Journal of Machine Learning Research 27 (2026) 1-54 Submitted 2/25; Revised 4/26; Published 4/26

Bridging Domain Invariance and Diversity: A Fine-Grained Risk Bound for Domain Generalization

Xi Wang WANGXI7@SXU.EDU.CN
Institute of Intelligent Information Processing
Shanxi University, Taiyuan, 030006, China

Liang Bai * BAILIANG@SXU.EDU.CN
Institute of Intelligent Information Processing
Shanxi University, Taiyuan, 030006, China

Xian Yang XIAN.YANG@MANCHESTER.AC.UK
Alliance Manchester Business School
University of Manchester, Manchester, M13 9PL, UK

Richard Yi Da Xu XUYIDA@HKBU.EDU.HK
Department of Mathematics
Hong Kong Baptist University, Hong Kong SAR, China

Jiye Liang LJY@SXU.EDU.CN
Institute of Intelligent Information Processing
Shanxi University, Taiyuan, 030006, China

Editor: Zhihua Zhang

Abstract

Domain-invariant representation learning and domain augmentation algorithms are two principal methodological paradigms for addressing domain generalization. They are widely employed in the machine learning literature to enhance domain invariance and domain diversity, respectively. However, existing risk bounds for domain generalization do not simultaneously capture the contributions of both approaches. This limitation arises because bounds derived directly in the original latent space are typically too coarse-grained and ambiguous to characterize how invariance and diversity jointly influence generalization. Since these two properties are often regarded as being inherently contradictory, it becomes difficult to disentangle and rigorously characterize their individual effects. To address this issue, we first observe that the latent representation space can be decomposed into several distinct subspaces, each exhibiting different characteristics and therefore being better suited for analyzing the respective roles of domain invariance and domain diversity. Building on this observation, we propose a unified analytical framework for domain generalization. Specifically, we introduce a Tri-Space Latent Representation and establish its unique decomposability via a direct-sum decomposition. Under this decomposition, each data representation can be uniquely partitioned into three components: domain-invariant features, spurious invariant features, and domain-variant features. Within this framework, we derive a finer-grained bound on the target-domain risk, which consists of two principal terms corresponding to domain diversity and invariant factors. By theoretically analyzing these two terms, we show that domain-invariant representation learning and domain augmentation are both effective and, crucially, compatible strategies for addressing domain

*. Corresponding author

generalization. Finally, we design two sets of experiments to empirically validate the relationship between domain invariance and domain diversity, and to examine their respective effects on domain generalization performance.

Keywords: Domain generalization, Learning theory, Domain invariance, Domain diversity, Tri-space latent representation

1. Introduction

Machine learning models have long relied on the assumption that data are independent and identically distributed (i.i.d.). While this premise has underpinned many advances, it often breaks down in practice when models face test data drawn from a different distribution than the training data, leading to significant performance degradation. Such distribution shifts—caused by factors like environmental changes, sensor variations, or temporal evolution—are well-documented in the literature Khosla et al. (2012); Hendrycks and Dietterich (2019); Moreno-Torres et al. (2012). These challenges have motivated the development of methods capable of handling real-world distribution mismatches.

In response to the limitations of the i.i.d. assumption in practical scenarios, domain adaptation has emerged as a valuable technique. It aims to adapt models—typically trained on labeled data from a source domain—to perform accurately in a target domain with a different data distribution. Often, the target domain contains only unlabeled data, a situation not encountered during training. Both the theoretical foundations and practical applications of domain adaptation are extensively discussed in Ben-David et al. (2010); Ganin and Lempitsky (2015); Ganin et al. (2016). Building on this idea, domain generalization addresses an even wider set of challenges. As detailed in Blanchard et al. (2011); Muandet et al. (2013); Ghifary et al. (2015), domain generalization involves training models on data collected from multiple domains, each with distinct distributions, with the goal of achieving robust performance across both known and previously unseen domains. This approach seeks consistent generalization across a broad spectrum of domain shifts, and comprehensive overviews can be found in Zhou et al. (2022); Wang et al. (2022).

The mainstream methodologies for domain generalization primarily fall into two categories. The first is domain-invariant representation learning, which aims to extract cross-domain invariant features by enforcing distribution consistency constraints across source domains, as exemplified by the approach introduced in Zhou et al. (2021). The second is domain augmentation, which enhances the diversity of training data in either the sample space or the feature space (Xu et al., 2021; Zhou et al., 2024). By exposing the model to a broader range of variations during training, this strategy improves its adaptability to unseen domains. In recent years, both lines of research have made notable progress in their respective directions. However, achieving reliable domain generalization remains challenging given current performance levels. Therefore, establishing a unified theoretical framework for systematically analyzing these approaches has become increasingly crucial. Currently, a significant line of theoretical research in domain generalization—particularly within the worst-case or arbitrary-target formulation—focuses on establishing upper bounds for the risk on any target domain within a specific set (Albuquerque et al., 2019; Sicilia et al., 2023). These studies underscore the importance of learning domain-invariant features and enhancing domain diversity to improve generalization performance. However, existing risk

bounds for domain generalization fail to simultaneously capture the contributions of both approaches. This limitation stems from the fact that bounds derived directly in the original latent space are generally too coarse-grained to characterize how invariance and diversity jointly influence generalization. Since these two properties are often regarded as inherently contradictory, it becomes difficult to disentangle and rigorously characterize their individual effects within such a unified yet overly broad space. To address this fundamental issue, we propose a novel generalization framework built upon a structural decomposition of the latent space. We first observe that the latent representation space can be decomposed into several distinct subspaces, each exhibiting different characteristics and thus being more suitable for analyzing the respective roles of domain invariance and domain diversity. Building on this insight, we formally define a Tri-Space latent representation composed of three key components: domain-invariant features that follow consistent distributions across all domains, spurious invariant features that maintain invariant distributions only within source domains, and domain-variant features that exhibit distributional shifts across domains. We establish the unique decomposability of this representation via a direct-sum decomposition of the space. Under this structured framework, we derive a finer-grained target domain risk bound and perform a term-by-term analysis, revealing two principal factors governing generalization performance: domain diversity and domain invariance. Through theoretical examination of these terms, we demonstrate that domain-invariant representation learning and domain augmentation are not only effective but also mutually complementary strategies. Finally, we design two sets of experiments to empirically validate the relationship between domain invariance and diversity, and to assess their combined impact on generalization performance. The primary contributions of this paper are listed as follows:

- We propose a Tri-space latent representation method which uses the direct sum decomposition to define and prove a data latent representation uniquely consisting three subspaces with distinct distributional properties. This approach helps us understand how different features contribute to domain generalization.
- We derive a finer-grained target domain risk bound for domain generalization composed of domain invariance and diversity factors, which can simultaneously explain domain-invariant representation learning and domain augmentation algorithms.
- We abstract two categories of algorithms—domain-invariant representation learning and domain augmentation—into rigorous mathematical definitions, and based on these definitions together with our fine-grained bound, we carry out a unified analysis of both types of algorithms.
- We perform experiments to investigate the relationship between domain invariance, domain diversity, and domain generalization, thereby validating our theoretical findings.

The remainder of the paper is organized as follows: Section 2 discusses related work. Section 3 details the symbols and formulas used in this paper. In Section 4, we present a novel method to derive a finer-grained target domain risk bound for domain generalization, and based on it, explain the roles of domain-invariant representation learning and domain augmentation algorithms in solving the domain generalization problem. In Section 5, we make some extensions to our theory. The paper concludes in Section 6.

2. Related Work

This section reviews existing research relevant to our study, with particular focus on domain-invariant representation learning, domain augmentation, and generalization theory. These research directions collectively provide the theoretical and empirical foundation for the unified framework we propose.

2.1 Domain-Invariant Representation Learning

The core objective of these methods is to minimize distribution discrepancies among multiple source domains by capturing invariant features. To this end, researchers have developed various techniques: for example, Li et al. (2018a) employed Maximum Mean Discrepancy (MMD) to reduce feature distribution divergence; Wang et al. (2021) and Li et al. (2020) utilized KL divergence to align domain-agnostic posterior distributions and to align source domain features with Gaussian distributions, respectively. As demonstrated by Zhou et al. (2021), optimal transport methods leverage the Wasserstein distance (OT distance) to align marginal distributions across source domains. Furthermore, domain adversarial learning (explored in Ganin et al. (2016), Li et al. (2018b) and Rahman et al. (2020)) employs adversarial strategies to promote the development of domain-agnostic features effective for novel and unseen domains.

In contrast to approaches that learn invariant features by matching representation distributions across domains, Invariant Risk Minimization (IRM) Arjovsky et al. (2019) has emerged as a highly influential framework that introduces a paradigm shift in formulating domain invariance. Rather than aligning feature distributions across domains, IRM approaches the problem from the perspective of predictor invariance—it aims to learn feature representations that allow an optimal classifier to remain consistent across all training domains. This perspective reframes the objective from distribution matching to learning causal mechanisms that are stable across different environments. The theoretical insights of IRM have inspired considerable follow-up work, including extensions such as Krueger et al. (2021), which emphasizes risk consistency across domains to learn invariant feature representations. Collectively, these approaches underscore the importance of capturing underlying causal structures that endure through distribution shifts. Meanwhile, Shi et al. (2021) achieves invariance by promoting gradient direction consistency across domains.

2.2 Domain Augmentation

Domain augmentation serves as a key strategy for enhancing domain diversity, thereby supporting robust domain generalization. Researchers have explored various methods, with particular attention to image-level and feature-level modifications. Volpi et al. (2018) and Shankar et al. (2018) leverage adversarial gradients to modify training inputs, thereby increasing domain diversity. The RandConv method proposed by Xu et al. (2021) introduces transformations via a randomly initialized convolutional layer, generating instances with new styles. Peng et al. (2022) utilizes adversarial training to create "fictitious" yet "challenging" sample populations and frames the model training within a meta-learning scheme.

At the feature level, techniques such as MixStyle Zhou et al. (2024) achieve style augmentation by mixing feature statistics across different instances. Similarly, Mancini et al.

(2020) applies the Mixup strategy at both pixel and feature levels to blend instances from different domains, further enriching domain diversity and enhancing the robustness of domain generalization. Although these methods effectively enhance domain diversity, there is currently a lack of systematic theoretical integration linking these methods with domain-invariant strategies to form a comprehensive solution for domain generalization.

2.3 Generalization Theory

Generalization theory in domain adaptation primarily involves measuring the distribution discrepancy between source and target domains. The seminal work by Ben-David et al. (2006) introduced the $\mathcal{H}$-divergence as a metric for assessing distribution distances. Mansour et al. (2009) expanded the applicability of these metrics to bounded loss functions, thereby broadening the theoretical landscape.

However, domain generalization faces the challenge of estimating risk without access to target domain data. Early theoretical explorations by Blanchard et al. (2011) laid the groundwork for this field. Blanchard et al. (2021) provided an upper bound for the maximum average risk estimation error of hypotheses within a closed ball in a reproducing kernel Hilbert space. Ye et al. (2021) introduced the concepts of "variation" and "informativeness" of features to derive a domain generalization error bound and provided a rigorous definition for the "learnability" of domain generalization. Li et al. (2022) defined a specialized Rademacher complexity term for the domain generalization problem, indicating that models with lower complexity contribute to better domain generalization. Recent theoretical works by Albuquerque et al. (2019) and Sicilia et al. (2023) propose methods that establish reference frames for the target domain using source domains and bound the risk on arbitrary target domains within these reference frames, thereby providing new perspectives for the theoretical analysis of domain generalization.

3. Preliminaries

3.1 Notations and Setup

3.1.1 Fundamental Notations

This subsection presents the notation used throughout the paper, with key symbols summarized in Table 1. We consider a standard network architecture composed of an encoder, responsible for feature extraction, and a classifier, which consists of a fully-connected layer followed by an activation function.

Let $\mathcal{X}$ denote the raw input space. The Encoder maps instances from $\mathcal{X}$ to a representation space $\mathcal{Z} \subset \mathbb{R}^{d_z}$, where d_z is the dimensionality of $\mathcal{Z}$. The label space is denoted by $\mathcal{Y}$. Following common practice in the field, and since most existing analyses focus on binary classification, we take $\mathcal{Y} = \{0, 1\}$. The function $f : \mathcal{Z} \to \mathcal{Y}$ acts as a deterministic labeling function—an assumption widely adopted in prior work (e.g., Zhao et al. (2019); Albuquerque et al. (2019)) and formally restated in Section 4.2. That is, for any pair $(z, y) \in \mathcal{Z} \times \mathcal{Y}$, the label satisfies $y = f(z)$.

A domain is defined as a tuple $\langle D, f \rangle$, where D is a probability distribution over $\mathcal{Z}$, and f is the deterministic labeling function for that domain. We also consider a hypothesis

Notation	Description
Data and Latent Spaces	
$x; \mathcal{X}$	Data sample and raw input space
$z; \mathcal{Z}$	Feature representation and latent space
Hypothesis and Domains	
$h; \mathcal{H}$	Hypothesis and hypothesis space
$s; S$	Source domain and set of source domains
$t; T$	Target domain and set of target domains
d	Arbitrary domain in $S \cup T$
$\|S\|$	Number of source domains
Domain Characteristics	
$f_d; D_d$	Labeling function and distribution of domain d
$\mathcal{D}; \mathcal{P}$	Space of domain distributions and corresponding probability law
$O; U$	Reference for target domain distributions
Loss and Risk Functions	
$\ell(\cdot)$	Loss function
$\varepsilon_{D_d}(h)$	Risk of hypothesis h on domain d
$\hat{\varepsilon}_{D_d}(h)$	Empirical risk on domain d's data
Domain Weights and Samples	
π_s	Weight coefficient for source domain s
$z_{s,i}$	i-th data point from source domain s
Distribution Divergence Measures	
$d_{\mathcal{H}}[\cdot, \cdot]$	$\mathcal{H}$-divergence between distributions
$d_{\tilde{\mathcal{H}}}[\cdot, \cdot]$	$\tilde{\mathcal{H}}$-divergence between distributions
$d_{\mathcal{H}\Delta\mathcal{H}}[\cdot, \cdot]$	$\mathcal{H}\Delta\mathcal{H}$-divergence between distributions

Table 1: Basic Symbols in the Problem Setting

$h : \mathcal{Z} \to [0, 1]$, which outputs the predicted probability of class 1, and belongs to a hypothesis class $\mathcal{H}$.

We assume access to multiple source domains during training. Each source domain is formally represented as a tuple $\langle D_s, f_s \rangle$, which we abbreviate simply as s for notational convenience. The set of all source domains is denoted by S, with s_i representing the i-th source domain and $|S|$ indicating the total number of source domains. Similarly, at test time, the model is evaluated on an unseen target domain, formally represented as $\langle D_t, f_t \rangle$ and abbreviated as t. The set of all possible target domains is denoted by T.

The risk associated with a hypothesis h in a domain $\langle D, f \rangle$ is defined as follows:

$$\varepsilon_D(h) = E_{z \sim D}[\ell\left(h(z), f(z)\right)], \tag{1}$$

where the ℓ is the loss function, and we assume it has a linear upper bound with respect to the prediction error on the compact set $[0, 1]$. In Section 4.2, we will formally define this assumption.

The empirical risk on the training data of source domain s is given by $\hat{\varepsilon}_{D_s}(h) = \frac{1}{n}\sum_{j=1}^{n} \ell(h(z_{s,i}), f_s(z_{s,i}))$, where n is the number of data points in a source domain, and $z_{s,i}$ denotes the i-th data point from the source domain s.

3.1.2 Problem Setup

Below we introduce the basic setup for the theoretical analysis of the domain generalization problem, that is, which theoretical quantity we aim to bound.

In the domain generalization literature, two primary theoretical frameworks have been established for analyzing generalization performance. The first framework follows the strict definition introduced by Muandet et al. (2013), which involves bounding the *domain generalization error* defined with respect to a meta-distribution over domains:

$$\mathrm{Err}(h) = \left| E_{d\sim\mathcal{P}}[\varepsilon_d(h)] - \frac{1}{|S|}\sum_{s\in S} \hat{\varepsilon}_{D_s}(h) \right|, \tag{2}$$

where $\mathcal{P}$ denotes the meta-distribution over all possible domain distributions.

However, defining a rigorous probability measure $\mathcal{P}$ on the space of all domain distributions faces fundamental measure-theoretic challenges. The crux of the problem is that the space of all domain distributions is only an intuitive concept and lacks a strict mathematical definition. As an abstract space composed of probability distributions, it possesses an infinite-dimensional and nonlinear structure. How to ensure the existence of a "meta-distribution" on this space and provide its concrete mathematical formulation remains an open and debatable problem. Some studies have attempted to circumvent this difficulty. For instance, Li et al. (2022) does not explicitly specify the precise mathematical definitions of the space of all domain distributions or the meta-distribution. Instead, it treats the domain space as an abstract space akin to a sample space. However, this approach does not explicitly account for the potential impact that the infinite-dimensional, nonlinear nature of the domain distribution space may have on generalization performance. On the other hand, Blanchard et al. (2021) technically bypasses the issue by leveraging the Azuma–McDiarmid inequality, shifting the analysis from the distribution space to finite samples, thereby avoiding the need for an explicit definition of a meta-distribution. Although this strategy is practical, it does not fully resolve the core theoretical challenge of rigorously constructing a probability measure on the space of domain distributions.

Therefore, in this work, we adopt the second established framework for domain generalization analysis. This alternative approach, exemplified by Albuquerque et al. (2019) and Sicilia et al. (2023) and systematically categorized in Wang et al. (2022), addresses the domain generalization problem through *arbitrary-target domain risk analysis*. Instead of assuming a meta-distribution, this framework aims to bound the risk on any possible target domain from a set T, providing guarantees that hold for arbitrary unseen domains during training. This formulation not only avoids the aforementioned measure-theoretic difficulties but also offers strong interpretability and practical insights into domain generalization. Accordingly, we analyze domain generalization by deriving bounds for the risk $\varepsilon_{D_t}(h)$ on any target domain $t \in T$, while evaluating performance on source domains through $\varepsilon_{D_s}(h)$. It should be noted that while our framework bears some resemblance in form to the bounds in multi-source domain adaptation, the problem we address in this paper is domain gen-

eralization, not multi-source domain adaptation. The rationale behind this terminological choice is discussed in greater detail in Appendix J.

3.2 Risk Bound for Domain Generalization extended from Domain Adaptation

Inspired by domain adaptation theory, several existing studies have adopted analytical approaches from domain adaptation to investigate the problem of domain generalization. Domain adaptation theory is built upon the definition of distribution divergence measure between domains. The most fundamental measure is $\mathcal{H}$-divergence, defined as:

$$d_{\mathcal{H}}(D_s, D_t) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{\mathbf{x} \sim D_s} \mathbb{I}[h(\mathbf{x}) = 1] - \mathbb{E}_{\mathbf{x} \sim D_t} \mathbb{I}[h(\mathbf{x}) = 1] \right|, \tag{3}$$

where $\mathcal{H} : \mathcal{X} \to \{0, 1\}$. This distribution discrepancy measure provides a preliminary quantification of the differences between data distributions in binary classification tasks.

