{17}$$

where $\beta > 1$, contingent upon the expressive capacity of the model. The InfoMin principle underscores that maximizing task-relevant representation information is beneficial. However, the representations extracted by the DIR extractor prove insufficient for the unknown target domain.

(In IID scenarios, Eq.3) *For an instance $X_{\mathcal{D}_1}$ drawn from the source support $\mathcal{D}_1$ (and vice versa), we have*

$$\begin{aligned} I(f_Y^*(X_{\mathcal{D}_1});Y) &\overset{\text{D.2}}{=} I(X_{\mathcal{D}_1};Y) \\ &\overset{\text{P.1}}{=} I(R_0;Y)+I(X_{\mathcal{D}_1};Y|R_0) \\ &\overset{\text{P.3}}{=} I(R_0;Y)+I(\mathcal{D}_1;Y|R_0)+I(X_{\mathcal{D}_1};Y|\mathcal{D}_1) \\ &\overset{\text{A.1}}{=} I(R_0;Y)+I(\mathcal{D}_1;Y|R_0) \end{aligned} \tag{18}$$

We compare Eq.1 with Eq. 3 and have:

$$I(R_0;Y)+I(\mathcal{D}_1;Y|R_0) - H(R_0) = I(\mathcal{D}_1;Y|R_0) \tag{19}$$

We find that the problems faced under the IID scenario are mainly attributed to $I(\mathcal{D}_{pT};Y;\mathcal{D}_s|R_0) > 0$, namely, the absence of DOI.

(In OOD scenarios, Eq.4) *For an unseen instance $X_{pT} \in \mathcal{D}_{pT}$ outside the source support, we have*

$$\begin{aligned} I(f_Y^*(X_{pT});Y) &\overset{\text{D.2}}{=} I(X_{pT};Y) \\ &\overset{\text{P.2}}{=} I(X_{pT};\mathcal{D}_s;Y) + I(X_{pT};Y|\mathcal{D}_s) \\ &\overset{\text{P.3}}{=} I(X_{pT};\mathcal{D}_s;Y) + I(\mathcal{D}_{pT};Y|\mathcal{D}_s) + I(X_{pT};Y|\mathcal{D}_s,\mathcal{D}_{pT}) \\ &\overset{\text{P.1}}{=} I(X_{pT};R_0;Y) + I(X_{pT};Y;\mathcal{D}_s|R_0) + I(\mathcal{D}_{pT};Y|\mathcal{D}_s)+I(X_{pT};Y|\mathcal{D}_s,\mathcal{D}_{pT}) \\ &\overset{\text{P.3}}{=} I(R'_0;Y) + I(\mathcal{D}_{pT};Y|\mathcal{D}_s) + I(\mathcal{D}_{pT};Y;\mathcal{D}_s|R_0) + I(X_{pT};Y|\mathcal{D}_s,\mathcal{D}_{pT}) \\ &\overset{\text{A.1}}{=} I(R'_0;Y) + I(\mathcal{D}_{pT};Y|\mathcal{D}_s) + I(\mathcal{D}_{pT};Y;\mathcal{D}_s|R_0) \end{aligned} \tag{20}$$

We compare Eq.1 with Eq. 4 and have:

$$\begin{aligned} & I(R_0'; Y) + I(\mathcal{D}_{pT}; Y|\mathcal{D}_s) + I(\mathcal{D}_{pT}; Y; \mathcal{D}_s|R_0) - H(R_0) \\ = & \underbrace{- I(R_0; Y|R_0')}_{\text{case 1}} + \underbrace{I(\mathcal{D}_{pT}; Y|\mathcal{D}_s)}_{\text{case 2}} + \underbrace{I(\mathcal{D}_{pT}; Y; \mathcal{D}_s|R_0)}_{\text{case 3}} \end{aligned} \tag{21}$$

We found that the problems faced under the OOD scenarios are more complex than IID scenarios, may include excess noise or lack of information. These are the three cases discussed in the Sec.3.2.