Tackling domain generalization problems is challenging because explicit information about the target domain is often unavailable. Existing study address this challenge by selecting an appropriate reference for the target domain. As detailed in Albuquerque et al. (2019), one common approach is to use the convex hull of the source distributions as the reference for the distribution of target domain. This is denoted as $\Lambda_S = \{\bar{D} : \bar{D}(\cdot) = \sum_{s \in S} \pi_s D_s(\cdot), \pi_s \in \triangle_{|S|-1}\}$, where $\triangle_{|S|-1}$ represents the $(|S| - 1)$-dimensional simplex. Using this reference, the risk on any target domain $t \in T$ is bounded as:

$$\varepsilon_{D_t}(h) \leq \underbrace{\sum_{s \in S} \pi_s \varepsilon_{D_s}(h)}_{\text{weighted source risk}} + \theta + \epsilon + \min\{E_{\bar{D}_t}[|f_\pi - f_t|], E_{D_t}[|f_t - f_\pi|]\}. \tag{4}$$

In this formula, $\bar{D}_t$ is the distribution that minimizes the divergence $d_{\mathcal{H}}[D_t, \sum_{s \in S} \pi_s D_s]$ over the coefficients $\pi_1, ..., \pi_{|S|}$, and the weights $\{\pi_s\}_{s \in S}$ in the weighted source risk are determined by it. Here, θ is defined as $d_{\tilde{\mathcal{H}}}[\bar{D}_t, D_t]$. The $d_{\tilde{\mathcal{H}}}[\cdot, \cdot]$ denotes the $\mathcal{H}$-divergence relative to the function class $\tilde{\mathcal{H}}$. This function class, from Zhao et al. (2019), is defined as $\tilde{\mathcal{H}} := \{\text{sgn}(|h(\mathbf{x}) - h'(\mathbf{x})| - l) \mid h, h' \in \mathcal{H},\ 0 < l < 1\}$, where $\mathcal{H} : \mathcal{X} \to [0, 1]$. The term $\epsilon = \max_{s, s' \in S} d_{\tilde{\mathcal{H}}}[D_s, D_{s'}]$, it represents the maximum pairwise $\tilde{\mathcal{H}}$-divergence observed between pairs within the training domains. Additionally, the function $f_\pi(x) = \sum_{s \in S} \pi_s f_s(x)$ symbolizes the combined labeling function for any x in the support set of $\bar{D}_t$, using the weights $\{\pi_s : s \in S\}$ as determined by $\bar{D}_t$.

Similar to domain adaptation theory, this bound also offers algorithmic insights. It suggests that we should learn a latent space where two objectives are minimized simultaneously: first, the maximum distribution discrepancy among the source domains, and second, the minimum distribution discrepancy between the target domain and the convex combination of the source domains.

Furthermore, Sicilia et al. (2023) expand the scope of the target domain reference Λ_S in Albuquerque et al. (2019). Their reference is defined as the following formula:

$$O = \Big\{ D : \sum_s \pi_s d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D) \leq \max_{s, s'} d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_{s'}) \Big\}, \tag{5}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_t) = \sup_{h,h' \in \mathcal{H}} |\mathbb{E}_{\mathbf{x}\sim D_t}\mathbb{I}[h(\mathbf{x}) \neq h'(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim D_s}\mathbb{I}[h(\mathbf{x}) \neq h'(\mathbf{x})]|$ is defined in Ben-David et al. (2010). Within this condition, the risk on any target domain $t \in T$ is bounded as follows:

$$\varepsilon_{D_t}(h) \leq \sum_{s\in S} \pi_s \varepsilon_{D_s}(h) + \frac{1}{2} \min_{D\in O} d_{\mathcal{H}\Delta\mathcal{H}}(D_t, D) + \frac{1}{2} \max_{s,s'\in S} d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_{s'}) + \lambda_\pi, \tag{6}$$

where $\lambda_\pi = \sum_s \pi_s \lambda_s$ and $\lambda_s = \min_{h\in\mathcal{H}} \varepsilon_{D_s}[h] + \varepsilon_D[h]$ is the error of an ideal joint hypothesis for D and D_s. This term provides a sufficient but not necessary condition for model transferability between domains: the existence of a hypothesis function that performs well on both the source and target domains. Similarly, the weights $\{\pi_s\}_{s\in S}$ in $\sum_{s\in S} \pi_s \varepsilon_{D_s}(h)$ are determined by the (D, O) that minimizes $\min_{D\in O} d_{\mathcal{H}\Delta\mathcal{H}}(D_t, D)$.

Eq. (4) and Eq. (6) exhibit excellent interpretability for the role of domain-invariant representation learning, since the term ϵ in Eq. (4) and the term $\max_{s,s'} d_{\mathcal{H}\Delta\mathcal{H}}(D_s, D_{s'})$ in Eq. (6) emphasize the importance of extracting invariant feature representations, as the more similar the source domain features are, the smaller the maximum distribution discrepancy between them. Meanwhile, the term θ in Eq. (4) and term $\min_{D\in O} d_{\mathcal{H}\Delta\mathcal{H}}(D_t, D)$ in Eq. (6) emphasize the importance of augmenting diversity of features, because the more diverse the source domain features, the broader the convex combination of the source domains and the range O in Eq. (5), which results in a smaller minimum distribution discrepancy between the target domain. However, the simultaneous analysis of domain invariance and diversity across an undifferentiated latent space offers limited guidance for balancing these competing objectives, since extracting invariance and enhancing diversity represent fundamentally conflicting operations within a single representation space. This limitation stems not from inherent flaws in the theoretical bounds, but rather from their coarse-grained perspective that fails to distinguish between features requiring invariance and those benefiting from diversity, thereby hindering the theoretical unification of these algorithmic approaches. In the following, we present our main result to address this issue.

4. Main Results

In this section, we first define a Tri-space latent representation and prove its unique decomposability. Based on this, we derive a fine-grained target risk bound to overcome the limitation of existing error bounds. Due to the number of mathematical symbols employed in this paper, we have summarized the notations introduced in our work in Table 2.

4.1 Tri-space Latent Representation

In this subsection, we lay the groundwork for deriving the arbitrary-target domain risk bound for domain generalization by analyzing the structure of $\mathcal{Z}$, the latent space. Specifically, we define three linear subspaces of $\mathcal{Z}$, each corresponding to a type of feature with distinct distributional properties. We then decompose $\mathcal{Z}$ into the direct sum of these three subspaces. This decomposition corresponds to the unique representation of features, meaning that any feature can be uniquely represented as the sum of features from these three subspace. We call this the Tri-space latent representation.

Firstly, we introduce some symbols to be employed in the following text. We use $\mathcal{K}$ to denote the set of all features. It is imperative to discern the disparity between $\mathcal{K}$ and

Notation	Description
Feature Sets and Subspaces	
$k; \mathcal{K}$	A feature and the set of all features
Φ, Γ, Ξ	Three feature sets with different distribution properties
ϕ, γ, ξ	Three features with different distribution properties
$\mathcal{Z}^K$ $(K \in \{\Phi, \Gamma, \Xi\})$	Subspace of $\mathcal{Z}$ corresponding to K
D_d^K	Distribution of domain d in subspace $\mathcal{Z}^K$
$L(\cdot)$	Lebesgue measure
Hypothesis and Complexity	
$\hat{h}; \hat{\mathcal{H}}$	$\hat{h} \in \hat{\mathcal{H}} = \{h(3\cdot); h \in \mathcal{H}\}$
$\mathcal{R}_S^w(\mathcal{H})$	Weighted Rademacher complexity of $\mathcal{H}$
Algorithm Operations and Representations	
$\mathrm{Aug}(\cdot)$	Domain augmentation mapping
$\mathrm{Inv}(\cdot)$	Domain-invariant representation learning mapping
$S_{\mathrm{aug}}; \hat{S}_{\mathrm{aug}}$	Augmented source domain set and augmented sample set
$\mathcal{Z}_{\mathrm{inv}}$	Latent space after $\mathrm{Inv}(\cdot)$
$\mathcal{Z}_{\mathrm{inv}}^{\Gamma}; \mathcal{Z}_{\mathrm{inv}}^{\Phi}; \mathcal{Z}_{\mathrm{inv}}^{\Xi}$	Three Subspace after $\mathrm{Inv}(\cdot)$
U_{inv}	Target Domain Reference after $\mathrm{Inv}(\cdot)$
$\Gamma_{\mathrm{aug}}\Phi_{\mathrm{aug}}; \Xi_{\mathrm{aug}}$	Subset of three feature based on S_{aug}
U_{aug}	Target Domain Reference based on S_{aug}
Domain Generalization Factors	
R_Ξ	Domain invariance factor
R_Φ	Domain diversity factor
Distribution and Loss Function	
$d_{\hat{\mathcal{H}}}^{\Xi}(\cdot, \cdot)$	Distribution divergence measure defined on subspace Ξ
$A(\hat{S})$	Hypothesis output by algorithm A
$\mathcal{L}$	Linear upper bound of ℓ w.r.t. prediction error
L	$\mathsf{L}(x - y) = \ell(x, y)$
$\mathcal{L}_0'$	Derivative of L at 0
$R_\Phi(A(\hat{S}))$	Domain diversity factor relative to $A(\hat{S})$

Table 2: Symbols Defined in Subsequent Content

$\mathcal{Z}$. The space $\mathcal{Z}$ denotes the value space of feature vectors. Each feature corresponds to a coordinate axis of $\mathcal{Z}$. Formally, we treat $\mathcal{Z}$ as a Euclidean space R^{d_z} , and $\mathcal{K}$ as the index set of features, i.e., $\mathcal{K} = \{1, \ldots, d_z\}$. For any subset $K \subseteq \mathcal{K}$, $\mathcal{Z}^K$ denotes the subspace of $\mathcal{Z}$ spanned by the dimensions indexed by K, and D_s^K represents the marginal distribution of the feature vector in source domain s projected onto $\mathcal{Z}^K$.

Definition 1 (Domain-invariant Feature) *We define a feature subset of $\mathcal{K}$ as Γ, which satisfies the following conditions:*

$$\Gamma = \left\{k \in \mathcal{K} : D_s^k = D_t^k; \quad \forall s \in S, \forall t \in T\right\}. \tag{7}$$

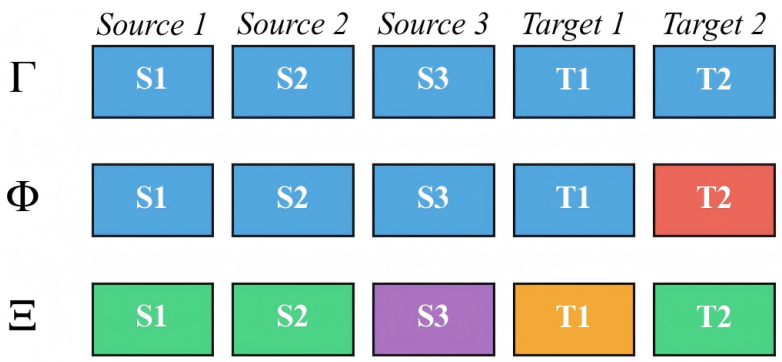


Figure 1: This figure illustrates the three types of features defined in our framework. Each row corresponds to a feature type, each column represents a domain, and different colors indicate different distributions.

We call it domain-invariant feature. Then we define the subspace of $\mathcal{Z}$ corresponds to the feature set Γ as $\mathcal{Z}^{\Gamma}$, and call it domain-invariant subspace.

The distribution of this feature remains consistent across both the source domains and all possible potential target domain. Consequently, we posit that it captures pure task-related information that aids the model in generalizing to any possible target domain.

Definition 2 (Spurious Invariant Feature) *We define a feature subset of $\mathcal{K}$ as Φ, which satisfies the following conditions:*

$$\Phi = \left\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k; \quad \forall s, s' \in S, \exists t \in T\right\}. \tag{8}$$

Then we call it spurious invariant feature and refer to the subspace $\mathcal{Z}^{\Phi}$ as spurious invariant subspace.

This feature contributes to the model's generalization across all source domains but does not facilitate the model's generalization to unseen target domains.

This definition stems from our concerns about domain-invariant representation learning, as the distribution shift from source to target domains can often be more pronounced than the variations among the source domains themselves. This observation leads to a reasonable assumption that not all invariances observed across source domains are beneficial for generalizing to target domains. Therefore, we preliminarily conjecture that spurious redundancy may exist in the invariances derived from aligning the source domains, which can severely harm domain generalization performance. To deepen the understanding of the subspace Φ, consider an illustrative example of cat and dog classification. In this scenario, all source domain images depict black cats and white dogs, whereas the target domain images feature white cats and black dogs. Notably, the color distribution is consistent across source domains but is entirely unrelated to the classification task. In this context, color information is a prime example of a feature in Φ. Although we could mitigate this by introducing images

of cats and dogs in different colors, practical limitations exist. This is because features in Φ typically consist of complex components like image background and style, and we only have access to a limited number of source domains during training. Consequently, it is reasonable to expect a substantial presence of such features in image datasets. In Appendix H, we provide a more comprehensive synthetic example that concretely instantiates the three types of features depicted in Figure 1 and further demonstrates how spurious invariant features can still lead to generalization failure even after domain-invariant representation learning.

Definition 3 (Domain-variant Feature) *We define a feature subset of $\mathcal{K}$ as Ξ, which satisfies the following conditions:*

$$\Xi = \left\{k \in \mathcal{K} : \exists s, s' \in S, D_s^k \neq D_{s'}^k\right\}. \tag{9}$$

Then we call it domain-variant feature and refer to the subspace $\mathcal{Z}^{\Xi}$ as domain-variant subspace.

This feature captures information that varies across source domains, making it less conducive to the model's generalization in any domain.

In the following, we prove that there exists a latent space that can be represented as the direct sum of these three subspaces. We present the following theorem.

Theorem 1 *The space $\mathcal{Z}$ can be expressed as the direct sum of $\mathcal{Z}^{\Gamma}$, $\mathcal{Z}^{\Phi}$ and $\mathcal{Z}^{\Xi}$, i.e.*

$$\mathcal{Z} = \mathcal{Z}^{\Gamma} \oplus \mathcal{Z}^{\Phi} \oplus \mathcal{Z}^{\Xi}, \tag{10}$$

where the $\oplus$ represents direct sum operator.

The proof can be found in Appendix A.

Based on the properties of direct sum decomposition, we can directly obtain a unique data representation, which we call the Tri-space Latent Representation. So we give the following Corollary.

Corollary 1 (Tri-space Latent Representation) *For any feature vector $z \in \mathcal{Z}$, we can uniquely representation it into the sum of features in $\mathcal{Z}^{\Gamma}$, $\mathcal{Z}^{\Phi}$ and $\mathcal{Z}^{\Xi}$, i.e.*

$$z = \gamma + \phi + \xi, \tag{11}$$

where $\gamma \in \mathcal{Z}^{\Gamma}$, $\phi \in \mathcal{Z}^{\Phi}$ and $\xi \in \mathcal{Z}^{\Xi}$.

Many studies connect domain generalization with Causality Schölkopf et al. (2021); Kuang et al. (2020); Yang et al. (2021). They suggest that features can be categorized as causal or spurious correlation features, with the former benefiting and the latter impairing the model's domain generalization. For example, in Yang et al. (2021), causal features are defined as the Markov blanket MB(Y) of the label Y, which includes the parents (direct causes), children (direct effects), and spouses (other parents of the class variable's children) of the label Y. In Cai et al. (2019), features are encoded into domain latent variables and semantic

latent variables, with the assumption that the semantic latent variables capture the causal features. Our definition aligns closely with this idea, where feature γ represents causal features, while features ϕ and ξ represent spurious correlation features. Our contribution lies in further separating the spurious correlation features and linking them explicitly to domain invariance and diversity.

4.2 Fine-grained Target Domain Risk Bound for Domain Generalization

Before introducing our fine-grained target domain risk bound, we begin by stating the assumptions underlying our theoretical results. We then define a distributional discrepancy measure and a target-domain reference, both constructed from the feature Ξ.

Assumption 1 *The loss function $\ell(\cdot,\cdot)$ has a linear upper bound with respect to the prediction error on the compact set $[0,1]$, i.e. for any $x, y \in [0,1]$, $\ell(x,y) \leq \mathcal{L}|x-y|$, where the $\mathcal{L}$ is a constant.*

Remark. Most loss functions in machine learning satisfy this property, such as 0-1 loss, MSE loss and cross-entropy loss.

Assumption 2 *For each domain $d \in S \cup T$, the label y can be uniquely determined by the domain-invariant feature γ. In other words, there exists a deterministic function $f_d : \mathcal{Z} \to [0,1]$ that accurately captures the relationship between y and γ, i.e., $y = f_d(\gamma)$. Without loss of generality, we set $f_d(\phi) = f_d(\xi) = 0$.*

Remark. This assumption implies that optimal predictions for all domains rely solely on features that remain invariant across them. Although some literature Bui et al. (2021) discusses the role of domain-specific features—which vary across domains yet remain predictive—in domain generalization, such considerations are beyond the scope of Assumption 2 and are excluded from this discussion. We acknowledge this as a limitation of our study.

Definition 4 *We define a distribution discrepancy metric $d^{\Xi}_{\hat{\mathcal{H}}}(D, D')$ between two distributions D and D' in the feature subspace $\mathcal{Z}^{\Xi}$ as follows:*

$$d^{\Xi}_{\hat{\mathcal{H}}}(D, D') = \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Xi}} \left(|D(\xi) - D'(\xi)|\right) \left(|\hat{h}(\xi)|\right) L(\mathrm{d}\xi). \tag{12}$$

In this context, $\hat{\mathcal{H}} = \{h(3\cdot) : h(\cdot) \in \mathcal{H}\}$; $L(\cdot)$ denotes the Lebesgue measure. These notations will be used throughout the following sections.

We provide the following interpretation of our newly defined distribution discrepancy measure, $d^{\Xi}_{\hat{\mathcal{H}}}(\cdot,\cdot)$, which is motivated by our fine-grained framework. This measure acts as a "localized measure" specifically for the domain-variant subspace $\mathcal{Z}^{\Xi}$, directly assessing the potential impact arising from incorrect predictions due to the model's reliance on non-generalizable features ξ, when facing distribution shifts in the $\mathcal{Z}^{\Xi}$ space. It achieves this by quantifying a compound effect:

- The magnitude of distribution shift at ξ: $|D(\xi) - D'(\xi)|$.

- The model's sensitivity to this shift: $|\hat{h}(\xi)|$.

Conceptually, $|\hat{h}(\xi)|$ acts as a weighting coefficient applied to the distribution shift at frequency ξ, thereby specifically quantifying harmful distribution shifts—those that undermine generalization capability. Since we assume in Assumption 2 that the label function does not depend on the feature ξ (i.e., $f_d(\xi) = 0$), $|\hat{h}(\xi)|$ essentially measures the prediction error arising from the model's erroneous reliance on this feature. Its magnitude at any point directly determines how detrimental a distribution shift at that point is to generalization performance: a larger $|\hat{h}(\xi)|$ indicates a more damaging shift, whereas if $|\hat{h}(\xi)| = 0$, the model is insensitive to the shift at ξ and the shift does not affect domain generalization. Hence, we treat $|\hat{h}(\xi)|$ as a pointwise weighting coefficient that helps identify and quantify those distribution shifts that are most harmful to domain generalization.

Definition 5 *Leveraging the domain variant feature and the newly introduced distributional discrepancy measure, we redefine a reference for the target domain.*

$$U = \left\{ D : \sum_{s \in S} \pi_s d^{\Xi}_{\hat{\mathcal{H}}}(D, D^{\Xi}_s) \leq \max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_{s'}, D^{\Xi}_s) \right\}, \tag{13}$$

where D^{Ξ}_s is the distribution of feature Ξ in source domain s.