A.2 Relationship between Existing Theory

We compare our approach with the theoretical foundations behind the current state-of-the-art DRM method. Similarly considering domain-specific information, the theoretical framework underlying the DRM method is:

Proposition 3. *Let $\{\mathcal{D}_i, f_i\}_{i=1}^K$ and $\mathcal{D}_\mathcal{T}, f_\mathcal{T}$ be the empirical distributions and corresponding labeling function for source and target domain, respectively. For any hypothesis $\hat{f} \in \mathcal{F}$, given mixed weights $\boldsymbol{\alpha} \in \mathbf{R}^K : \|\boldsymbol{\alpha}\|_1 = 1 \wedge \boldsymbol{\alpha} \geq 0$, we have:*

$$\epsilon_t(\hat{f}) \leq \sum_{i=1}^K \alpha_i \big(\mathbb{E}_{X\sim\mathcal{D}_i}[\frac{P_\mathcal{T}(X)}{P_i(X)}|\hat{f} - f_i|] + \mathbb{E}_{X\sim\mathcal{D}_\mathcal{T}}|f_i - f_\mathcal{T}|\big) \tag{22}$$

According to the derivation of [49],

$$\begin{aligned} & \sum_{i=1}^K \left(\mathbb{E}_{X\sim\mathcal{D}_i}[\alpha_i \frac{P_\mathcal{T}(X)}{P_i(X)}|\hat{f} - f_i| + \alpha_i \mathbb{E}_{\mathcal{D}_\mathcal{T}}[|f_i - f_\mathcal{T}|]]\right) \\ = & \sum_{i=1}^K \left(\mathbb{E}_{X\sim\mathcal{D}_\mathcal{T}}[\alpha_i|\hat{f} - f_i| + \alpha_i|f_i - f_\mathcal{T}|]\right) \end{aligned} \tag{23}$$

The ideal hypothesis

$$f^* = \arg\min_{\hat{f}\in\mathcal{F}}[\sum_{i=1}^K \left(\mathbb{E}_{X\sim\mathcal{D}_\mathcal{T}}[\alpha_i|\hat{f} - f_i| + \alpha_i|f_i - f_\mathcal{T}|]\right)] \tag{24}$$

We define $\mathcal{F}_s := \{f_i\}_{i=1}^K$, where a broader hypothesis space is denoted as $\mathcal{F}_s \subseteq \mathcal{F}_s^+$. Next, we have

$$\begin{aligned} \inf_{\boldsymbol{\alpha}\in\mathbf{R}^K} \sum_{i=1}^K \left(\alpha_i|\hat{f}^* - f_i| + \alpha_i|f_i - f_\mathcal{T}|\right) & \geq \min_{\mathcal{F}_s} |f_i(X) - f_\mathcal{T}(X)| \\ & \geq \min_{\mathcal{F}_s^+} |f_i(X) - f_\mathcal{T}(X)| \end{aligned} \tag{25}$$

In the generalization bound of DRM, f^* is subject to constraints imposed by the source labeling functions $\mathcal{F}_s$ that minimizes the empirical risk on each source domain.

Our method investigates tighter generalization bounds through two distinct avenues. Firstly, employing the domain-disentangled invariant representation learning for deconstructing the hypothesis space of the source and facilitating flexible combinations. This framework transforms the original source hypothesis space $\mathcal{F}_s$, initially comprised of only K points, into a surface $\mathcal{F}_s^+$ (also contains the source hypothesis). Secondly, we introduce the inference stage by adopting a strategy for selecting auxiliary policies to aid in the optimal search for α.

B BASELINE DETAILS AND ADDITIONAL COMPARATIVE EXPERIMENTS

In this section, we introduce the methods in more detail, providing insights and highlighting related differences.