Remark. This definition is based on the fact that $D^{\Gamma}_s = D^{\Gamma}_{s'}$ and $D^{\Phi}_s = D^{\Phi}_{s'}$ for $s, s' \in S$, it is a direct simplification of Eq. (5). It essentially relaxes the constraint on the dimensions of feature γ and ϕ, thereby expanding the range of the set O, i.e. $O \subseteq U$.

Theorem 2 *Given the Assumption 1 and 2, for every hypothesis h in the class $\mathcal{H}$, the risk on any target domain $t \in T$ can be bounded as:*

$$\begin{aligned} \varepsilon_{D_t}(h) \leq & \underbrace{\sum_{s \in S} \pi_s \varepsilon_{D_s}(h)}_{\text{term 1}} + \underbrace{\mathcal{L}\left(\frac{1}{3} \min_{D \in U} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t, D)\right)}_{\text{term 2}} + \underbrace{\frac{1}{3} \max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s, D^{\Xi}_{s'})}_{\text{term 3}} \\ & + \underbrace{\frac{B}{3} \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi)}_{\text{term 4}} + \underbrace{\min\left\{ \sum_s \pi_s E_{D_t}[|f_s - f_t|], \sum_s \pi_s E_{D_s}[|f_s - f_t|] \right\}}_{\text{term 5}} \Big). \end{aligned} \tag{14}$$

Here, D^{Ξ}_s is the feature distribution of ξ on the feature subspace $\mathcal{Z}^{\Xi}$ in the source domain s. Similarly, in the following text, D^{Φ}_s and D^{Γ}_s are the distributions of ϕ and γ. $B = \sup_{\phi \in \mathcal{Z}^{\Phi}} |D^{\Phi}_t(\phi) - D^{\Phi}(\phi)|$ is a constant, where $D^{\Phi} = D^{\Phi}_s, s \in S$. We provide an approximate estimation method for B in Appendix I. Moreover, similar to Eq. (6), the weights $\{\pi_s\}_{s \in S} \subset \triangle_{|S|-1}$ are determined by the (D, U) that minimizes $\min_{D \in U} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t, D)$.

The proof can be found in Appendix B.

Our bound consists of five terms, each contributing to the arbitrary-target domain risk in a distinct manner. The simplest ones are the term 1 and term 5, which are present in previous works. The term 1 is the weighted average of source domain risks, reflecting the risk

when training on the source domains. The term 5 reflects the error arising from differences in labeling functions across domains, as also discussed in Albuquerque et al. (2019). In most scenarios within domain adaptation/generalization applications, this typically reduces to zero, as the covariate shift assumption is generally assumed to hold David et al. (2010).

In the following, we conduct a term-by-term analysis of the remaining three terms, examining how domain-invariant representation learning and domain augmentation algorithms respectively affect the values of these terms. Our analysis base on the mathematical definitions we established for these two algorithms.

4.3 A Unified Analysis of Domain Augmentation and Domain-Invariant Representation Learning

In this subsection, we leverage our derived bound to present a unified analysis of two algorithmic approaches: domain augmentation and domain-invariant representation learning. Thanks to the direct-sum decomposition of the latent space, our bound is more fine-grained than those in prior work. This allows for a clearer understanding of how different feature types affect generalization capabilities and helps to elucidate the underlying mechanisms of both approaches. We first provide formal mathematical definitions of the two algorithms. We then analyze the remaining three terms in our bound in light of these definitions.

4.3.1 Mathematical Definitions for Domain Augmentation and Domain-invariant Representation Learning

Since prior work has not offered formal definitions of domain augmentation and domain-invariant representation learning, we take the first step toward filling this gap by introducing precise definitions for both concepts.

Definition 6 (Domain Augmentation) *Let $\mathcal{Z}$ be the latent space (i.e., the output of Encoder). Let S be the set of source domains, and let $\hat{S} = \bigcup_{s\in S} \hat{S}_s$ be a finite sample set drawn from these domains, where $\hat{S}_s \subset \mathcal{Z}$ denotes the samples from a specific domain s. Let $g : \hat{S} \to S$ be the mapping that assigns each sample to its source domain, i.e., $g(z) = s$ for all $z \in \hat{S}_s$. Domain augmentation is a mapping on the latent space*

$$\begin{aligned} \mathrm{Aug} : \mathcal{Z} &\longrightarrow \mathcal{Z} \\ z &\longmapsto z_{\mathrm{aug}}, \end{aligned} \tag{15}$$

and satisfies

1. *$|\hat{S}_{\mathrm{aug}}| = |\hat{S}|$, where $\hat{S}_{\mathrm{aug}} = \mathrm{Aug}(\hat{S}), i.e., \{\mathrm{Aug}(z) : z \in \hat{S}\}$;*
2. *$S \subset S_{\mathrm{aug}} \subseteq S \cup T$, where $S_{\mathrm{aug}} = g(\mathrm{Aug}(\hat{S}))$ denotes the set of source domains after augmentation.*

Remark. These two conditions reflect the fundamental requirements of domain augmentation algorithms: 1. the original sample size remains unchanged: Since increasing sample size generally improves generalization regardless of the underlying function, therefore, we impose this condition to ensure that our demonstration of generalization capability is independent of the number of data samples used. 2. The diversity of sample domains increases

after the augmentation process. Because all augmentation operations are applied to the data samples $\hat{S}$, we seek to ensure that the underlying domain distributions responsible for generating these samples also become more diverse. We argue that this phenomenon reflects real-world conditions: for example, Zhou et al. (2024) suggests that differences in image domains are specifically reflected in image styles. Therefore, the MixStyle algorithm expands the range of sample styles by mixing style statistics from two samples belonging to different domains. Alternatively, Xu et al. (2021) argues that variations in image backgrounds lead to domain differences, so random convolutions are used to arbitrarily alter image backgrounds.

From the augmented source domain set S_{aug}, we obtain new augmented features corresponding to the domain-invariant, spurious-invariant, and domain-variant components. These features are defined as follows:

$$\Gamma_{\text{aug}} = \left\{k \in \mathcal{K} : D_s^k = D_t^k; \forall s \in S_{\text{aug}}, \forall t \in T\right\}, \tag{16}$$

$$\Phi_{\text{aug}} = \left\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k; \forall s, s' \in S_{\text{aug}}, \exists t \in T\right\}, \tag{17}$$

$$\Xi_{\text{aug}} = \left\{k \in \mathcal{K} : \exists s, s' \in S_{\text{aug}}, D_s^k \neq D_{s'}^k\right\}, \tag{18}$$

At the same time, Domain-invariant representation learning constrains the representation learning process through distribution-consistent regularization, enabling the encoder to obtain a representation space with distribution consistency. It can thus be defined as a mapping acting on the latent space, whose output is a new latent space with distribution consistency.

Definition 7 (Domain-invariant Representation Learning) *Let $\mathcal{Z}$ be the original latent space. Let $(\mathcal{Z}, \Sigma)$ be the corresponding measurable space, and let $\mathcal{P}(\mathcal{Z})$ denote the space of probability measures (distributions) on $(\mathcal{Z}, \Sigma)$. For each source domain $s \in S$, let $D_s \in \mathcal{P}(\mathcal{Z})$ be the probability measure induced by domain s on $\mathcal{Z}$. Domain-invariant representation learning is a measurable mapping*

$$\text{Inv} : (\mathcal{Z}, \Sigma) \to (\mathcal{Z}_{\text{inv}}, \Sigma_{\text{inv}}), \tag{19}$$

where $\mathcal{Z}_{\text{inv}} = \text{Inv}(\mathcal{Z})$ is the image of Inv*, and $\Sigma_{\text{inv}} = \{B \subseteq \mathcal{Z}_{\text{inv}} : \text{Inv}^{-1}(B) \in \Sigma\}$ is the σ-algebra induced by* Inv*. The mapping* Inv *must satisfy the following condition: its induced push-forward map*

$$\text{Inv}_{\#} : \mathcal{P}(\mathcal{Z}) \to \mathcal{P}(\mathcal{Z}_{\text{inv}}), \quad \text{Inv}_{\#}(D)(B) = D \circ \text{Inv}^{-1}(B), \quad \forall B \in \Sigma_{\text{inv}}, D \in \mathcal{P}(\mathcal{Z}) \tag{20}$$

makes all elements in the set $\{D_s\}_{s \in S}$ identical, i.e.,

$$\text{Inv}_{\#}(D_s) = \text{Inv}_{\#}(D_{s'}) \quad \forall s, s' \in S. \tag{21}$$

Building upon Theorem 1, the latent space $\mathcal{Z}_{\text{inv}}$ admits a direct sum decomposition

$$\mathcal{Z}_{\text{inv}} = \mathcal{Z}_{\text{inv}}^{\Gamma} \oplus \mathcal{Z}_{\text{inv}}^{\Phi} \oplus \mathcal{Z}_{\text{inv}}^{\Xi}. \tag{22}$$

To simplify our notation, we define $\mathcal{Z}_{\text{inv}}^J = \text{Inv}(\mathcal{Z}^J)$ for $J \in \{\Gamma, \Phi, \Xi\}$, each associated with the probability measure $(\text{Inv})_{\#} D_s^J$ (abbreviated as $D_s^{J_{\text{inv}}}$) for every source domain $s \in S$.

Just as in the domain augmentation setting, where we defined $\Gamma_{\text{aug}}, \Phi_{\text{aug}}, \Xi_{\text{aug}}$, we now introduce the corresponding feature subsets for domain-invariant features too: namely $\Gamma_{\text{inv}}, \Phi_{\text{inv}}$, and Ξ_{inv}. Subsequently, the target domain reference U, as defined in Definition 5, then becomes:

$$U_{\text{inv}} = \left\{ D : \sum_{s \in S} \pi_s \, d_{\hat{\mathcal{H}}}^{\Xi_{\text{inv}}}(D, D_s^{\Xi_{\text{inv}}}) \leq \max_{s,s' \in S} d_{\hat{\mathcal{H}}}^{\Xi_{\text{inv}}}(D_{s'}^{\Xi_{\text{inv}}}, D_s^{\Xi_{\text{inv}}}) \right\}. \tag{23}$$

Next, we examine how domain-invariant representation learning affects the term 2 of Eq. (14). By denoting $R(U, \Xi) = \min_{D \in U} d_{\hat{\mathcal{H}}}^{\Xi}(D_t^{\Xi}, D)$, we prove this relationship in Theorem 3.

Theorem 3 $R(U_{\text{inv}}, \Xi_{\text{inv}}) = 0$.

The proof can be found in Appendix C.

Theorem 3 is important because they show that after domain-invariant representation learning, i.e., $U \to U_{\text{inv}}$ and $\Xi \to \Xi_{\text{inv}}$, we can, at least theoretically, eliminate the term 2 in Eq. (14).

4.3.2 Domain Invariance and Domain Diversity terms

In the following, we describe in detail the remaining bounded terms in Eq. (14), namely "term 3" and "term 4". We derive a factor affecting domain generalization from each of these two terms and use them to analyze the roles of the two types of algorithms.

Definition 8 (Domain Invariance Factor, aka., "term 3") *We define the domain invariance factor R_Ξ as follows:*

$$R_\Xi = \max_{s,s' \in S} d_{\hat{\mathcal{H}}}^{\Xi}(D_s^{\Xi}(\xi), D_{s'}^{\Xi}(\xi)), \tag{24}$$

where D_s^{Ξ} is the feature distribution of ξ on the feature subspace $\mathcal{Z}^{\Xi}$ in the source domain s. Its definition is given by Definition 4. Similarly, in the following text, D_s^{Φ} and D_s^{Γ} are the distributions of ϕ and γ.

The term 3 in our bound includes a component (R_Ξ) that measures the maximum distribution discrepancy between source domains, calculated specifically using the feature ξ. Theoretically, it can be eliminated by aligning the source domain distributions in a common representation space during training – a process known as domain-invariant representation learning. Similar argument can be found in Zhao et al. (2019); Albuquerque et al. (2019) and Sicilia et al. (2023). This indicates that domain-invariant representation learning can not only eliminate $R(U, \Xi)$, but also remove $\max_{s,s'} d_{\hat{\mathcal{H}}}^{\Xi}(D_s^{\Xi}, D_{s'}^{\Xi})$. Both of these terms are solely related to feature ξ, reflecting the underlying principle of such methods, enhancing domain generalization by optimizing feature ξ.

Definition 9 (Domain Diversity Factor, aka., "term 4") *We define the domain diversity factor R_Φ as follows:*

$$R_\Phi = \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi). \tag{25}$$

The term 4 of our bound includes a component that captures the error induced by the feature ϕ in the subspace $\mathcal{Z}^{\Phi}$. In Theorem 4, we aim to formally quantify and prove its relationship with domain augmentation.

Theorem 4 *$R_{\Phi} \geq R_{\Phi_{\text{aug}}}$, the "$=$" is taken when $\Phi_{\text{aug}} = \Phi$.*

The proof can be found in Appendix D.

Theorem 4 states that as the diversity of source domains increases during training, the domain diversity factor may decrease. This result pinpoints the source of the improved generalization capability observed in domain augmentation algorithms, showing that optimizing the feature ϕ is the key contributing factor. However, the "$\geq$" relation in Theorem 4 provides only a partial conclusion—it ensures that domain augmentation does not increase the domain diversity factor, but does not guarantee a strict decrease. Thus, simply adding more source domains with different distributions does not necessarily improve domain generalization. In Theorem 4, we further identify the condition under which equality holds, namely $\Phi_{\text{aug}} = \Phi$. Intuitively, since domain augmentation increases the diversity of domain-related features—which include both ϕ and ξ—if the augmentation affects only ξ while leaving ϕ unchanged, the domain diversity factor will remain unchanged. Therefore, we conclude that the effectiveness of domain augmentation in improving domain generalization fundamentally depends on its ability to meaningfully diversify the feature ϕ.

Remark. This also highlights a trade-off between diversity and invariance. If augmentation is applied indiscriminately, it may not only fail to reduce R_{Φ} but also risk enlarging the domain-variant feature set Ξ_{aug}, potentially increasing the divergence term $\max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_{s}, D^{\Xi}_{s'})$ (i.e., R_{Ξ}), which reflects greater distribution discrepancy among source domains. Consequently, the net effect on the generalization bound is governed by the balance between the reduction in R_{Φ} and the potential increase in R_{Ξ}, underscoring the importance of designing augmentation strategies that selectively target spurious features while preserving domain-invariant structures.

Finally, and most importantly, the above analysis shows that domain-invariant representation learning and domain augmentation algorithms optimize different components of our bound, and neither can fully substitute for the other. This leads to the following new conclusions, which distinguish our work from prior studies:

Conclusion 1 *From the detailed analysis of each term in the domain generalization error bound in Theorem 2, we conclude that:*

(1) *Domain generalization performance is positively correlated with domain diversity.*

(2) *Domain-invariant representation learning and domain augmentation algorithms are complementary. More importantly, increasing the domain diversity of source domains via domain augmentation can help mitigate the limitations caused by spurious invariant features in domain-invariant representation learning.*

5. Theoretical Extensions

In Section 4.2, we present the main contribution of this paper—an upper bound on the risk over any target domain. This section extends our analysis through four structured subsections. First, we bridge Theorem 2 with classical generalization theory by introducing the concept of weighted Rademacher complexity, leading to a refined version of Theorem 2 applied to target domain risk. Second, we derive a lower bound for the arbitrary-target domain risk and analyze the tightness of the upper bound through comparative evaluation. Third, we establish a domain generalization error bound grounded in algorithmic stability. Finally, we discuss both theoretical and methodological similarities and differences between our framework and Invariant Risk Minimization.

5.1 Arbitrary-Target Domain Risk Bound with Weighted Rademacher Complexity

In Theorem 2, we use the weighted risk of source domains to bound the target domain risk. However, in practice, we can only obtain the empirical risk of the source domain, which detaches Theorem 2 from practical applications. In this subsection, we bridge Theorem 2 with classical generalization theory by using the weighted empirical risk of the source domain and Rademacher complexity to bound the arbitrary-target domain risk. By leveraging the concept of weighted Rademacher complexity (which has been established in prior work, e.g., Bartlett and Mendelson (2002); Koltchinskii (2001); Liu et al. (2015); Kuck et al. (2018); Sun et al. (2011)), we extend the methodology of complexity analysis to the scenario of domain generalization.

Since Rademacher complexity depends on the data distribution, it varies across different domains. Therefore, we adopt the weighted Rademacher complexity for source domains, a notion previously studied in the literature.

Definition 10 *For source domain set S, the weighted Rademacher complexity is defined as follow:*

$$\mathcal{R}_S^w(\mathcal{H}) = \sum_{s\in S} \pi_s E_{D_s}\big[\hat{\mathcal{R}}_s(\mathcal{H})\big] = \sum_{s\in S} \pi_s \mathcal{R}_s(\mathcal{H}), \tag{26}$$

where the $\mathcal{R}_s(\mathcal{H})$ is the Rademacher complexity on the source domain s, the $\hat{\mathcal{R}}_s(\mathcal{H})$ is the empirical Rademacher complexity defined as:

$$\hat{\mathcal{R}}_s(\mathcal{H}) = E_\sigma\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(z_{s,i})\Big]. \tag{27}$$

Based on the definition of weighted Rademacher complexity, the following theorem presents a version of Theorem 2 concerning the domain generalization error.

Theorem 5 *Given any δ within the range $(0, \frac{1}{|S|})$, for every hypothesis h in the class $\mathcal{H}$, and considering any target domain $t \in T$, the target domain risk $\varepsilon_{D_t}(h)$ can be bounded*

with a confidence level of at least $1-|S|\delta$. *This bound is expressed as:*

$$\varepsilon_{D_t}(h) \leq \sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h) + \mathcal{L}\Big(4\mathcal{R}_S^w(\mathcal{H}) + 6\sqrt{\frac{\ln(2/\delta)}{2n}} + \frac{1}{3}\min_{D\in U} d_{\hat{\mathcal{H}}}^{\Xi}(D_t^{\Xi}, D) + \frac{1}{3}\max_{s,s'} d_{\hat{\mathcal{H}}}^{\Xi}(D_s^{\Xi}, D_{s'}^{\Xi}) + \frac{B}{3}\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\hat{\mathcal{Z}}^{\Phi}}\hat{h}(\phi)L(\mathrm{d}\phi) + \min\big\{\sum_s \pi_s E_{D_t}[|f_s - f_t|], \sum_s \pi_s E_{D_s}[|f_s - f_t|]\big\}\Big). \tag{28}$$

The proof can be found in Appendix E.

5.2 The Lower Bound of Arbitrary-Target Domain Risk

For the sake of theoretical completeness, we derive a lower bound on the arbitrary-target domain risk based on the classical generalization lower bound framework from Anthony and Bartlett (2009). This lower bound reveals the fundamental limitations of domain generalization imposed by model complexity (VC dimension), the number of source domains, sample size and domain diversity. Compared to the upper bound (Theorem 5), this lower bound exhibits similar dependencies on these key parameters, providing a theoretical basis for evaluating the tightness of our risk bounds.

We first introduce the theoretical framework of this classic lower bound for generalization error.

Theorem 6 (Anthony and Bartlett (2009)) *Let* $\mathcal{H}$ *be a class of functions with VC dimension* $d_{VC} > 1$, *for any learning algorithm* A, *there exists a distribution* D *makes the following expression holds:*

$$P_{Z\sim D^n}\Big(\varepsilon_D(A(Z)) - \inf_{h\in\mathcal{H}}\varepsilon_D(h) > \sqrt{\frac{d_{VC}}{320n}}\Big) > \frac{1}{64}, \tag{29}$$

where $A(Z)$ *is the hypothesis output by the learning algorithm* A *based on the training sample set* Z *draw from* D.

To reasonably utilize this framework for deriving the lower bound, we additionally introduce two definitions.

Definition 11 *We define the domain diversity factor and a risk gap caused by distribution discrepancy relative to the algorithmic output* $A(\hat{S})$ *as follow:*

$$R_{\Phi}(A(\hat{S})) = \int_{\mathcal{Z}^{\Phi}} A(\hat{S})(\phi)L(\mathrm{d}\phi), \tag{30}$$

$$\Delta^{\Xi}(D_1^{\Xi}, D_2^{\Xi}) = \int_{\mathcal{Z}^{\Xi}} A(\hat{S})(\xi)\big(D_1^{\Xi}(\xi) - D_2^{\Xi}(\xi)\big)L(\mathrm{d}\xi). \tag{31}$$

Remark on the risk gap $\Delta^{\Xi}(\cdot,\cdot)$: This quantity differs from a standard distance measure. It is specifically designed to quantify the impact of a distribution shift on the model's risk within the domain-variant subspace $\mathcal{Z}^{\Xi}$, rather than to measure the distribution difference itself. Its value represents the expected change in the model's output when the feature distribution changes, thus capturing the interaction between the distribution shift and the model's reliance on domain-variant features.

Since deriving a lower bound is challenging, we make additional assumptions to streamline the derivation process.

Assumption 3 (Covariate Shift Assumption) *For any $s, s' \in S$, f_s and $f_{s'}$ are the labeling function of source domains s and s', for any possible target domain D_t:*

$$f_s(z) = f_{s'}(z) = f_t(z), \quad z \in \mathcal{Z}. \tag{32}$$

Remark. The covariate shift assumption is the most general setting in domain generalization and is widely applied in domain generalization works. It assumes that all domains adhere to identical classification rules, thereby ensuring consistency in their labeling functions. Based on this assumption, any distribution shift can be attributed to potential differences in the marginal feature distributions across individual domains, which is referred to as covariate shift.

Assumption 4 *For any $x, y \in [0,1]$, there exists a function of the prediction error that is differentiable at zero and evaluates to zero at that point, i.e. $\mathsf{L}(x-y) = \ell(x,y)$, $\mathsf{L}(0) = 0$ and $\mathsf{L}'(0) = \mathcal{L}_0'$.*

Assumption 5 *We assume that the hypothesis space $\mathcal{H}$ is a linear hypothesis class.*

Theorem 7 (Lower Bound of Target Domain Risk) *Under the Assumption 3, 4 and 5, if the VC dimension d of the hypothesis space $\mathcal{H}$ satisfies $d_{VC} > 1$, then for any learning algorithm A, there exists a set of source domain distributions $\{D_s, s \in S\}$ such that the following expression holds:*

$$P_{\hat{S}\sim(D_a)^{|S|n}}\left(\varepsilon_{D_t}(A(\hat{S})) > \Omega\right) > \frac{1}{64}, \tag{33}$$

the lower bound:

$$\Omega = \inf_{h\in\mathcal{H}} \varepsilon_{D_a}(h) + \sqrt{\frac{d_{VC}}{320|S|n}} + \max\left\{0, \quad \mathcal{L}_0'\bar{B}R_{\Phi}(A(\hat{S})) + \frac{\mathcal{L}_0'}{|S|}\sum_{s\in S}\Delta^{\Xi}(D_t^{\Xi}, D_s^{\Xi})\right\}, \tag{34}$$

where D_a is the mixture distribution of the source domains, i.e. $D_a = \frac{1}{|S|}\sum_{s\in S} D_s$. $\bar{B} = \sup_{\phi\in\mathcal{Z}^{\Phi}} |D_t^{\Phi}(\phi'') - D^{\Phi}(\phi'')|$ is a constant, where $\phi'' \in \Phi$ is a number generated by the Mean Value Theorems for Definite Integrals.

The proof can be found in Appendix F.

This lower bound demonstrates consistency with the upper bound (Theorem 5) in terms of key parameter dependencies, indicating that our risk bounds are tight with respect to these parameters. Specifically:

- For sample size n, both the upper and lower bounds contain an $O(1/\sqrt{n})$ term, which aligns with the optimal sample complexity in statistical learning theory.
- Regarding model complexity, the upper bound employs the weighted Rademacher complexity $\mathcal{R}_S^w(\mathcal{H})$, while the lower bound utilizes the VC dimension d_{VC} of the hypothesis space, both reflecting the impact of model complexity on generalization performance from different perspectives.
- The presence of the domain diversity-related term $R_\Phi(A(\hat{S}))$ in the lower bound exhibits similarity to the dependence on the diversity factor in the upper bound, collectively demonstrating the significant role of domain diversity in domain generalization.

5.3 Arbitrary-Target Domain Risk Bound with Algorithmic Stability

In this subsection, we conduct further analysis within the framework of algorithmic stability. Algorithmic stability is a fundamental concept in statistical learning theory that provides an alternative to uniform convergence-based generalization bounds. Unlike traditional approaches that depend on the complexity of the hypothesis class, stability-based bounds characterize generalization through the sensitivity of a learning algorithm to perturbations in the training dataset. Next, we present the classical notion of stability and its corresponding upper bounds.

Definition 12 (Uniform Stability with Replacement) *Let $\mathcal{Z}$ be an instance space, $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ a training set of n samples drawn i.i.d. from a distribution $\mathcal{D}$, and $S^{(i)} = (z_1, \ldots, z_i', \ldots, z_n)$ a modified training set where the i-th sample z_i is replaced by an independent sample $z_i' \sim \mathcal{D}$. A learning algorithm $A : \mathcal{Z}^n \to \mathcal{H}$ has β-uniform stability with replacement with respect to a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ if for all such pairs $(S, S^{(i)})$ and for all instances $z \in \mathcal{Z}$, we have:*

$$\forall S \in \mathcal{Z}^n, \forall i \in \{1, \ldots, n\}, \left|\ell(A(S), z) - \ell(A(S^{(i)}), z)\right| \leq \beta. \tag{35}$$

The parameter $\beta = \beta(n)$ typically depends on the sample size n and decreases as n increases.

The above concepts thus lead to a stability-based generalization error bound. The following classical theorem establishes the connection between uniform stability and generalization performance:

Theorem 8 (Bousquet and Elisseeff (2002)) *Let A be a learning algorithm with β-uniform stability with respect to a loss function ℓ that satisfies $0 \leq \ell(h, z) \leq M$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. Let $\hat{S} = (z_1, \ldots, z_n)$ be a training set drawn i.i.d. from distribution D, and let $A(\hat{S})$ be the hypothesis output by the algorithm. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the random draw of S, the following bound holds:*

$$\left|\varepsilon(A(\hat{S})) - \hat{\varepsilon}(A(\hat{S}))\right| \leq \beta + (2n\beta + M)\sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{36}$$

Next, we extend the aforementioned stability framework to domain generalization.

Theorem 9 (Stability-Based Domain Generalization Bound) *Let $S = \{s_1, \ldots, s_{|S|}\}$ be a collection of source domains, where each $\hat{S}_s = \{z_{s,1}, \ldots, z_{s,n}\}$ contains n i.i.d. samples from domain s, and $\hat{S} = \cup_{s\in S}\hat{S}_s$. Let A be a learning algorithm with β-uniform stability with replacement on the source domains with respect to a bounded loss function $0 \le \ell(h, z) \le M$. Then under the covariate shift assumption 3, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the multi-source sample $\hat{S}$, for any target domain $t \in T$ with distribution D_t, we have:*

$$\begin{aligned}\left|\varepsilon_{D_t}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right| \le \beta + (2|S|n\beta + M)\sqrt{\frac{\ln(1/\delta)}{2|S|n}} + \mathcal{L}\Big(\frac{1}{3}\min_{D\in U} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t, D) \\ + \frac{1}{3}\max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s, D^{\Xi}_{s'}) + \frac{B}{3}\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\hat{\mathcal{Z}}^{\Phi}} \hat{h}(\phi)L(\mathrm{d}\phi)\Big).\end{aligned} \tag{37}$$

The proof can be found in Appendix G.

5.4 Discussion: Connections and Distinctions with Invariant Risk Minimization

To situate our work more comprehensively within the broader landscape of domain generalization theory, this section compares our approach with the theory of Invariant Risk Minimization (IRM)—an important and influential framework. Invariant Risk Minimization (IRM) Arjovsky et al. (2019) provides a complementary perspective on domain generalization by leveraging principles of causal inference. Although IRM and our proposed framework share the fundamental objective of learning domain-invariant representations, they exhibit significant differences in their theoretical foundations and methodologies.

First, IRM is built upon causal inference principles, positing that optimal predictors should depend solely on causal features that remain invariant across different environments. The theoretical guarantees of IRM primarily focus on the identifiability of causal features—specifically, establishing conditions under which true causal features can be reliably recovered from observational data Kamath et al. (2021). Although IRM also aims to learn invariance across source domains, it achieves this by enforcing consistency among optimal predictors across source domains, which is independent of the distributional differences between them. In other words, IRM does not explicitly reduce inter-domain distribution shifts. In contrast, our theoretical analysis is grounded in a distributional divergence framework, which constitutes a fundamental distinction between IRM and our work.

Furthermore, another fundamental distinction lies in their conceptualization of domain diversity. IRM implicitly requires sufficient "environment diversity" as a prerequisite for causal feature identification. Conversely, our framework provides explicit quantification of domain diversity through the R_Φ factor, which appears directly in our generalization bounds. This quantitative characterization precisely explains how insufficient diversity leads to dependence on spurious invariant features $\phi \in \Phi$, thereby offering a mechanistic explanation for failures in invariant learning.

Third, although IRM presents compelling theoretical arguments for why invariant predictors should generalize well, its analysis does not yield explicit generalization error bounds

in the conventional statistical learning sense. Our framework addresses this gap by providing explicit bounds that incorporate both hypothesis class complexity (via Rademacher complexity) and domain characteristics (via R_Ξ and R_Φ), thereby firmly connecting domain generalization with established learning theory concepts.

Despite these fundamental differences, both frameworks offer complementary insights for algorithmic development. IRM's emphasis on classifier consistency across domains aligns with our conceptualization of domain-invariant representation learning. However, our Tri-Space decomposition provides additional guidance by explicitly distinguishing between different feature types, suggesting principled methodologies to overcome IRM's limitations when training domains exhibit limited diversity. Moreover, from a theoretical perspective, although they originate from different starting points, both frameworks contribute valuable viewpoints to a unified understanding of domain generalization. IRM provides causal intuition explaining why invariant features generalize effectively, while our Tri-Space framework offers a statistical characterization of how different features influence generalization risk. Together, they provide complementary theoretical tools that address both feature identifiability and generalization performance, thereby advancing the theoretical foundations of domain generalization research.

6. Experiments

In this section, we conduct three sets of experiments to validate our theory from three perspectives: (1) whether domain generalization performance improves with greater source domain diversity, (2) whether domain diversity and invariance independently contribute to domain generalization performance. Specifically, we investigate whether incorporating domain augmentation alongside domain invariance extraction can further enhance performance, and (3) whether the distribution alignment regularization in domain-invariant representation learning genuinely reduces distribution divergence measure. First, we provide a brief overview of the datasets used in these experiments.

Digits-DG Zhou et al. (2020) contains 4 different digit datasets including MNIST LeCun et al. (1998), MNIST-M Ganin and Lempitsky (2015), SVHN Netzer et al. (2011) and SYN Ganin and Lempitsky (2015), which differ drastically in font style, stroke color and background.

PACS Li et al. (2017) consists of four domains, namely Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images) and Sketch (3,929 images). Each domain contains seven categories.

Office-Home Venkateswara et al. (2017) contains four domains,which are Artistic, Clipart, Product and Real-World. Each domain has 65 classes, which are related to office and home objects. There are around 15,500 images in total. The domain variations mainly take place in background, view point and image style.

To facilitate a fair comparison with previous studies, we adopt the leave-one-domain-out protocol described in Li et al. (2017), Carlucci et al. (2019), and Li et al. (2019). In this method, one domain is designated as the test domain, while the remaining domains serve as source domains for model training. We set the batch size to 64 for each domain and kept all other hyperparameters unchanged. Results are averaged over three runs with dif-

Table 3: Target domain accuracy at varying levels of domain augmentation

Digits-DG / #Domain	MNIST	MNIST-M	SVHN	SYN	Average
3	91.90%	37.00%	48.03%	72.23%	62.29%
4	95.38%	59.96%	54.88%	77.42%	71.91%
6	96.82%	63.65%	64.93%	87.04%	78.11%
PACS / #Domain	**Sketch**	**Photo**	**Cartoon**	**Art**	**Average**
3	61.58%	95.86%	70.48%	72.63%	75.14%
4	62.39%	96.02%	73.36%	76.04%	76.95%
6	65.28%	96.10%	75.83%	80.12%	79.33%
Office-Home / #Domain	**Artistic**	**Clipart**	**Product**	**Real-World**	**Average**
3	58.10%	43.66%	71.15%	74.74%	61.91%
4	59.43%	43.96%	70.40%	74.26%	62.01%
6	59.73%	44.83%	71.75%	74.74%	62.76%

ferent random seeds, and all experimental code is implemented using the publicly available Dassl.pytorch toolbox.

6.1 Impact of Domain Diversity on Domain Generalization

In this subsection, we validate the impact of increasing domain diversity on domain generalization performance using experiments based on the MixStyle algorithm Zhou et al. (2024), a state-of-the-art domain augmentation technique. The core idea of MixStyle is to apply Mixup Zhang (2018) to the style statistics of samples from different domains within ResNet residual blocks. This generates new domain samples while maintaining the original sample count.

Applying MixStyle to different subsets of the source domain set S—each containing three domains—generates varying degrees of domain augmentation. For example, applying MixStyle to samples from two source domains generates one new domain, while applying it to all three produces samples representing three distinct new domains. In this experiment, we investigate how the number of newly generated domains affects domain generalization performance. Specifically, we compare the results of three experimental setups: (1) No MixStyle augmentation; (2) Applying MixStyle to two source domains; (3) Applying MixStyle to all three source domains.

The experimental results are summarized in Table 3. In Table 3, each row corresponds to an experimental setup, and each column represents a distinct target domain. Except for the cases where Product and Real-World serve as target domains in the Office-Home dataset, all experimental results consistently demonstrate that increasing domain diversity—by generating more new domains through MixStyle—enhances domain generalization performance. This effect is particularly pronounced in the Digits-DG dataset, where the number of newly introduced domains significantly influences test accuracy across all target domains. These findings provide strong empirical support for Theorem 4 and Conclusion 1

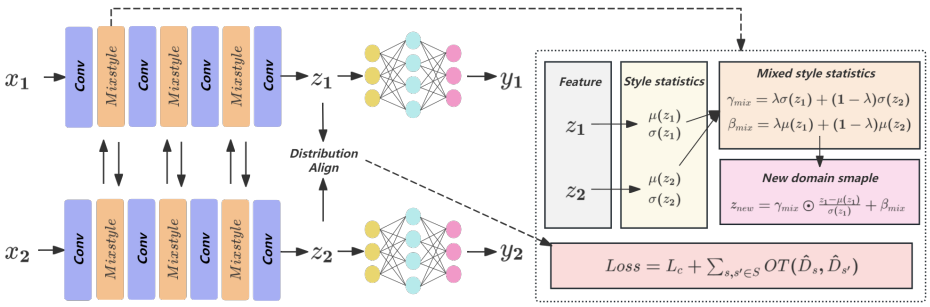


Figure 2: **Experiment Architecture**. The figure illustrates the experimental framework that simultaneously uses Mixstyle to enhance domain diversity and aligns the domain-wise OT distances to extract domain-invariant representations. μ and σ represent the style statistics of the samples, L_c denotes the classification loss, $\hat{D}_s$ and $\hat{D}_{s'}$ represent the distributions of samples from different source domains, and $OT(\cdot)$ refers to the optimal transport distance.

(1). However, when Product and Real-World are used as target domains, the improvements from MixStyle are less significant. This is likely due to the limited intrinsic domain diversity in these cases, which restricts the effectiveness of MixStyle in generating sufficiently diverse new domains to enhance generalization.

6.2 Complementary Effects of Domain Invariance and Diversity

In this subsection, we experimentally demonstrate that domain diversity and domain invariance exhibit complementary effects on domain generalization. Specifically, combining domain augmentation with domain invariance extraction further enhances domain generalization performance. The experiment utilizes MixStyle and Optimal Transport (OT) distance, also known as the Wasserstein distance, which is a mathematical framework for quantifying the discrepancy between probability distributions. This discrepancy is measured by calculating the minimum "transportation cost" required to transform one distribution into the other. To enhance domain diversity, MixStyle is applied between the residual blocks of the ResNet-18 convolutional layers. Additionally, by incorporating the OT distance between different domains—computed in the latent space after ResNet—into the loss function, we constrain the extraction of cross-domain invariant feature representations. The architecture of this experiment is illustrated in Figure 2.

We design four experimental setups to validate our conclusions by comparing their results: (1) using both MixStyle to augment domain diversity and OT distance to extract domain invariance; (2) using only MixStyle to augment domain diversity; (3) using only OT distance to extract domain invariance; and (4) using neither domain diversity augmentation nor domain invariance extraction. Each experiment was trained for 70 epochs, with testing conducted on the target domain after each epoch. The domain generalization error (the test error on the target domain minus the average training error on the source domain)

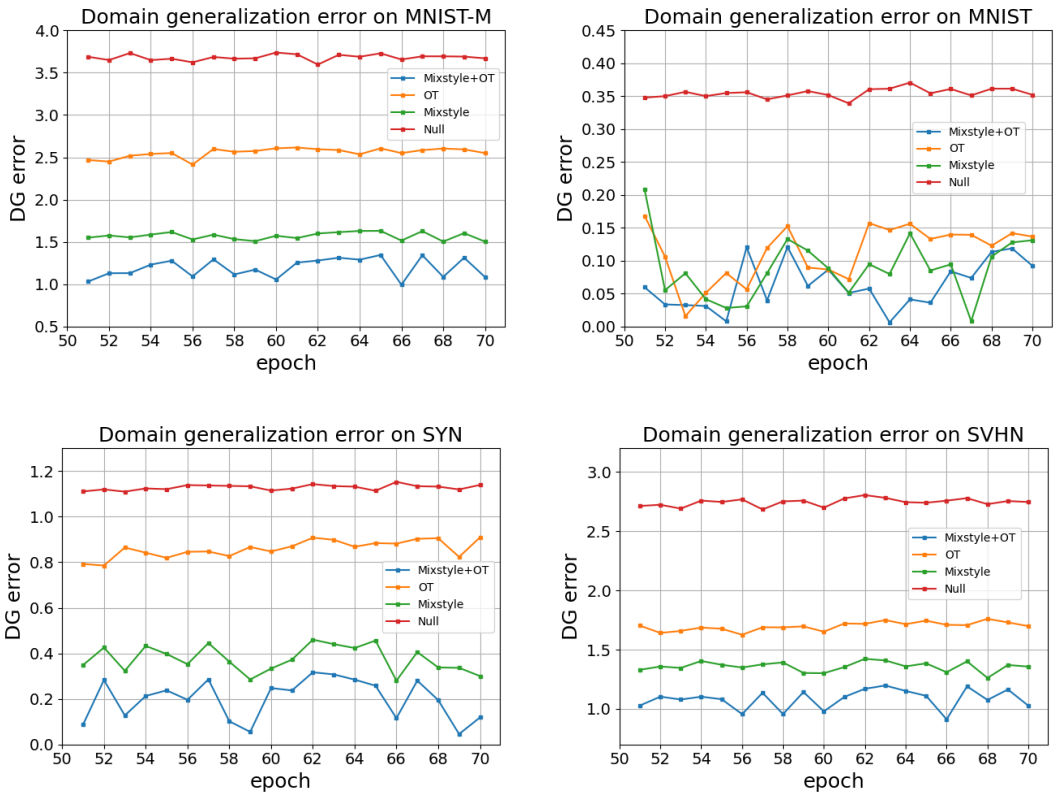


Figure 3: Comparison of Domain Generalization Errors of Four Different Methods on the Digits-DG Dataset

for epochs 51-70 was recorded. The experimental results are visualized as line plots in Figures 3, 4, and 5.