- **ERM[39]** aggregates data from all source domains and aims to minimize the cross entropy loss associated with classification.
- **DANN[12]** implements a domain discriminator and employs adversarial learning to harmonize feature distributions across source domains.
- **CORAL[36]** acquires a nonlinear transformation that aligns correlations among layer activations in deep neural networks.
- **GroupDRO[33]** enhances the impact of domains with higher errors during ERM by adjusting the weights of minibatches.
- **IRM[1]** does not aim to match representations across all domains; instead, it enforces that the optimal classifier over the representation space remains consistent across all domains.
- **VREx[17]** mitigates risk disparities across diverse training domains to enhance the model's resilience to drastic distribution shifts, encompassing both causal and anti-causal factors.
- **ARM[46]** extends MLDG by adding a module that calculates domain embeddings. These embeddings help the prediction module gather insights about the input distribution.
- **SWAD[4]** discovers flatter minima and reduces overfitting through a dense and overfit-aware approach to randomly selecting weights.
- **SagNet[26]** separates style features from class information, avoiding biased predictions and emphasizing content.
- **AND-mask[29]**checks if all gradients for parameters agree on a direction before updating the network. The update occurs only when there is unanimous agreement among the gradients.
- **SAND-mask[34]** verifies agreement in gradient direction and encourages consistency in their magnitudes. This additional step helps discover invariances across different training domains.
- **Fish[35]** matches the gradient across different domains. This is achieved by maximizing the inner product between gradients from different domains.
- **SelfReg[16]** introduces a class-specific domain perturbation layer (CDPL), which facilitates the successful implementation of mixture augmentation, even with only positive pairs without negative pairs. SelfReg utilizes contrastive learning to learn domain-invariant representations, while attempting to disregard domain-orthogonal invariant information as much as possible. In contrast, our approach fundamentally differs as we aim to maximize the effective utilization of domain-orthogonal invariant information. This method essentially leverages contrastive learning to acquire domain-invariant representations while minimizing consideration of domain-specific classification information. In contrast, DDIR fundamentally differs as we aim to maximize the effective utilization of domain-orthogonal invariant information.
- **PCL[45]** is also a contrastive learning-based method that eases the positive alignment problem by substituting sample-to-sample relations with proxy-to-sample relations. Essentially, PCL also focuses on extracting domain-invariant representations.
- **Fishr[31]** is a regularization method that minimizes differences in gradient variances at the domain level, demonstrating that Fishr aligns domain-level risks and Hessians, leading to reduced inconsistencies between different domains.

- **RIDG[5]** leverages the reasoning concept to achieve a strong output goal. It works by ensuring agreement between the reasoning matrix of each sample and its average value. † We performed experiments on the RotatedMNIST according to the official code.
- **CCFP[18]** introduces modules that have adjustable feature changes and semantic consistency restrictions. Compared to the runner-up CCFP, DDIR reveals similar performance on PACS but outperforms CCFP by 0.8% on Office Home, 0.8% on RotatedMNIST (97.8%->98.6%) and 1.5% on DomainNet.
- **DRM[49]** introduces domain-specific classifiers for each domain, where the target domain data is processed through all classifiers before final classification decisions are made. In contrast to our method, DRM lacks a dedicated domain-invariant classifier to guarantee the exclusion of extraneous redundancies in classifying distant support instances (The first quadrant in Fig.3). The irrelevance between different classifiers obtaining representation contained information is also not involved.
- **mDSDI[3]** based on decoupling task-relevant representations. However, mDSDI requires a dual-backbone network, rendering a less equitable comparison. DDIR outperform mDSDI performance using a single backbone network (Tab.4). Moreover, this method concatenates DIR and all domain specific representations before assigning the task classifier for classification. This introduces two issues. Firstly, during the testing phase, it introduces redundant information, impacting the task classifier's decision-making particularly in the cases of the first, second and fourth quadrants in Fig.3. Secondly, during training, for a given training data, the task relevance of domain-specific representations from other domains is low, thereby diminishing the attention of the trained classifier to domain-specific representations.

Table 6: LODO results on Domainbed *with oracle selection.* The results of other methods can be found in [49].

	RMNIST	PACS	VLCS	DomainNet
ERM	97.8 ± 0.1	86.7 ± 0.3	77.6 ± 0.3	41.3 ± 0.1
DRM	98.3 ± 0.1	88.4 ± 0.9	79.5 ± 2.4	42.7 ± 0.1
DDIR	98.6 ± 0.1	88.9 ± 0.3	79.7 ± 0.9	43.2 ± 0.2

C DATASET DETAILS

In this section, we elaborate on the datasets, with detailed parameter settings provided for each dataset in Tab.7.

1) **RotatedMNIST (RMNIST)** [13] dataset comprises six domains, each containing 11,667 images across 10 classes (digits 0-9). In each domain, the images undergo rotations at angles of 0, 15, 30, 45, 60, and 75 degrees. The set of images at each specific angle forms a domain, denoted as $\texttt{Rm}_0$, $\texttt{Rm}_{15}$, $\texttt{Rm}_{30}$, $\texttt{Rm}_{45}$, $\texttt{Rm}_{60}$ and $\texttt{Rm}_{75}$.