Figures 3, 4, and 5 display the domain generalization errors for setups (1), (2), (3), and (4) on the Digits-DG, PACS, and Office-Home datasets. Each figure plots the domain generalization error on the vertical axis against the number of epochs on the horizontal axis. The different experimental setups are represented by red, yellow, green, and blue curves. The red curve corresponds to the domain generalization error when MixStyle is used for domain augmentation while simultaneously aligning the source domain distributions using OT distance. The yellow curve shows the error when only MixStyle is applied, the green curve represents the error when only OT distance is used for aligning the source domain distributions, and the blue curve indicates the error when neither domain augmentation nor alignment is applied.

For the Digits-DG dataset (using MNIST-M, SVHN, and SYN domains as target domains), the PACS dataset (using Sketch and Cartoon as target domains), and the Office-Home dataset (using Artistic and Clipart as target domains), the results align with our conclusions. Specifically, setup (1) achieves a lower domain generalization error than both setups (2) and (3), while setups (2) and (3) outperform setup (4). For the Real-World

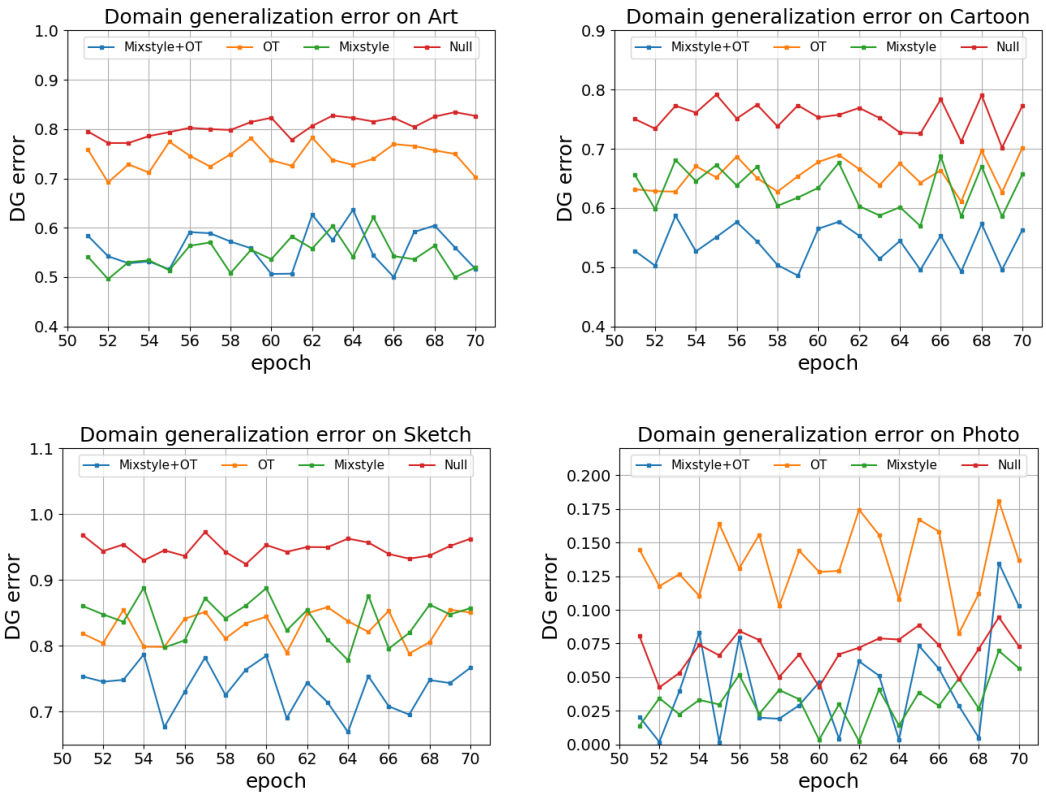


Figure 4: Comparison of Domain Generalization Errors of Four Different Methods on the PACS Dataset

domain in Office-Home, the results are generally consistent, though some fluctuations occur at certain points due to training instability, preventing the full realization of the expected outcome at those moments. The experimental results reveal several key insights. First, they demonstrate that both domain invariance extraction and domain diversity enhancement are effective strategies for domain generalization. Moreover, they confirm that these two factors independently contribute to improved domain generalization, highlighting their complementary roles. More importantly, the results suggest that enhancing source domain diversity through augmentation while simultaneously extracting domain-invariant features can significantly boost performance. This provides strong support for Conclusion 1 (2).

For the MNIST domain in Digits-DG and the Product domain in Office-Home, the positions of the yellow, green, and blue lines are nearly overlapping. This indicates that neither domain generalization method performs well in these cases. The domain generalization error in these cases likely arises from other factors, such as model complexity and sample randomness. For the Art domain in PACS, the experimental results generally align with our conclusions, though the blue and green lines fluctuate at roughly the same level. This suggests that for these target domains, extracting domain invariance using OT distance does not improve upon the performance achieved by MixStyle alone. We hypothesize that

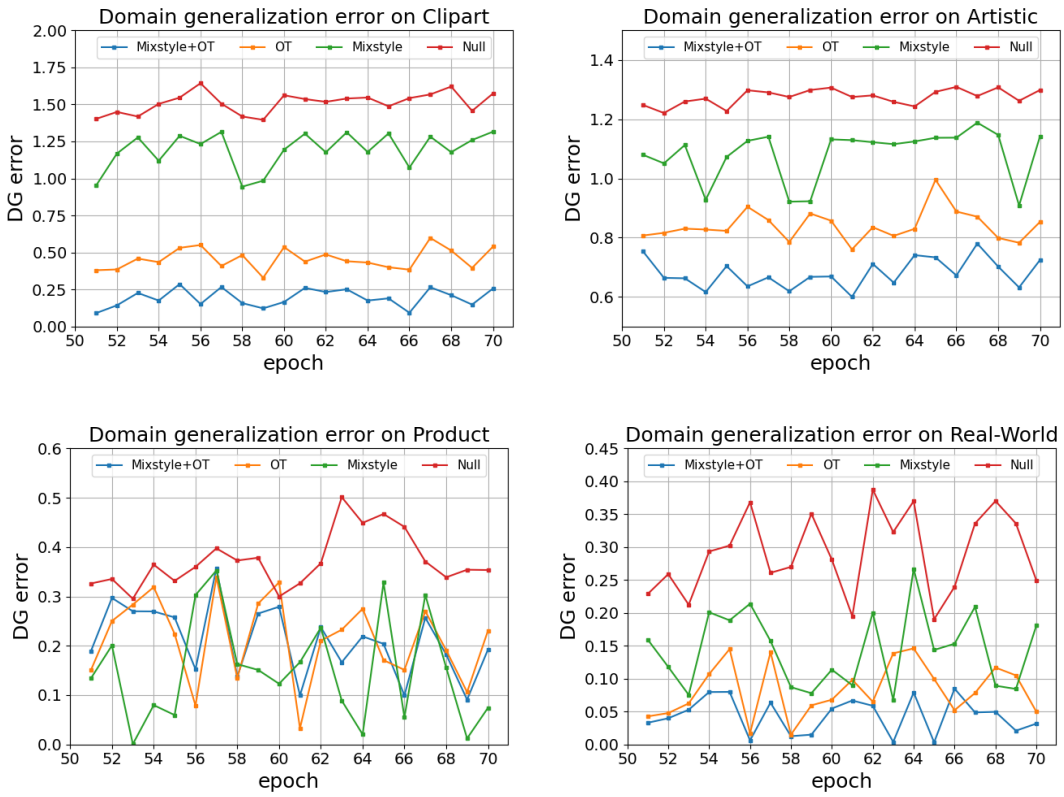


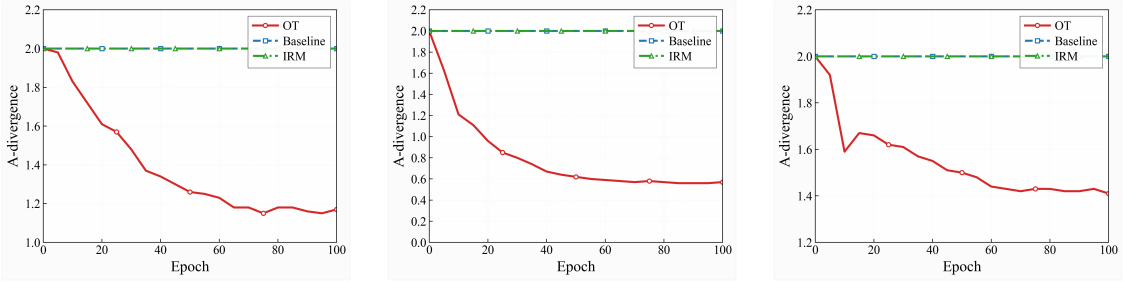
Figure 5: Comparison of Domain Generalization Errors of Four Different Setups on the Office-Home Dataset

this is due to the relatively small contribution of the domain invariance factor in these cases. However, for the Photo domain in PACS, the results show that extracting domain invariance using OT distance leads to a larger domain generalization error and a significantly higher training error. We speculate that this occurs because extracting domain invariance disrupts category-specific information that is beneficial for classification, leading to results that contradict our theoretical expectations.

6.3 Evolutionary Trends of Inter-Source Distribution Divergence under Domain-Invariant Learning Methods

In this subsection, we conducted additional experiments to analyze whether distribution consistency regularization in domain-invariant representation learning genuinely reduces distribution divergence measures. This investigation aims to provide empirical evidence for the fundamental mechanism of domain-invariant learning methods.

We compute the pairwise Optimal Transport (OT) distances between source domains and use their average as a regularization term to constrain the representation learning pro-



(a) PACS (b) Digits-DG (c) Office-Home

Figure 6: Evolutionary Trends of source domains $\mathcal{A}$-Divergence under Domain-Invariant Learning Methods.

cess, thereby enabling the latent space to extract cross-domain invariant feature representations. We monitor the changes in distribution divergence measures as training progresses.

Due to the computational challenges associated with $\mathcal{H}\Delta\mathcal{H}$-divergence, we employ the widely-used Proxy $\mathcal{A}$-divergence (PAD) as our empirical divergence measure. Following Ben-David et al. (2006), the $\mathcal{A}$-divergence is approximated by training a domain classifier to distinguish between two domains (e.g., labeling source domain samples as 0 and target domain samples as 1) and calculating based on the classifier's error rate: $d_{\mathcal{A}}(D_s, D_t) = 2(1-2\epsilon(h_d))$, where $\epsilon(h_d)$ is the error rate of the domain classifier h_d. A lower $\mathcal{A}$-divergence indicates greater similarity between domains, making them harder to distinguish.

Our experimental results, which illustrate the evolutionary trends of distribution divergence across the three datasets (PACS, Digits-DG, and Office-Home), are consolidated and presented in Figure 6. In each figure, the horizontal axis represents the training epoch, while the vertical axis denotes the $\mathcal{A}$-divergence. Each plot contains three curves in different colors, illustrating the evolution of the $\mathcal{A}$-divergence over 1–100 training epochs under different methods: the red curve corresponds to the method using OT distance as a source-domain distribution consistency regularizer, the green curve corresponds to the IRM algorithm (using the IRMv1 regularizer), and the blue curve corresponds to the baseline (ERM). The results demonstrate that the $\mathcal{A}$-divergence between source domains remains unchanged and consistently maximal for the baseline. In contrast, the source distribution consistency regularizer effectively reduces the $\mathcal{A}$-divergence, while the IRMv1 regularizer does not lead to such a reduction. This indicates that the invariance learned by IRM does not correspond to distributional invariance across source domains, which aligns with our discussion in Section 5.4.

7. Conclusion

This paper proposes a unified theoretical framework encompassing domain-invariant representation learning and domain augmentation methods. By introducing a theoretical framework of tri-space latent representation, we have, for the first time, derived a fine-grained

target domain risk bound capable of uniformly characterizing the roles of both domain invariance and domain diversity. This bound theoretically confirms that these two properties, previously considered conflicting, can and should coexist within the same generalization framework. Based on this fine-grained bound, we systematically elucidate the mechanisms through which these two classes of algorithms contribute to domain generalization, achieving theoretical unification. An important underpinning of this work is the axiomatic formalization of the two core paradigms through the definitions of Domain Augmentation and Domain-invariant Representation Learning, which provide the precise mathematical premises for our analysis. Experimental results validate our theoretical findings, confirming the significant interplay between domain invariance and diversity and their combined impact on domain generalization performance.

From a practical perspective, our theoretical bounds provide clear guidance for designing more effective domain generalization algorithms:

- **Feature Disentanglement:** The tri-space decomposition indicates that explicitly modeling and disentangling domain-invariant features γ, spurious invariant features ϕ, and domain-variant features ξ can yield more robust representations. To realize this structure, future algorithms could explore incorporating causally disentangled representations, using predefined causal graphs among the three feature types, class labels, and domain labels as a structural inductive bias. This provides a principled framework for achieving the decomposition through structural constraints or regularization terms.

- **Balanced Optimization:** Our bounds show that domain-invariant learning and domain augmentation optimize different terms within the risk bound. Therefore, a hybrid approach that simultaneously aligns source domain distributions (reducing R_{Ξ}) and enhances diversity (reducing R_{Φ}) is theoretically justified. How to reasonably combine these two methods for better domain generalization performance warrants further investigation. Based on our bound theory, one feasible path is to design an adaptively weighted multi-objective training strategy to dynamically balance the optimization goals of distribution alignment and diversity enhancement.

Despite these advances, this study has several limitations:

- **Deterministic Labeling Assumption:** We assume labels are completely determined by domain-invariant features (Assumption 2), which may not hold in real-world scenarios where labels depend on domain-specific cues.

- **Linear Hypothesis Class Limitation:** Some derivations (e.g., the lower bound proof) rely on a linear hypothesis class, which may limit direct applicability to deep nonlinear models.

- **Binary Classification Framework:** Our analysis is primarily developed for binary classification, as this represents the standard and most tractable setting for establishing core theoretical results in domain generalization/adaptation. Extending the exact form of our risk bound to multi-class settings would require the incorporation of concepts such as margin theory.

In summary, despite certain limitations, this work, by establishing a unified theory and formalizing core concepts, not only deepens the understanding of domain generalization but also lays a solid foundation for developing more effective and interpretable domain generalization methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62432006, U25A20529, 62376141, 62276159) and the Fundamental Research Program of Shanxi Province (No. 202303021223004).

References

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

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

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

Olivier Bousquet and Andr'e Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

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

Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, page 2060, 2019.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International conference on learning representations*, 2019.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 158–171. Springer, 2012.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

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

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4485–4492, 2020.

Jonathan Kuck, Ashish Sabharwal, and Stefano Ermon. Approximate inference via weighted rademacher complexity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

Da Li, Henry Gouk, and Timothy Hospedales. Finding lost dg: Explaining domain generalization via model complexity. 2022.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018a.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018b.

Jianwei Liu, Jiajia Zhou, and Xionglin Luo. Multiple source domain adaptation: A sharper bound using weighted rademacher complexity. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 546–553. IEEE, 2015.

Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *International Conference on Computational Learning Theory*, 2009.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45 (1):521–530, 2012.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

Xi Peng, Fengchun Qiao, and Long Zhao. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1775–1787, 2022.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124, 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *International conference on learning representations*, 2018.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve. *Machine Learning*, 112 (7):2685–2721, 2023.

Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. *Advances in neural information processing systems*, 24, 2011.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Ziqi Wang, Marco Loog, and Jan Van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9756–9763. IEEE, 2021.

Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *International conference on learning representations*, 2021.

Shuai Yang, Kui Yu, Fuyuan Cao, Lin Liu, Hao Wang, and Jiuyong Li. Learning causal representations for robust domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2750–2764, 2021.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021.

Hongyi Zhang. mixup: Beyond empirical risk minimization. *International conference on learning representations*, 2018.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR, 2019.

F Zhou, Z Jiang, C Shui, B Wang, and B Chaib-draa. Domain generalization with optimal transport and metric learning. *Neurocomputing*, 456:469–480, 2021.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13025–13032, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4396–4415, 2022.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132 (3):822–836, 2024.

Appendix A. Proof of Theorem 1

Proof The proof of theorem 1 is divided into four steps.

(1) The proof of "$\mathcal{K} = \Gamma \cup \Phi \cup \Xi$"

For the sake of proof, we further mathematically represent the definition formulas of the three feature subsets.

$$\begin{aligned}
\Gamma &= \left\{k \in \mathcal{K} : D_s^k = D_t^k; \forall s \in S, \forall t \in T\right\} \\
&= \bigcap_{s\in S}\bigcap_{t\in T}\{k \in \mathcal{K} : D_s^k = D_t^k\} \\
\Phi &= \left\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k; \forall s \in S, \exists t \in T\right\} \\
&= \bigcap_{s,s'\in S}\bigcup_{t\in T}\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k\} \\
\Xi &= \left\{k \in \mathcal{K} : \exists s, s' \in S, D_s^k \neq D_{s'}^k\right\} \\
&= \bigcup_{s,s'\in S}\{k \in \mathcal{K} : D_s^k \neq D_{s'}^k\}
\end{aligned} \tag{38}$$

Through these representation, we can derive the following relationship:

$$\begin{aligned}
&\complement_{\mathcal{K}}\Xi = \complement_{\mathcal{K}}\left(\bigcup_{s,s'\in S}\{k \in \mathcal{K} : D_s^k \neq D_{s'}^k\}\right) \\
&= \bigcap_{s,s'\in S}\complement_{\mathcal{K}}\{k \in \mathcal{K} : D_s^k \neq D_{s'}^k\} = \bigcap_{s,s'\in S}\{k \in \mathcal{K} : D_s^k = D_{s'}^k\} \\
&= \bigcap_{s,s'\in S}\left(\left(\bigcap_{t\in T}\{k \in \mathcal{K} : D_s^k = D_{s'}^k = D_t^k\}\right)\bigcup\left(\bigcup_{t\in T}\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k\}\right)\right) \\
&= \left(\bigcap_{s\in S}\bigcap_{t\in T}\{k \in \mathcal{K} : D_s^k = D_t^k\}\right)\bigcup\left(\bigcap_{s,s'\in S}\bigcup_{t\in T}\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k\}\right) \\
&= \Gamma \cup \Phi
\end{aligned} \tag{39}$$

where the symbols $\complement_B A$ represents the complement of subset A in the set B. So, $\mathcal{K} = \Gamma \cup \Phi \cup \Xi$.

(2) The proof of "$\Gamma \cap \Phi = \varnothing; \Gamma \cap \Xi = \varnothing; \Phi \cap \Xi = \varnothing$"

From (1), we know $\complement_{\mathcal{K}}\Xi = \Gamma \cup \Phi$, so $\Phi \cap \Xi = \varnothing$ and $\Gamma \cap \Xi = \varnothing$.

For any $k \in \Phi$, $\exists t \in T$, $\forall s \in S$, $D_s^k \neq D_t^k$, contradicts the condition of Γ, so $\Gamma \cap \Phi = \varnothing$.

(3) The proof of "$\mathcal{Z} = \mathcal{Z}^\Gamma + \mathcal{Z}^\Phi + \mathcal{Z}^\Xi$"

For any $z_1 \in \mathcal{Z}^\Gamma + \mathcal{Z}^\Phi + \mathcal{Z}^\Xi$, there exists $\gamma \in \mathcal{Z}^\Gamma$, $\phi \in \mathcal{Z}^\Phi$, and $\xi \in \mathcal{Z}^\Xi$, make the $z_1 = \gamma + \phi + \xi$. Because $\mathcal{Z}^\Gamma, \mathcal{Z}^\Phi, \mathcal{Z}^\Xi \subseteq \mathcal{Z}$, $\gamma, \phi, \xi \in \mathcal{Z}$, $\gamma + \phi + \xi \in \mathcal{Z}$, i.e. $z_1 \in \mathcal{Z}$. So, we have that $\mathcal{Z}^\Gamma + \mathcal{Z}^\Phi + \mathcal{Z}^\Xi \subseteq \mathcal{Z}$.

For any $z_2 \in \mathcal{Z}$, because $\mathcal{K} = \Gamma \cup \Phi \cup \Xi$ and $\Gamma \cap \Phi = \varnothing; \Gamma \cap \Xi = \varnothing; \Phi \cap \Xi = \varnothing$, we know that the Γ, Φ, Ξ represent different and all dimensions of the latent space $\mathcal{Z}$, so $\exists z_2^\gamma \in \mathcal{Z}^\Gamma$, $z_2^\phi \in \mathcal{Z}^\Phi$, $z_2^\xi \in \mathcal{Z}^\Xi$, make that $z_2 = z_2^\gamma + z_2^\phi + z_2^\xi$. So $z_2 \in \mathcal{Z}^\Gamma + \mathcal{Z}^\Phi + \mathcal{Z}^\Xi$, and $\mathcal{Z} \subseteq \mathcal{Z}^\Gamma + \mathcal{Z}^\Phi + \mathcal{Z}^\Xi$.

In summary, we have $\mathcal{Z} = \Gamma + \Phi + \Xi$.

(4) The proof of "$\mathcal{Z} = \mathcal{Z}^\Gamma \oplus \mathcal{Z}^\Phi \oplus \mathcal{Z}^\Xi$"

Because $\mathcal{Z}^\Gamma, \mathcal{Z}^\Phi, \mathcal{Z}^\Xi$ are feature subspaces of $\mathcal{Z}$ with different dimensions, we have that $\mathcal{Z}^\Gamma \cap \mathcal{Z}^\Phi = \{0\}$, $\mathcal{Z}^\Gamma \cap \mathcal{Z}^\Xi = \{0\}$, $\mathcal{Z}^\Xi \cap \mathcal{Z}^\Phi = \{0\}$, and $\mathcal{Z} = \mathcal{Z}^\Gamma + \mathcal{Z}^\Phi + \mathcal{Z}^\Xi$, we have

$\mathcal{Z} = \mathcal{Z}^{\Gamma} \oplus \mathcal{Z}^{\Phi} \oplus \mathcal{Z}^{\Xi}$. ■

Appendix B. Proof of Theorem 2

Proof For any δ within the range $(0, \frac{1}{|S|})$ and every hypothesis h in the class $\mathcal{H}$, with a confidence level of at least $1 - |S|\delta$ we have that:

$$
\begin{aligned}
&\Big|\varepsilon_{D_t}(h) - \sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&= \Big|\int_{\mathcal{Z}} D_t(z)\ell(h(z), f_t(z))L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\ell(h(z), f_s(z))L(\mathrm{d}z)\Big| \\
&\le \underbrace{\mathcal{L}\Big|\int_{\mathcal{Z}} D_t(z)\Big(|h(z) - f_t(z)|\Big)L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_t(z)|\Big)L(\mathrm{d}z)\Big|}_{A} \\
&\quad + \underbrace{\mathcal{L}\Big|\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_t(z)|\Big)L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_s(z)|\Big)L(\mathrm{d}z)\Big|}_{B},
\end{aligned}
\tag{40}
$$

the "$\le$" is due to the Assumption 1. In addition, we have that:

$$
\begin{aligned}
&\Big|\varepsilon_{D_t}(h) - \sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&\le \Big|\int_{\mathcal{Z}} D_t(z)\ell(h(z), f_t(z))L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\ell(h(z), f_s(z))L(\mathrm{d}z)\Big| \\
&\le \underbrace{\mathcal{L}\Big|\int_{\mathcal{Z}} D_t(z)\Big(|h(z) - f_t(z)|\Big)L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z) - f_s(z)|\Big)L(\mathrm{d}z)\Big|}_{C} \\
&\quad + \underbrace{\mathcal{L}\Big|\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z) - f_s(z)|\Big)L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_s(z)|\Big)L(\mathrm{d}z)\Big|}_{D}.
\end{aligned}
\tag{41}
$$

Next, we process the four terms (A, B, C, D) on the R.H.S. of the Eq. (40) and Eq. (41), respectively.

(A) First, we process the formula indexed by A in Eq. (40). The labeling function acts as the true predictive function between the data and the label. Its classification rule should depend solely on feature γ for prediction, and be independent of domain-related features

(such as ϕ and ξ). Therefore, we use $f_t(\gamma)$ to replace $f_t(z)$.

$$\begin{aligned}
&\Big|\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)-\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)\Big|\\
=&\Big|\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)-\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)\Big|\\
\leq&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)\\
=&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_t(\gamma)|\Big)L(\mathrm{d}z)
\end{aligned}\tag{42}$$

Because feature Γ,Φ,Ξ is from different dimensions, and they encompass all the features, we can represent the space as a direct product, i.e. $\mathcal{Z}=\mathcal{Z}^{\Gamma}\times\mathcal{Z}^{\Phi}\times\mathcal{Z}^{\Xi}$. Then we use the Fubini's theorem to this formula:

$$\begin{aligned}
&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_t(\gamma)|\Big)L(\mathrm{d}z)\\
=&\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Gamma}\times\mathcal{Z}^{\Phi}\times\mathcal{Z}^{\Xi}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_t(\gamma)|\Big)L(\mathrm{d}z)\\
=&\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Gamma}}\int_{\mathcal{Z}^{\Phi}}\int_{\mathcal{Z}^{\Xi}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_t(\gamma)|\Big)L(\mathrm{d}\gamma)L(\mathrm{d}\xi)L(\mathrm{d}\phi)\\
=&\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Gamma}}L(\mathrm{d}\gamma)\int_{\mathcal{Z}^{\Phi}}L(\mathrm{d}\phi)\int_{\mathcal{Z}^{\Xi}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_t(\gamma)|\Big)L(\mathrm{d}\xi)
\end{aligned}\tag{43}$$

Based on the linear property of the fully connected layer and the convex function property of the activation function, we have that:

$$\begin{aligned}
&\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Gamma}}L(\mathrm{d}\gamma)\int_{\mathcal{Z}^{\Phi}}L(\mathrm{d}\phi)\int_{\mathcal{Z}^{\Xi}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_t(\gamma)|\Big)L(\mathrm{d}\xi)\\
\leq&\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Gamma}}L(\mathrm{d}\gamma)\int_{\mathcal{Z}^{\Phi}}L(\mathrm{d}\phi)\int_{\mathcal{Z}^{\Xi}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(3\gamma)\\
&+h(3\phi)+h(3\xi)-f_t(\gamma)|\Big)L(\mathrm{d}\xi)\\
\leq&\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Gamma}}\Big(\Big|D_t^{\Gamma}(\gamma)-D_s^{\Gamma}(\gamma)\Big|\Big)\Big(|h(3\gamma)-f_t(\gamma)|\Big)L(\mathrm{d}\gamma)+\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Phi}}\Big(\Big|D_t^{\Phi}(\phi)\\
&-D_s^{\Phi}(\phi)\Big|\Big)\Big(|h(3\phi)|\Big)L(\mathrm{d}\phi)+\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^{\Xi}}\Big(\Big|D_t^{\Xi}(\xi)-D_s^{\Xi}(\xi)\Big|\Big)\Big(|h(3\xi)|\Big)L(\mathrm{d}\xi).
\end{aligned}\tag{44}$$

In the derivation process above, $L(\cdot)$ denotes the Lebesgue measure. Symbols $\{D_s^J : s \in S, J \in \{\Gamma, \Phi, \Xi\}\}$ represent the source domain distributions of the variables γ, ϕ, ξ. $\{D_t^J : J \in \{\Gamma, \Phi, \Xi\}\}$ are the same meaning but in target domain. Building upon the distribution properties of the three feature subspaces in Section 4.1, we can further process Eq. (44) in three steps.

(1) Since, for any $s \in S$, $D_s^\Gamma = D_t^\Gamma$, we have $|D_t^\Gamma(\gamma) - D_s^\Gamma(\gamma)| = 0$ for any $\gamma \in \mathcal{Z}^\Gamma$, so,

$$\sum_{s\in S} \pi_s \int_{\mathcal{Z}^\Gamma} \Big(\Big|D_t^\Gamma(\gamma) - D_s^\Gamma(\gamma)\Big|\Big)\Big(|h(3\gamma) - f_t(\gamma)|\Big) L(\mathrm{d}\gamma) = 0 \tag{45}$$

So, the second term of Eq. (44) is 0.

(2) We define that $\hat{h}(\cdot) = h(3\cdot)$, and a function space $\hat{\mathcal{H}} = \{h(3\cdot) : h(\cdot) \in \mathcal{H}\}$, according Definition 4, we have that:

$$\begin{aligned}
&\sum_{s\in S} \pi_s \int_{\mathcal{Z}^\Xi} \Big(\Big|D_t^\Xi(\xi) - D_s^\Xi(\xi)\Big|\Big)\Big(|h(3\xi)|\Big) L(\mathrm{d}\xi) \\
\leq &\sum_{s\in S} \pi_s \sup_{\hat{h}\in\hat{\mathcal{H}}} \int_{\mathcal{Z}^\Xi} \Big(\Big|D_t^\Xi(\xi) - D_s^\Xi(\xi)\Big|\Big)\Big(|\hat{h}(\xi)|\Big) L(\mathrm{d}\xi) \\
= &\sum_{s\in S} \pi_s d_{\hat{\mathcal{H}}}^\Xi(D_s^\Xi, D_t^\Xi)
\end{aligned} \tag{46}$$

We define that $D_{\min}^\Xi = \mathrm{argmin}_{D\in U}\, d_{\hat{\mathcal{H}}}^\Xi(D, D_t^\Xi)$, and then:

$$\begin{aligned}
\sum_{s\in S} \pi_s d_{\hat{\mathcal{H}}}^\Xi(D_s^\Xi, D_t^\Xi) &\leq \sum_{s\in S} \pi_s d_{\hat{\mathcal{H}}}^\Xi(D_{\min}^\Xi, D_t^\Xi) + \sum_{s\in S} \pi_s d_{\hat{\mathcal{H}}}^\Xi(D_s^\Xi, D_{\min}^\Xi) \\
&\leq \min_{D\in U} d_{\hat{\mathcal{H}}}^\Xi(D, D_t^\Xi) + \max_{s,s'} d_{\hat{\mathcal{H}}}^\Xi(D_s^\Xi, D_{s'}^\Xi)
\end{aligned} \tag{47}$$

Then combining Eq. (46) and Eq. (47), we set the upper bound of the last term in Eq. (44) as:

$$\frac{1}{3}\min_{D\in U} d_{\hat{\mathcal{H}}}^\Xi(D, D_t^\Xi) + \frac{1}{3}\max_{s,s'} d_{\hat{\mathcal{H}}}^\Xi(D_s^\Xi, D_{s'}^\Xi) \tag{48}$$

(3) Since $D_s^\Phi = D_{s'}^\Phi; D_t^\Phi \neq D_s^\Phi$ for any $s \in S$, we define $D^\Phi = D_s^\Phi$ for any $s \in S$.

$$\begin{aligned}
&\sum_{s\in S} \pi_s \int_{\mathcal{Z}^\Phi} \Big(\Big|D_t^\Phi(\phi) - D_s^\Phi(\phi)\Big|\Big)\Big(|h(3\phi)|\Big) L(\mathrm{d}\phi) \\
= &\int_{\mathcal{Z}^\Phi} \Big(\Big|D_t^\Phi(\phi) - D^\Phi(\phi)\Big|\Big)\Big(|h(3\phi)|\Big) L(\mathrm{d}\phi) \\
\leq &\sup_{\hat{h}\in\hat{\mathcal{H}}} \int_{\mathcal{Z}^\Phi} \Big(\Big|D_t^\Phi(\phi) - D^\Phi(\phi)\Big|\Big)\Big(|\hat{h}(\phi)|\Big) L(\mathrm{d}\phi).
\end{aligned} \tag{49}$$

Because the hypothesis function $\hat{h}(\phi) \geq 0$, we can apply the Mean Value Theorems for Definite Integrals to the integration of above equation:

There exists $\phi' \in \mathcal{Z}^{\Phi}$,

$$\begin{aligned}
&\int_{\mathcal{Z}^{\Phi}} \Big(\Big|D_t^{\Phi}(\phi) - D^{\Phi}(\phi)\Big|\Big)\Big(|\hat{h}(\phi)|\Big)L(\mathrm{d}\phi) \\
&= \Big(\Big|D_t^{\Phi}(\phi') - D^{\Phi}(\phi')\Big|\Big)\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi)L(\mathrm{d}\phi) \\
&\leq B\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi)L(\mathrm{d}\phi),
\end{aligned} \tag{50}$$

where $B = \sup_{\phi\in\mathcal{Z}^{\Phi}} |D_t^{\Phi}(\phi) - D^{\Phi}(\phi)|$ is a constant.
Combining Eq. (49) and Eq. (50), we set the upper bound of the third term in Eq. (44) as:

$$\frac{1}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi)L(\mathrm{d}\phi) \tag{51}$$

Finally, combining Eq. (44), Eq. (48) and Eq. (51), the formula indexed by B in Eq. (40) has a upper bound:

$$\frac{1}{3}\min_{D\in U} d_{\hat{\mathcal{H}}}^{\Xi}(D, D_t^{\Xi}) + \frac{1}{3}\max_{s,s'} d_{\hat{\mathcal{H}}}^{\Xi}(D_s^{\Xi}, D_{s'}^{\Xi}) + \frac{1}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi)L(\mathrm{d}\phi) \tag{52}$$

(B) Then, we process the formula indexed by B in Eq. (40).

$$\begin{aligned}
&\Big|\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_t(z)|\Big)L(\mathrm{d}z) - \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_s(z)|\Big)L(\mathrm{d}z)\Big| \\
&\quad\leq \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big|\Big(|h(z) - f_t(z)|\Big) - \Big(|f_s(z) - h(z)|\Big)\Big|L(\mathrm{d}z) \\
&\quad\leq \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z) - f_t(z) + f_s(z) - h(z)|\Big)L(\mathrm{d}z) \\
&\quad= \sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|f_t(z) - f_s(z)|\Big)L(\mathrm{d}z) \\
&\quad= \sum_{s\in S}\pi_s E_{D_s}[|f_t(z) - f_s(z)|]
\end{aligned} \tag{53}$$

(C) The simplification of formula C is similar to that of formula B.

$$
\begin{aligned}
&\Big|\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)-\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_s(z)|\Big)L(\mathrm{d}z)\Big| \\
=&\Big|\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_t(z)|\Big)L(\mathrm{d}z)-\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_s(z)|\Big)L(\mathrm{d}z)\Big| \\
&\quad\le\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big|\Big(|h(z)-f_t(z)|\Big)-\Big(|f_s(z)-h(z)|\Big)\Big|L(\mathrm{d}z) \\
&\quad\le\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_t(z)+f_s(z)-h(z)|\Big)L(\mathrm{d}z) \\
&\quad=\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|f_t(z)-f_s(z)|\Big)L(\mathrm{d}z) \\
&\quad=\sum_{s\in S}\pi_s E_{D_t}[|f_t(z)-f_s(z)|]
\end{aligned}
\tag{54}
$$

(D) Since the derivation of formula D differs from that of formula A only in the labeling function, their derivations are similar in many respects. Here, we provide a brief outline of its derivation process.

$$
\begin{aligned}
&\Big|\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_t(z)\Big(|h(z)-f_s(z)|\Big)L(\mathrm{d}z)-\sum_{s\in S}\pi_s\int_{\mathcal{Z}} D_s(z)\Big(|h(z)-f_s(z)|\Big)L(\mathrm{d}z)\Big| \\
\le&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(z)-f_s(z)|\Big)L(\mathrm{d}z) \\
=&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_s(\gamma)|\Big)L(\mathrm{d}z)
\end{aligned}
\tag{55}
$$

Based on the linear property of the fully connected layer and the convex function property of the activation function, we have that:

$$
\begin{aligned}
&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_s(\gamma)|\Big)L(\mathrm{d}z) \\
\le&\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^\Gamma}\Big(|D_t^\Gamma(\gamma)-D_s^\Gamma(\gamma)|\Big)\Big(|h(3\gamma)-f_s(\gamma)|\Big)L(\mathrm{d}\gamma) \\
&+\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^\Phi}\Big(|D_t^\Phi(\phi)-D_s^\Phi(\phi)|\Big)\Big(|h(3\phi)|\Big)L(\mathrm{d}\phi) \\
&+\frac{1}{3}\sum_{s\in S}\pi_s\int_{\mathcal{Z}^\Xi}\Big(|D_t^\Xi(\xi)-D_s^\Xi(\xi)|\Big)\Big(|h(3\xi)|\Big)L(\mathrm{d}\xi).
\end{aligned}
\tag{56}
$$

Simplifying the three terms on the right-hand side of the Eq. (56) yields:

$$\begin{aligned}&\sum_{s\in S}\pi_s\int_{\mathcal{Z}}\Big(|D_t(z)-D_s(z)|\Big)\Big(|h(\gamma+\phi+\xi)-f_s(\gamma)|\Big)L(\mathrm{d}z)\\&\leq\frac{1}{3}\min_{D\in U}d^{\Xi}_{\hat{\mathcal{H}}}(D,D^{\Xi}_t)+\frac{1}{3}\max_{s,s'}d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s,D^{\Xi}_{s'})+\frac{1}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^{\Phi}}\hat{h}(\phi)L(\mathrm{d}\phi).\end{aligned}\tag{57}$$

Based on the above, we obtain the risk bound on target domain t as:

$$\begin{aligned}\varepsilon_{D_t}(h)\leq&\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)+\frac{\mathcal{L}}{3}\min_{D\in U}d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t,D)+\frac{\mathcal{L}}{3}\max_{s,s'}d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s,D^{\Xi}_{s'})\\&+\frac{\mathcal{L}}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^{\Phi}}\hat{h}(\phi)L(\mathrm{d}\phi)+\mathcal{L}\min\big\{\sum_s\pi_sE_{D_t}[|f_s-f_t|],\sum_s\pi_sE_{D_s}[|f_s-f_t|]\big\}.\end{aligned}\tag{58}$$

■

Appendix C. Proof of Theorem 3

Proof Let $\mathcal{Z}_{\text{inv}}$ be the invariant latent space obtained through domain-invariant representation learning $\text{Inv}(\cdot)$. According to Definition 7, we have the distribution invariance condition:

$$(\text{Inv})_{\#}D_s=(\text{Inv})_{\#}D_{s'},\quad\forall s,s'\in S,\tag{59}$$

where D^{Z}_s denotes the distribution over $\mathcal{Z}$ for source domain s.