2) **PACS** [19] dataset comprises 9,991 images in 7 classes, distributed across four domains: art painting (A), cartoon (C), photo (P) and sketch (S).

3) **Office-Home** [32] dataset consists of 15,588 images categorized into 65 classes across four domains: art (Ar), clip art (Cl), product (Pr) and real-world (Rw).

4) **VLCS** [10] dataset consists of 10,729 images categorized into 5 classes across four domains: Caltech101 (C), LabelMe (L), SUN09 (S), VOC2007 (V).

5) **DomainNet (DNet)** [30] dataset consists of 569,010 images categorized into 345 different classes across six domains: Clipart (C), Sketch (S), Quickdraw (Q), Real (R), Infograph (I) and Painting (P).

6) **Terra Incognita (TeInc)** [2] dataset consists of 24,788 images categorized into 10 different classes, covering four domains: wild animal images captured at locations L100, L38, L43, and L46.

Table 7: Experimental setup details.

	RMNIST	PACS	OfficeHome	VLCS	TeInc	DNet
batch size N/K	64	16				12
learning rate	1.0 E-03	5.0 E-05				
weight decay	0	1.0 E-05				
total step	5000					15000
input size	$3 \times 28 \times 28$	$3 \times 224 \times 224$				
z_c size	128	2048				
R size	32	512				

D ADDITIONAL RESULTS

D.1 Results on ResNet-18 Backbone

We employ the ImageNet-pretrained ResNet-18 [15], with the dimension of high-dimensional feature z_c is 512. This dimension matches the size of the representations R used for DI and DOI. We utilize the ResNet-18 backbone on both the PACS and Office-Home datasets. Our findings indicate that the DDIR method, when compared to optimal results, achieves an average accuracy improvement of 1.5% and 0.6% on the respective datasets. Additionally, in comparison to ERM, our approach exhibits enhancements of 4.5% and 3.3% on the respective datasets.

Table 8: Leave-one-domain-out results with ResNet-18.

	P	A	C	S	**AVG.**	Ar	Cl	Pr	Rw	**AVG.**
ERM	78.0	73.4	94.1	73.6	79.8	52.2	48.7	69.9	71.7	60.6
DANN	79.0	72.5	94.4	70.8	79.2	51.8	47.1	69.1	70.2	59.5
CORAL	81.5	75.4	95.2	74.8	81.7	55.1	49.7	71.8	73.1	62.4
IRM	76.9	75.1	94.3	77.4	80.9	49.7	46.8	67.5	68.1	58.0
VREx	74.4	75.0	93.3	78.1	80.2	51.1	47.4	69.0	70.5	59.5
SelfReg	82.5	74.4	95.4	74.9	81.8	55.1	49.2	72.2	73.0	62.4
RIDG	82.4	76.7	95.3	76.7	82.8	56.6	50.3	72.5	73.8	63.3
GroupDRO	77.7	76.4	94.0	74.8	80.7	52.6	48.2	69.9	71.5	60.6
SagNet	82.9	73.2	94.6	76.1	81.7	55.3	49.6	72.1	73.2	62.5
ARM	79.4	75.0	93.3	78.1	80.2	51.3	48.5	68.0	70.5	59.5
Fish	80.9	75.9	95.0	76.2	82.0	54.6	49.6	71.3	72.4	62.0
DDIR-ODS	82.6	77.7	95.9	80.9	84.3	56.1	53.4	72.1	74.1	63.9

D.2 Results on Terra Incognita Dataset

In Tab.9, we further implemented our method (with ODS approach) on the Terra Incognita dataset (DomainBed with Training-domain mode l selection). Since the absence of this Terra Incognita dataset in many comparative methods, we have placed this experiment in the appendix instead of Tab.1.Our method achieves an average

accuracy of 49.5% (±0.8), surpassing not only all current state-of-the-art comparative methods but also showing a 3.4% improvement in average accuracy compared to ERM.

Table 9: Leave-one-domain-out results on Terra Incognita.

	ERM	IRM	GDRO	CORAL	mDSDI	ARM	VREx	**DDIR**
AVG.	46.1	47.6	43.2	47.6	48.1	45.5	46.4	49.5 ± 0.8

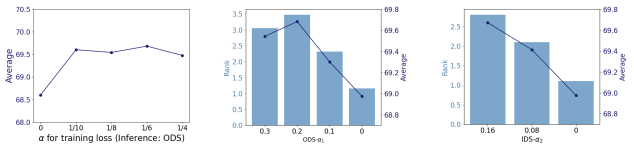
D.3 Results on SWAD-based Codebase

In Tab.10, we further implemented our method on the SWAD-based codebase, surpassing other SWAD-based methods and achieving superior performance.