Let D^{inv}_s denote the distribution over $\mathcal{Z}_{\text{inv}}$ for source s, i.e., $D^{\text{inv}}_s=(\text{Inv})_{\#}D_s$. The above condition implies:

$$D^{\text{inv}}_s=D^{\text{inv}}_{s'},\quad\forall s,s'\in S.\tag{60}$$

By Corollary 1, each $z_{\text{inv}}\in\mathcal{Z}_{\text{inv}}$ admits a unique Tri-space decomposition:

$$z_{\text{inv}}=\gamma_{\text{inv}}+\phi_{\text{inv}}+\xi_{\text{inv}},\tag{61}$$

with $\gamma_{\text{inv}}\in\mathcal{Z}^{\Gamma}_{\text{inv}}$, $\phi_{\text{inv}}\in\mathcal{Z}^{\Phi}_{\text{inv}}$, $\xi_{\text{inv}}\in\mathcal{Z}^{\Xi}_{\text{inv}}$.

Let $D^{\Gamma_{\text{inv}}}_s$, $D^{\Phi_{\text{inv}}}_s$, and $D^{\Xi_{\text{inv}}}_s$ denote the distributions of these subspaces. Since D^{inv}_s is invariant across sources, and

$$D^{\Gamma_{\text{inv}}}_s=D^{\Gamma_{\text{inv}}}_{s'},\quad D^{\Phi_{\text{inv}}}_s=D^{\Phi_{\text{inv}}}_{s'},\quad\forall s,s'\in S.\tag{62}$$

we have:

$$D^{\Xi_{\text{inv}}}_s=D^{\Xi_{\text{inv}}}_{s'},\quad\forall s,s'\in S.\tag{63}$$

However, by definition, Ξ_{inv} corresponds to domain-variant features. The equality $D^{\Xi_{\text{inv}}}_s=D^{\Xi_{\text{inv}}}_{s'}$ for all s,s' contradicts the very notion of domain-variance, unless $\mathcal{Z}^{\Xi}_{\text{inv}}$ is trivial. Thus we must have:

$$\mathcal{Z}^{\Xi}_{\text{inv}}=\{0\},\quad\text{i.e., }\Xi_{\text{inv}}=\emptyset.\tag{64}$$

Now, consider the risk term $R(U_{\text{inv}}, \Xi_{\text{inv}})$:

$$\begin{aligned}
R(U_{\text{inv}}, \Xi_{\text{inv}}) &= \min_{D \in U_{\text{inv}}} d_{\hat{\mathcal{H}}}^{\Xi_{\text{inv}}}(D_t^{\Xi_{\text{inv}}}, D) \\
&= \min_{D \in U_{\text{inv}}} d_{\hat{\mathcal{H}}}^{\Xi_{\text{inv}}}((\text{Inv})_{\#} D_t^{\Xi}, D) \\
&= \min_{D \in U_{\text{inv}}} \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Xi_{\text{inv}}}} \left| \hat{h}(\xi)\, d\big((\text{Inv})_{\#} D_t^{\Xi}\big)(\xi) - \hat{h}(\xi)\, dD(\xi) \right| \\
&= \min_{D \in U_{\text{inv}}} \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Xi_{\text{inv}}}} \left| \hat{h}(\xi) \right| \left| D_t^{\Xi} \circ \text{Inv}^{-1}(\xi) - D(\xi) \right| L(\mathrm{d}\xi) \qquad (65) \\
&\leq \min_{D \in U_{\text{inv}}} \sup_{\hat{h} \in \hat{\mathcal{H}}} \max_{\xi \in \mathcal{Z}^{\Xi_{\text{inv}}}} \left| D_t^{\Xi} \circ \text{Inv}^{-1}(\xi) - D(\xi) \right| \left| \hat{h}(\xi) \right| \cdot L(\mathcal{Z}^{\Xi_{\text{inv}}}) \\
&= \min_{D \in U_{\text{inv}}} \sup_{\hat{h} \in \hat{\mathcal{H}}} \max_{\xi \in \mathcal{Z}^{\Xi_{\text{inv}}}} \left| D_t^{\Xi} \circ \text{Inv}^{-1}(\xi) - D(\xi) \right| \left| \hat{h}(\xi) \right| \cdot 0 \\
&= 0.
\end{aligned}$$

■

Appendix D. Proof of Theorem 4

Proof Let S_{aug} be the set of source domain after the domain augmentation. According to Definition 6, we have $S \subset S_{\text{aug}}$, so:

$$\begin{aligned}
&\{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k; \quad \forall s, s' \in S_{\text{aug}}, \exists t \in T\} \\
&\subseteq \{k \in \mathcal{K} : D_s^k = D_{s'}^k \neq D_t^k; \quad \forall s, s' \in S, \exists t \in T\},
\end{aligned} \qquad (66)$$

so $\Phi_{\text{aug}} \subseteq \Phi$.

When $\Phi_{\text{aug}} \subset \Phi$, let $\Phi' = \Phi - \Phi_{\text{aug}}$. Then we have $\mathcal{Z}^{\Phi} = \mathcal{Z}^{\Phi'} \times \mathcal{Z}^{\Phi_{\text{aug}}}$.

Let ϕ', ϕ_{aug} and ϕ represents the variables in $\mathcal{Z}^{\Phi'}$, $\mathcal{Z}^{\Phi_{\text{aug}}}$, $\mathcal{Z}^{\Phi}$ and $\phi = (\phi_{\text{aug}}, \phi')$.

We express the hypothesis function $\hat{h}$ as a composition of an activation function $\sigma : (0,1) \to [0,1]$ and a fully-connected layer $\text{FC} : \mathcal{Z}^{\Phi} \to (0,1)$ and $\hat{h}(\cdot) = \sigma \circ \text{FC}(\cdot)$. Then we have:

$$\begin{aligned}
\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) &= \int_{\mathcal{Z}^{\Phi}} \sigma(\text{FC}(\phi)) L(\mathrm{d}\phi) \\
&= \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} \int_{\mathcal{Z}^{\Phi'}} \sigma(\text{FC}(\phi_{\text{aug}}, \phi')) L(\mathrm{d}\phi_{\text{aug}}) L(\mathrm{d}\phi')
\end{aligned} \qquad (67)$$

Due to the linearity property of the multilayer perceptron and the monotonically increasing property of the activation function, we have:

$$\begin{aligned}
\sigma(\text{FC}(\phi_{\text{aug}}, \phi')) &= \sigma(\text{FC}(\phi_{\text{aug}}, 0) + \text{FC}(0, \phi') + C) \\
&> \sigma(\text{FC}(\phi_{\text{aug}}, 0)) \\
&= \hat{h}(\phi_{\text{aug}})
\end{aligned} \qquad (68)$$

where the $C \geq 0$ is a constant. Then, we use the Fubini's theorem:

$$\begin{aligned}\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) &= \int_{\mathcal{Z}^{\Phi_{\text{aug}}} \times \mathcal{Z}^{\Phi'}} \sigma(\mathrm{FC}(\phi_{\text{aug}}, \phi')) L(\mathrm{d}\phi) \\ &= \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} \int_{\mathcal{Z}^{\Phi'}} \sigma(\mathrm{FC}(\phi_{\text{aug}}, \phi')) L(\mathrm{d}\phi_{\text{aug}}) L(\mathrm{d}\phi') \\ &> \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} \int_{\mathcal{Z}^{\Phi'}} \hat{h}(\phi_{\text{aug}}) L(\mathrm{d}\phi_{\text{aug}}) L(\mathrm{d}\phi') \\ &= \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} \hat{h}(\phi_{\text{aug}}) L(\mathrm{d}\phi_{\text{aug}})\end{aligned} \tag{69}$$

So, we have:

$$\sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) > \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} \hat{h}(\phi_{\text{aug}}) L(\mathrm{d}\phi_{\text{aug}}), \tag{70}$$

i.e. $R_{\Phi} > R_{\Phi_{\text{aug}}}$.

When $\Phi_{\text{aug}} = \Phi$, we have $\mathcal{Z}^{\Phi} = \mathcal{Z}^{\Phi_{\text{aug}}}$, and then:

$$\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) = \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} \hat{h}(\phi_{\text{aug}}) L(\mathrm{d}\phi_{\text{aug}}) \tag{71}$$

then we have that:

$$\sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) = \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\mathcal{Z}^{\Phi_{\text{aug}}}} |\hat{h}(\phi_{\text{aug}})| L(\mathrm{d}\phi_{\text{aug}}). \tag{72}$$

So, in this case, $R_{\Phi} = R_{\Phi_{\text{aug}}}$. ■

Appendix E. Proof of Theorem 5

First, we present the fundamental generalization bound with Rademacher complexity.

Theorem 10 (Generalization Bound with Rademacher Complexity) *For hypothesis class $\mathcal{H} : \mathcal{Z} \to [0, 1]$, n i.i.d. samples $\{z_{s,i}, i = 1, \cdots, n\}$ from the source domain s and any $\delta > 0$, with probability at least $1 - \delta$ we have:*

$$\sup_{h \in \mathcal{H}} \left| E_{D_s}[h(z)] - \frac{1}{n} \sum_{i=1}^{n} h(z_{s,i}) \right| \leq 2\mathcal{R}_s(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{73}$$

where $\mathcal{R}_s(\mathcal{H})$ is the Rademacher complexity on the domain s.
Next, we will prove Theorem 5.

Proof

$$
\begin{aligned}
&\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h)\Big| \\
&=\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)+\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)-\sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h)\Big| \\
&\le\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big|+\Big|\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)-\sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h)\Big| \\
&\le\sum_{s\in S}\pi_s|\varepsilon_{D_s}(h)-\hat{\varepsilon}_{D_s}(h)|+\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&\le\sup_{h\in\mathcal{H}}\sum_{s\in S}\pi_s\Big|E_{D_s}[\ell(h(z),f_s(z))]-\frac{1}{n}\sum_{i=1}^{n}\ell(h(z_{s,i}),f_s(z_{s,i}))\Big|+\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&\le\sup_{h\in\mathcal{H}}\sum_{s\in S}\pi_s\Big|\mathcal{L}E_{D_s}[|h(z)-f_s(z)|]-\frac{\mathcal{L}}{n}\sum_{i=1}^{n}|h(z_{s,i})-f_s(z_{s,i})|\Big|+\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&\le\mathcal{L}\Big|\sup_{h\in\mathcal{H}}\sum_{s\in S}\pi_s\Big(E_{D_s}[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_{s,i})\Big)+\sup_{h\in\mathcal{H}}\sum_{s\in S}\pi_s\Big(E_{D_s}[f_s(z)]-\frac{1}{n}\sum_{i=1}^{n}f_s(z_{s,i})\Big)\Big| \\
&\quad+\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big|,
\end{aligned}
\tag{74}
$$

where the last inequality follows from the absolute value inequality $\big||a|-|b|\big|\le|a+b|$. Because for any $s\in S$, the labeling function $f_s\in\mathcal{H}$, we have:

$$
\sup_{h\in\mathcal{H}}\sum_{s\in S}\pi_s\Big(E_{D_s}[f_s(z)]-\frac{1}{n}\sum_{i=1}^{n}f_s(z_{s,i})\Big)\le\sup_{h\in\mathcal{H}}\sum_{s\in S}\pi_s\Big(E_{D_s}[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_{s,i})\Big), \tag{75}
$$

Substituting this into $\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h)\Big|$, and according to Theorem 10, for any $\delta\in[0,\frac{1}{|S|}]$, we have:

$$
\begin{aligned}
&\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h)\Big| \\
&\le 2\mathcal{L}\sum_{s\in S}\pi_s\sup_{h\in\mathcal{H}}\big|E_{D_s}[h(z)]-\frac{1}{n}\sum_{i=1}^{n}h(z_{s,i})\big|+\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&\le 4\mathcal{L}\sum_{s\in S}\pi_s\mathcal{R}_s(\mathcal{H})+6\mathcal{L}\sqrt{\frac{\ln(2/\delta)}{2n}}+\Big|\varepsilon_{D_t}(h)-\sum_{s\in S}\pi_s\varepsilon_{D_s}(h)\Big| \\
&\le\mathcal{L}\Big(4\mathcal{R}_S^w(\mathcal{H})+6\sqrt{\frac{\ln(2/\delta)}{2n}}+\frac{1}{3}\min_{D\in U}d_{\hat{\mathcal{H}}}^{\Xi}(D_t^{\Xi},D)+\frac{1}{3}\max_{s,s'}d_{\hat{\mathcal{H}}}^{\Xi}(D_s^{\Xi},D_{s'}^{\Xi}) \\
&\quad+\frac{1}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^\Phi}\hat{h}(\phi)L(\mathrm{d}\phi)+\min\Big\{\sum_s\pi_sE_{D_t}[|f_s-f_t|],\sum_s\pi_sE_{D_s}[|f_s-f_t|]\Big\}\Big).
\end{aligned}
\tag{76}
$$

So, we have:

$$\varepsilon_{D_t}(h) \leq \sum_{s\in S}\pi_s\hat{\varepsilon}_{D_s}(h) + \mathcal{L}\Big(4\mathcal{R}_S^w(\mathcal{H}) + 6\sqrt{\frac{\ln(2/\delta)}{2n}} + \frac{1}{3}\min_{D\in U} d_{\hat{\mathcal{H}}}^{\Xi}(D_t^{\Xi}, D) + \frac{1}{3}\max_{s,s'} d_{\hat{\mathcal{H}}}^{\Xi}(D_s^{\Xi}, D_{s'}^{\Xi}) + \frac{1}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^{\Phi}}\hat{h}(\phi)L(\mathrm{d}\phi) + \min\big\{\sum_s \pi_s E_{D_t}[|f_s - f_t|], \sum_s \pi_s E_{D_s}[|f_s - f_t|]\big\}\Big). \tag{77}$$

This inequality holds with probability $1 - |S|\delta$. ■

Appendix F. Proof of Theorem 7

Proof Applying Theorem 6 to the scenario with multiple source domains, and then there exists a set of source domain distributions such that:

$$P_{\hat{S}\sim(D_a)^{|S|n}}\Big(\varepsilon_{D_a}(A(\hat{S})) - \inf_{h\in\mathcal{H}}\varepsilon_{D_a}(h) > \sqrt{\frac{d_{VC}}{320|S|n}}\Big) > \frac{1}{64}, \tag{78}$$

where D_a is the average distribution of source domains, i.e. $D_a = \frac{1}{|S|}\sum_{s\in S} D_s$, $\hat{S}$ is the set of $|S|n$ training samples draw from the source domains.

Due to the Assumption 3, we have:

$$\begin{aligned}
\varepsilon_{D_t}(A(\hat{S})) &= \varepsilon_{D_a}(A(\hat{S})) + \int_{\mathcal{Z}} \ell\big(A(\hat{S})(z), f(z)\big)\big(D_t(z) - D_a(z)\big)L(\mathrm{d}z) \\
&= \varepsilon_{D_a}(A(\hat{S})) + \int_{\mathcal{Z}} \ell\big(A(\hat{S})(\gamma) + A(\hat{S})(\phi) + A(\hat{S})(\xi), f(z)\big)\big(D_t(z) - D_a(z)\big)L(\mathrm{d}z) \\
&= \varepsilon_{D_a}(A(\hat{S})) + \int_{\mathcal{Z}} \mathsf{L}\big(A(\hat{S})(\gamma) + A(\hat{S})(\phi) + A(\hat{S})(\xi) - f(z)\big)\big(D_t(z) - D_a(z)\big)L(\mathrm{d}z) \\
&\geq \varepsilon_{D_a}(A(\hat{S})) + \int_{\mathcal{Z}} \Big(\mathcal{L}_0'\big(A(\hat{S})(\gamma) - f(\gamma)\big) + \mathcal{L}_0' A(\hat{S})(\phi) + \mathcal{L}_0' A(\hat{S})(\xi)\Big)\big(D_t(z) - D_a(z)\big)L(\mathrm{d}z).
\end{aligned} \tag{79}$$

The second equality holds due to the Assumption 5. The third equality holds due to the Assumption 4, and the "$\geq$" holds due to the second-order Taylor expansion of the function L about zero. According the distribution property of the features γ, ϕ and ξ we defined in Section 4.1, we have:

$$\begin{aligned}
&\int_{\mathcal{Z}} \Big(\mathcal{L}_0'\big(A(\hat{S})(\gamma) - f(\gamma)\big) + \mathcal{L}_0' A(\hat{S})(\phi) + \mathcal{L}_0' A(\hat{S})(\xi)\Big)\big(D_t(z) - D_a(z)\big)L(\mathrm{d}z) \\
&= \int_{\mathcal{Z}^{\Gamma}}\int_{\mathcal{Z}^{\Phi}}\int_{\mathcal{Z}^{\Xi}} \Big(\mathcal{L}_0'\big(A(\hat{S})(\gamma) - f(\gamma)\big) + \mathcal{L}_0' A(\hat{S})(\phi) + \mathcal{L}_0' A(\hat{S})(\xi)\Big)\big(D_t(z) - D_a(z)\big)L(\mathrm{d}\gamma)L(\mathrm{d}\phi)L(\mathrm{d}\xi) \\
&= \mathcal{L}_0'\int_{\mathcal{Z}^{\Gamma}} \big(A(\hat{S})(\gamma) - f(\gamma)\big)\big(D_t^{\Gamma}(\gamma) - D_a^{\Gamma}(\gamma)\big)L(\mathrm{d}\gamma) + \mathcal{L}_0'\int_{\mathcal{Z}^{\Phi}} A(\hat{S})(\phi)\big(D_t^{\Phi}(\phi) - D_a^{\Phi}(\phi)\big)L(\mathrm{d}\phi) \\
&\quad + \mathcal{L}_0'\int_{\mathcal{Z}^{\Xi}} A(\hat{S})(\xi)\big(D_t^{\Xi}(\xi) - D_a^{\Xi}(\xi)\big)L(\mathrm{d}\xi)
\end{aligned} \tag{80}$$