Table 10: SWAD-based methods on Office-Home and PACS.

	A	C	P	S	**AVG.**	Ar	Cl	Pr	Rw	**AVG.**
SWAD[4]	89.3	83.4	97.3	82.5	88.1	66.1	57.7	78.4	80.2	70.6
PCL[45]	90.2	83.9	98.1	82.6	88.7	67.3	59.9	78.7	80.7	71.6
CCFP[18]	87.5	81.3	96.4	81.4	86.6	68.0	58.6	79.7	81.9	72.1
DDIR	89.3	85.1	97.9	82.4	88.7	69.8	59.9	79.7	81.1	72.6

D.4 Sensitivity to Hyper-Parameter



(a) α (b) α_1 (c) α_2

Figure 8: Sensitivity to Hyper-Parameter on $\alpha, \alpha_1, \alpha_2$

D.4.1 Sensitivity to Hyper-Parameter $\alpha, \alpha_1, \alpha_2$. The hyper-parameters $\alpha, \alpha_1, \alpha_2$ primarily balance the weights of the DI branch network and the DOI branch during both training and inference. The sensitive results are shown in Fig. 7. "Rank" is calculated in the same way as described in Sec.5.3 proof-of-concept experiment.

During training process, the parameter α influences the ability of the high-dimensional common representation z_c to retain information. When $\alpha = 0$, the DOI branch network is not trained, leading the model to degrade into DIR learning. Conversely, when $\alpha = 1/K$, the model loses its ability to extract DI information. Therefore, it is essential to avoid overly large or small values for α.

During inference process, the decision-making process is influenced by two factors, α_1 and α_2, which determine the importance of DI and DOI. When either α_1 or α_2 is set to zero, the decision solely relies on the prediction from the DI branch. As α_1 or α_2 increases, the significance of DI information decreases. Our perspective aligns with the notion that DI information should carry greater weight than DOI information due to its higher likelihood of being task-relevant.

Table 11: Sensitivity to temperature τ.

	Ar	Cl	Pr	Rw	**AVG.**
$\tau = 0.05$	65.1 ± 0.6	56.7 ± 1.2	77.4 ± 1.2	79.3 ± 0.6	69.6 ± 0.9
$\tau = 0.07$	65.6 ± 1.5	56.2 ± 1.6	77.8 ± 0.1	79.2 ± 0.9	69.7 ± 0.2
$\tau = 0.10$	65.6 ± 0.9	56.3 ± 0.6	78.0 ± 0.6	79.4 ± 0.9	69.8 ± 0.4

D.4.2 Sensitivity to Temperature τ . In our experiments, we used the default temperature τ in the InstDisc [44] method, which initially introduced the InfoNCE contrastive loss and has since been widely used in similar studies. We conducted additional experiments varying the temperature τ, as shown in Tab.11. When $\tau = 0.05$, the average accuracy on Office-Home dataset is 69.6%, and when $\tau = 0.10$, the average accuracy is 69.8%. Our results suggest that temperature variations minimally impact DDIR accuracy within the range of 0.05 to 0.1. Additionally, our method outperformed the state-of-the-art methods at all three temperatures tested.

D.5 Training Process Visualization

Here, we present visualizations of different loss functions applied to each branch concerning the training step. As shown in Fig.7, we record changes in the representation space every 200 epochs for the initial 1400 epochs.

For the DIR branch, we compare the impact of including the $\mathcal{L}_{DI}$ loss function on the representation space. We observe that instances from the same domain without the $\mathcal{L}_{DI}$ are more concentrated in the representation space, indicating that $\mathcal{L}_{DI}$ better excludes domain-orthogonal invariant information.

For the DOIR branches, we compare our method without the $\mathcal{L}_{\mathcal{D}_k OI}$ loss function and with a replacement $\mathcal{L}^*_{\mathcal{D}_k OI}$, which excludes instances of different domains and the same class from the negative instance set. We observe that using the $\mathcal{L}^*_{\mathcal{D}_k OI}$ loss improves hierarchy compared to not using $\mathcal{L}_{\mathcal{D}_k OI}$. However, it falls short of completely distinguishing this information from that of other domains.

E ABBREVIATIONS

Table 12 lists all abbreviations used in the paper, serving as a quick reference for readers to understand them better.

Table 12: Abbreviations Summary

Abbreviation	**Full Form**
Specific Abbreviations	
DDIR	Domain-Disentangled Invariant Representation
DOI(R)	Domain-Orthogonal Invariant (Representation)
ODS / IDS	Optimal DOI Selection / Inclusive DOI Selection
Commonly Used Abbreviations	
DI(R)	Domain Invariant (Representation) [7, 42, 47]
DG	Domain Generalization [16, 18, 42, 47, 49]
IID	Independent and Identically Distributed [47, 49]
OOD	Out-of-Distribution [18, 47]
LODO	Leave-One-Domain-Out result [14, 42]
ID	In-Distribution result [47, 49]
HM	Harmonic Mean result [23, 47]

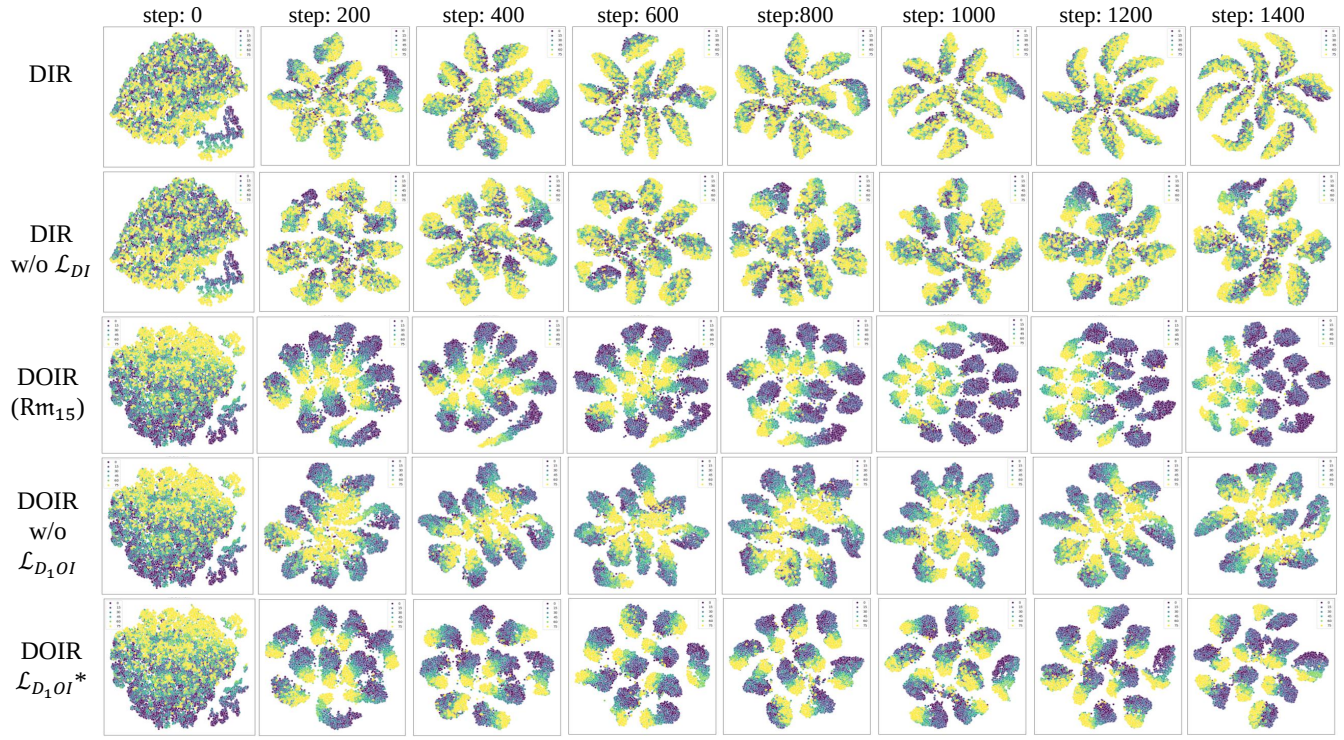


Figure 7: Evolution of the representation space throughout training across 1400 steps (recorded every 200 steps).