Because the distribution property of feature ξ, we have: $D_t^\Gamma(\gamma) - D_a^\Gamma(\gamma) = 0$. For the second integral, since $D_s^\Phi = D_{s'}^\Phi = D^\Phi$ for any $s, s' \in S$, we have $D_t^\Phi(\phi) - D_a^\Phi(\phi) = D_t^\Phi(\phi) - D^\Phi(\phi)$, we apply the Mean Value Theorems:

$$\begin{aligned}\int_{\mathcal{Z}^\Phi} A(\hat{S})(\phi)\big(D_t^\Phi(\phi) - D_a^\Phi(\phi)\big)L(\mathrm{d}\phi) &= \int_{\mathcal{Z}^\Phi} A(\hat{S})(\phi)\big(D_t^\Phi(\phi) - D^\Phi(\phi)\big)L(\mathrm{d}\phi) \\ &= \big(D_t^\Phi(\phi'') - D_a^\Phi(\phi'')\big)\int_{\mathcal{Z}^\Phi} A(\hat{S})(\phi)L(\mathrm{d}\phi) \\ &= \bar{B}\int_{\mathcal{Z}^\Phi} A(\hat{S})(\phi)L(\mathrm{d}\phi) \\ &= \bar{B}R_\Phi(A(\hat{S})),\end{aligned} \tag{81}$$

where the $\bar{B} = \big(D_t^\Phi(\phi'') - D_a^\Phi(\phi'')\big)$ is a constant,and $\phi'' \in \Phi$ is a number generated by the Mean Value Theorems for Definite Integrals. The final equality is due to Definition 11. Finally, because $D_a = \frac{1}{|S|}\sum_{s\in S} D_s$, we have:

$$\begin{aligned}\int_{\mathcal{Z}^\Xi} A(\hat{S})(\xi)\big(D_t^\Xi(\xi) - D_a^\Xi(\xi)\big)L(\mathrm{d}\xi) &= \int_{\mathcal{Z}^\Xi} A(\hat{S})(\xi)\big(D_t^\Xi(\xi) - \frac{1}{|S|}\sum_{s\in S} D_s^\Xi(\xi)\big)L(\mathrm{d}\xi) \\ &= \frac{1}{|S|}\sum_{s\in S}\int_{\mathcal{Z}^\Xi} A(\hat{S})(\xi)\big(D_t^\Xi(\xi) - D_s^\Xi(\xi)\big)L(\mathrm{d}\xi) \\ &= \frac{1}{|S|}\sum_{s\in S}\Delta^\Xi(D_t^\Xi, D_s^\Xi),\end{aligned} \tag{82}$$

the final equality is due to Definition 11.
So, we have:

$$\varepsilon_{D_t}(A(\hat{S})) - \varepsilon_{D_a}(A(\hat{S})) \geq \mathcal{L}_0'\bar{B}R_\Phi(A(\hat{S})) + \mathcal{L}_0'\frac{1}{|S|}\sum_{s\in S}\Delta^\Xi(D_t^\Xi, D_s^\Xi). \tag{83}$$

Because a model trained on the samples from source domains naturally outperforms on the source domain distribution compared to its performance on the target domain, we have:

$$\varepsilon_{D_t}(A(\hat{S})) - \varepsilon_{D_a}(A(\hat{S})) \geq 0, \tag{84}$$

then:

$$\varepsilon_{D_t}(A(\hat{S})) - \varepsilon_{D_a}(A(\hat{S})) \geq \max\left\{0, \quad \mathcal{L}_0'\bar{B}R_\Phi(A(\hat{S})) + \mathcal{L}_0'\frac{1}{|S|}\sum_{s\in S}\Delta^\Xi(D_t^\Xi, D_s^\Xi)\right\}. \tag{85}$$

Finally, we have:

$$\begin{aligned}\varepsilon_{D_t}(A(\hat{S})) &\geq \varepsilon_{D_a}(A(\hat{S})) + \max\left\{0, \quad \mathcal{L}_0'\bar{B}R_\Phi(A(\hat{S})) + \frac{\mathcal{L}_0'}{|S|}\sum_{s\in S}\Delta^\Xi(D_t^\Xi, D_s^\Xi)\right\} \\ &> \inf_{h\in\mathcal{H}}\varepsilon_{D_a}(h) + \sqrt{\frac{d_{VC}}{320|S|n}} + \max\left\{0, \quad \mathcal{L}_0'\bar{B}R_\Phi(A(\hat{S})) + \frac{\mathcal{L}_0'}{|S|}\sum_{s\in S}\Delta^\Xi(D_t^\Xi, D_s^\Xi)\right\},\end{aligned} \tag{86}$$

■

Appendix G. Proof of Theorem 9

First, we generalize this framework to the setting with multi-domain samples.

Theorem 11 (Generalization Bound for Uniformly Stable Algorithms) *Let A be a learning algorithm with β-uniform stability with respect to a loss function ℓ that satisfies $0 \leq \ell(h, z) \leq M$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. Let $\hat{S}_s = (z_{s,1}, \ldots, z_{s,n})$ be a training set drawn i.i.d. from the domain s, $s \in S$, $\hat{S} = \cup_{s \in S} \hat{S}_s$ and let $A(\hat{S})$ be the hypothesis output by the algorithm. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the random draw of S, the following bound holds:*

$$\left|\varepsilon_{D_a}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right| \leq \beta + (2|S|n\beta + M)\sqrt{\frac{\ln(1/\delta)}{2|S|n}}, \tag{87}$$

where S is the set of domains from which samples are drawn, $|S|$ is the number of elements in S, and $D_a = \frac{1}{|S|}\sum_{s \in S} D_s$ is the average distribution of source domain.

This theorem is a direct consequence of extending the stability-based generalization bound to the multi-domain training sample setting. Then we give the proof of the Theorem 9.

Proof

$$\begin{aligned}\left|\varepsilon_{D_t}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right| &= \left|\varepsilon_{D_t}(A(\hat{S})) - \varepsilon_{D_t}(A(\hat{S})) + \varepsilon_{D_t}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right| \\ &\leq \left|\varepsilon_{D_t}(A(\hat{S})) - \varepsilon_{D_t}(A(\hat{S}))\right| + \left|\varepsilon_{D_t}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right|.\end{aligned} \tag{88}$$

According to Theorem 2, we have:

$$\begin{aligned}\varepsilon_{D_t}(h) \leq \sum_{s \in S} \pi_s \varepsilon_{D_s}(h) + \mathcal{L}\Big(&\frac{1}{3}\min_{D \in U} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t, D) + \frac{1}{3}\max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s, D^{\Xi}_{s'}) \\ &+ \frac{1}{3}B \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\hat{\mathcal{Z}}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) + \min\Big\{\sum_s \pi_s E_{D_t}[|f_s - f_t|], \sum_s \pi_s E_{D_s}[|f_s - f_t|]\Big\}\Big).\end{aligned} \tag{89}$$

Let $h = A(\hat{S})$ and $\pi_s = \frac{1}{|S|}, s \in S$, we have

$$\begin{aligned}\left|\varepsilon_{D_t}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right| \leq \mathcal{L}\Big(&\frac{1}{3}\min_{D \in U} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t, D) + \frac{1}{3}\max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s, D^{\Xi}_{s'}) \\ &+ \frac{1}{3}B \sup_{\hat{h} \in \hat{\mathcal{H}}} \int_{\hat{\mathcal{Z}}^{\Phi}} \hat{h}(\phi) L(\mathrm{d}\phi) \\ &+ \min\Big\{\sum_s \pi_s E_{D_t}[|f_s - f_t|], \sum_s \pi_s E_{D_s}[|f_s - f_t|]\Big\}\Big).\end{aligned} \tag{90}$$

Due to the covariate shift assumption 3, we have

$$\min\Big\{\sum_s \pi_s E_{D_t}[|f_s - f_t|], \sum_s \pi_s E_{D_s}[|f_s - f_t|]\Big\} = 0 \tag{91}$$

Due to Theorem 11, we obtain the final result:

$$\begin{aligned}\left|\varepsilon_{D_t}(A(\hat{S})) - \hat{\varepsilon}_{D_a}(A(\hat{S}))\right| \leq \beta + (2|S|n\beta + M)\sqrt{\frac{\ln(1/\delta)}{2|S|n}} + \mathcal{L}\Big(\frac{1}{3}\min_{D\in U} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_t, D) \\ + \frac{1}{3}\max_{s,s'} d^{\Xi}_{\hat{\mathcal{H}}}(D^{\Xi}_s, D^{\Xi}_{s'}) + \frac{1}{3}B\sup_{\hat{h}\in\hat{\mathcal{H}}}\int_{\mathcal{Z}^{\Phi}} \hat{h}(\phi)L(\mathrm{d}\phi)\Big).\end{aligned} \tag{92}$$

■

Appendix H. A Synthetic Illustration of the Tri-Space Decomposition

In Section 4.1, we introduced the Tri-Space latent representation through Definitions 1, 2, and 3, and established its unique direct-sum decomposition in Theorem 1. Although mathematically precise, the abstract formulation may obscure the concrete meaning of the subspaces $\mathcal{Z}^{\Gamma}$, $\mathcal{Z}^{\Phi}$, and $\mathcal{Z}^{\Xi}$. To facilitate an intuitive understanding of the three types of features, we provide a fully specified synthetic example with explicit numerical parameters and analyze the behavior of a linear model trained on this data. This example directly instantiates the three feature types depicted in Figure 1 of the main manuscript, illustrates their distinct roles in generalization, and demonstrates the impact of spurious invariant features on domain generalization performance.

Data Generating Process

Consider a binary classification task where the label $Y \in \{0, 1\}$ indicates whether an image is a circle ($Y = 1$) or a square ($Y = 0$). Each observation is generated from three independent latent factors, which correspond exactly to the three subspaces of our framework. Let $\gamma \in \mathcal{Z}^{\Gamma}$ denote the shape attribute, $\phi \in \mathcal{Z}^{\Phi}$ the color attribute, and $\xi \in \mathcal{Z}^{\Xi}$ the background texture attribute. We assume the following generative model.

The shape γ is drawn independently of the domain from a Bernoulli distribution:

$$\gamma \sim \text{Bernoulli}(0.5). \tag{93}$$

We adopt the convention that squares correspond to $\gamma = 0$ and circles to $\gamma = 1$. The label is deterministically assigned as $Y = \gamma$. Thus, γ fully determines the classification target, satisfying Assumption 2 of the main text.

The color ϕ is a binary variable whose conditional distribution given γ depends on the domain. We adopt the convention that red corresponds to $\phi = 0$ and blue to $\phi = 1$. In all source domains $s \in S = \{s_1, s_2\}$, the color is perfectly correlated with the shape:

$$\phi \mid \gamma = \gamma \quad \text{(i.e., red squares and blue circles: when } \gamma = 0, \phi = 0\text{; when } \gamma = 1, \phi = 1\text{).} \tag{94}$$

In the target domain t, the correlation is reversed:

$$\phi \mid \gamma = 1 - \gamma \quad \text{(i.e., blue squares and red circles: when } \gamma = 0, \phi = 1\text{; when } \gamma = 1, \phi = 0\text{).} \tag{95}$$

The background texture ξ is a continuous scalar drawn from a Gaussian distribution whose mean and variance depend on the domain:

$$\xi \sim \mathcal{N}(\mu(d), \sigma^2(d)), \tag{96}$$

where $d \in S \cup \{t\}$ indexes the domain. Specifically, we set:

$$\begin{aligned} \mu(s_1) &= 0.0, & \sigma^2(s_1) &= 0.1; \\ \mu(s_2) &= 1.0, & \sigma^2(s_2) &= 0.5; \\ \mu(t) &= 2.0, & \sigma^2(t) &= 1.0. \end{aligned}$$

Then we verify that this generative process satisfies Definitions 1, 2, 3 of the main text.

- The marginal distribution of γ is Bernoulli(0.5) in every domain, i.e., $D_s^\gamma = D_t^\gamma$ for all $s \in S$ and $t \in T$. By construction, the label Y is a deterministic function of γ alone.
- The conditional distribution of ϕ given γ is identical across all source domains: $P(\phi = \gamma \mid \gamma) = 1$ for every $s \in S$. Hence, $D_s^\phi = D_{s'}^\phi$ for all $s, s' \in S$. In the target domain, however, the conditional distribution flips, so $D_t^\phi \neq D_s^\phi$. Color thus exemplifies a spurious correlation that appears invariant during training but fails to generalize.
- The distribution of ξ differs across source domains ($D_{s_1}^\xi \neq D_{s_2}^\xi$) because $\mu(s_1) \neq \mu(s_2)$ and $\sigma^2(s_1) \neq \sigma^2(s_2)$, and it shifts further in the target domain. Background texture is therefore a purely domain-variant attribute.

Learning Failure under Spurious Invariant Feature

We now employ this synthetic example to analyze, under the lens of domain-invariant representation learning, the detrimental impact of spurious invariant features on domain generalization. Domain-invariant representation learning aims to eliminate features whose distributions are inconsistent across source domains. In this example, the distribution of ξ exhibits different means and variances between source domains s_1 and s_2; consequently, an effective domain-invariant learning method will identify ξ as a domain-variant feature and remove it from the learned representation. In other words, after domain-invariant representation learning, the model relies exclusively on the remaining features γ and ϕ. Accordingly, we consider a linear model that takes features γ and ϕ as input and outputs a prediction $\hat{Y}$ for the label Y:

$$\hat{Y} = w_1\gamma + w_2\phi + b, \tag{97}$$

where w_1, w_2, and b are learnable parameters. In the source domains, since $\phi = \gamma$ holds deterministically, both γ and ϕ are perfectly linearly correlated with the label Y. Specifically, for any sample we have $\gamma = \phi$, and thus the true label $Y = \gamma$ can be equivalently written as $Y = \phi$. Any linear model satisfying $w_1 + w_2 = 1$ and $b = 0$ achieves zero training loss on the source domains. Consequently, the model may converge to various different solutions. In particular, consider the following two representative solutions:

- **Ideal model** (relying solely on γ): $w_1 = 1$, $w_2 = 0$, $b = 0$. The prediction function is $\hat{Y} = \gamma$, which coincides with the true model.

- **Spurious model** (relying solely on ϕ): $w_1 = 0$, $w_2 = 1$, $b = 0$. The prediction function is $\hat{Y} = \phi$. In the source domains, since $\phi = \gamma$, this model also fits the training data perfectly.

When evaluated on the target domain t, where $\phi = 1 - \gamma$, the two models exhibit drastically different generalization performance. For the ideal model, $\hat{Y} = \gamma$, so the prediction error satisfies $|\hat{Y} - Y| = |\gamma - \gamma| = 0$, and the accuracy remains 100%. For the spurious model, $\hat{Y} = \phi = 1 - \gamma$, whereas the true label is $Y = \gamma$. Hence, the prediction error is $|\hat{Y} - Y| = |(1 - \gamma) - \gamma| = |1 - 2\gamma|$. When $\gamma = 0$, the error is 1; when $\gamma = 1$, the error is also 1. If we adopt a threshold of 0.5 for binary classification, the spurious model makes completely incorrect predictions on the target domain, yielding an accuracy of 0%.

Connection to the Fine-Grained Risk Bound

This toy scenario illustrates the precise mechanism that our fine-grained risk bound (Theorem 2) is designed to capture. Domain-invariant representation learning successfully eliminates the domain-variant feature ξ. However, because the spurious invariant feature ϕ also exhibits invariance across source domains, it is erroneously retained and may induce the model to learn a spurious solution that depends on ϕ. The Tri-Space decomposition disentangles the sources of generalization error: the risk induced by the spurious invariant feature ϕ is encapsulated by the domain diversity factor R_Φ (Definition 9), while the risk stemming from the domain-variant feature ξ is reflected in the domain invariance factor R_Ξ (Definition 8). This example clearly demonstrates that even when ξ is removed, if the source domains lack sufficient diversity to expose the spurious invariance of ϕ, the model may still fail to generalize on the target domain. The example thus provides an intuitive and mathematically grounded foundation for the theoretical analysis developed in the main body of the paper.

Appendix I. Approximation Estimation of Constant B

In Theorem 2 of Section 4.2, we derive a fine-grained risk upper bound for arbitrary target domains, which involves a constant B obtained via the Mean Value Theorem for Definite Integrals. This constant, defined as $B = \sup_{\phi \in \mathcal{Z}^\Phi} |D_t^\Phi(\phi) - D^\Phi(\phi)|$, quantifies the maximum pointwise discrepancy between the probability density functions of the spurious invariant feature ϕ under the target domain distribution and the mixture distribution of the source domains. By definition, the distribution of feature ϕ inherently differs between the source and target domains. Consequently, the target domain distribution necessarily deviates from the mixed source domain distribution, implying that B is strictly bounded between 0 and 1.

In practice, we propose an approximation estimation method for this constant, this method is predicated on the perfect separation of subspace $\mathcal{Z}^\Phi$, that is, obtaining the disentangled representation of feature ϕ. Specifically, we partition $\mathcal{Z}^\Phi$ into M disjoint regions, denoted by $i \in [1, M]$, and estimate B as follows:

$$\max_{i \in [1,M]} \frac{n_S(i)/m - n_T(i)}{n}, \tag{98}$$

where n denotes the total number of samples in a domain, m is the number of source domains, $n_S(i)$ represents the total number of source domain samples falling within region i, and $n_T(i)$ denotes the total number of target domain samples in region i.

Appendix J. Domain Generalization or Multi-Source Domain Adaptation? A Discussion on Terminology Choice

This section addresses a potentially confusing terminological choice: whether the bound we establish pertains to domain generalization or multi-source domain adaptation.

Our theoretical framework, which bounds target risk using multiple source domains, may resemble multi-source domain adaptation (MSDA). However, it fundamentally aligns with the domain generalization (DG) paradigm for key reasons.

1. In MSDA, the target domain is a specific, known entity (even if unlabeled) that the model must adapt to. This is a "many-to-one" problem: multiple source domains are leveraged to improve performance on that single, fixed target distribution. In contrast, our work follows the domain generalization (DG) paradigm, which is inherently a "many-to-any" problem. Here, the target domain D_t represents any arbitrary domain from a potential set T of unseen domains. No information about any specific target domain is available during training. The goal is to learn a model from multiple source domains that performs well on any possible target domain from T, not a pre-specified one. Therefore, the theoretical bound we derive must hold uniformly for all $t \in T$.

2. Domain generalization theory has developed along two main analytical paths, as comprehensively surveyed in recent overviews (e.g., Wang et al. (2022)):

(1) **The DG Error based on Meta-Distribution:** This approach, introduced by Muandet et al. (2013), defines generalization error with respect to a meta-distribution $\mathcal{P}$ over domains. The goal is to bound the deviation between the expected risk over $\mathcal{P}$ and the empirical risk on the source domains.

(2) **The Arbitrary-Target Domain Risk Analysis Framework:** This approach, exemplified by Albuquerque et al. (2019) and Sicilia et al. (2023), does not assume a meta-distribution. Instead, it aims to provide an upper bound for the risk $\varepsilon_{D_t}(h)$ on any possible target domain within a set T, offering guarantees for arbitrary unseen domains.

Our work explicitly follows and contributes to the second path—the arbitrary-target domain risk analysis framework. This firmly establishes our work within the established theoretical landscape of domain generalization, not multi-source domain adaptation.

3. We adopt the arbitrary-target domain risk analysis framework for its mathematical tractability and strong algorithmic interpretability. First, it avoids the fundamental measure-theoretic challenges associated with rigorously defining a probability measure (meta-distribution) on the abstract, infinite-dimensional space of all domain distributions—a significant open problem noted in Section 3.1.2. Second, and more importantly, the risk bounds derived under this framework decompose into terms with direct, intuitive connections to algorithmic operations. For instance, terms involving maximum divergence among source domains relate to domain-invariant learning, while terms involving the distance from the target to a convex hull of sources relate to the benefits of domain diversity. This clear interpretability provides actionable insights for algorithm design and analysis, which is a central goal of our work.

4. We emphasize that the Tri-Space latent representation and the subsequent fine-grained risk bound is fundamentally independent of the choice between these two DG analysis frameworks. The Tri-Space decomposition is a structural characterization of the latent representation, and the derived bound reveals how different feature types affect generalization. In principle, these insights could also be integrated into a meta-distribution-based analysis, provided a well-defined meta-distribution exists. We chose the "arbitrary-target risk" framework for its mathematical tractability and its strong algorithmic interpretability.

In summary, while our theoretical setup shares superficial similarities with MSDA, it is conceptually and formally aligned with the domain generalization paradigm, specifically the arbitrary-target domain risk analysis strand. Our core contributions are agnostic to this framework choice and offer novel insights into the mechanics of generalization that we believe are valuable to the broader DG community.