COLLABORATIVE DUAL-EVIDENCE REPRESENTATION LEARNING FOR DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain Generalization (DG) seeks to induce a model from several source domains that generalizes to unseen target domains. Existing alignment-based DG methods strive to promote statistical invariance by reducing representation discrepancies across source domains. However, such alignment alone may not be sufficient for reliable generalization, since different representations may achieve similar levels of source-domain alignment while preserving different discriminative structures, and consequently behave differently on unseen domains. This motivates us to complement the statistical alignment with an additional inductive bias that provides a distinct discriminative signal for more reliable prediction. To this end, we propose **Co**llaborative **D**ual-**E**vidence **R**epresentation Learning (CoDER), which implements the two inductive biases through collaborative pathways and integrates dual predictive evidence for final prediction. Specifically, one pathway derives evidence from a CORAL-regularized representation, while the other utilizes a shared latent dictionary to generate context-aware modulation signals for this representation, yielding adaptive discriminative evidence. Each pathway employs a different classifier to provide Dirichlet evidence, which is fused at inference. Extensive experiments on five benchmarks demonstrate that CoDER consistently outperforms strong competitors, with evidence fusion *effectively enhancing* predictive performance and reliability.

1 INTRODUCTION

In real-world applications, models often encounter conditions that differ substantially from those observed during training, such as changes in weather, sensors, backgrounds, or populations Sanchez et al. (2023). These distribution shifts can severely degrade predictive performance Eastwood et al. (2022). Domain Generalization (DG) Zhang et al. (2021) seeks to induce a model from several source domains, enabling it to generalize to unseen target domains under these shifts.

To bridge the generalization gap, a widely adopted strategy is to suppress domain-specific variations, allowing a common predictor to operate across domains. Following this intuition, many existing DG methods seek invariance by reducing representation discrepancies across source domains, i.e., domain-invariant representation learning Ganin et al. (2016). However, recent studies reveal that distribution alignment does not necessarily eliminate domain dependence. For example, Zhu et al. (2022) demonstrated that adversarial alignment may reduce global discrepancies while leaving local structures insufficiently aligned. Furthermore, Cao & Chen (2024) discovered that samples in the learned representation space often still form domain-dependent clusters, indicating that substantial domain information remains encoded after alignment.

We view these phenomena as a direct consequence of existing alignment strategies primarily enforcing invariance at the *statistical* level. In other words, aligning source domains merely dictates how representations are distributed globally, but does not sufficiently constrain which predictive information should be retained, leaving room for domain-related information to remain encoded. As illustrated in Fig. 1, representations based on shape, color, or both may all satisfy the source-domain alignment objective, yet exhibit markedly different generalization performance on the target domain. Admittedly, standard task supervision objectives, such as cross-entropy loss Vapnik et al. (1998), can be used to rule out representations uninformative for source-domain prediction, as exemplified by the *Only Color* candidate in Fig. 1. However, multiple aligned and source-predictive representations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

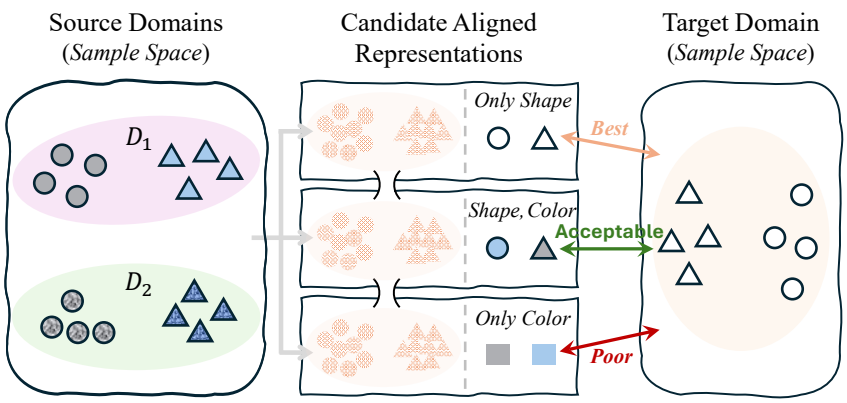

Figure 1: Illustration of a potential limitation of statistical alignment, where three candidate representations exhibit comparable alignment but preserve different information, resulting in different generalization on the target domain.

may remain, including those that retain both stable structures and domain-related information, such as the *Shape, Color*. Causal learning methods Arjovsky et al. (2019) further seek predictive mechanisms that remain stable across environments by modeling causal structures Salaudeen & Koyejo (2024). Nevertheless, the range of variations observed across source domains may be insufficient to ensure that the information identified as invariant remains reliable in unseen target domains.

Rather than undermining the value of statistically invariant representations, these observations suggest that relying on a single aligned representation may be insufficient across diverse unseen target domains, as the discriminative structures it preserves may vary in relevance across domain variations. We therefore retain statistical alignment as a relatively stable predictive basis while complementing it with an additional inductive bias that provides a distinct discriminative signal. The underlying intuition is simply that the same sample can be characterized from different perspectives, each highlighting distinct yet task-relevant cues that may provide complementary support for prediction. By jointly exploiting these perspectives, the model can reduce its reliance on any single representation and achieve more reliable predictions when the relevance of particular discriminative cues changes across domains.

To this end, we propose a **Co**llaborative **D**ual-**E**vidence **R**epresentation Learning framework, termed CoDER. This framework instantiates two distinct inductive biases through coordinated representation pathways, enabling them to capture different task-relevant cues and produce predictive evidence for subsequent integration. Specifically, CoDER employs a statistical-invariance pathway and a context-aware pathway built upon a shared encoder. The former maps the shared representation into a pathway-specific representation space through a projection head and applies CORAL regularization Sun & Saenko (2016) to learn a statistically aligned representation. The latter leverages the shared representation to query a shared latent dictionary and forms a sample-conditioned context from the latent bases. This context serves as a conditioning signal for generating adaptive scale-and-shift parameters Perez et al. (2018), which modulate the aligned representation and yield a distinct context-conditioned discriminative signal. Each pathway is equipped with an independent classifier that outputs class-specific nonnegative evidence to parameterize a Dirichlet distribution Sensoy et al. (2018). The resulting dual predictive evidence is subsequently integrated to produce the final prediction. In summary, our primary contributions can be briefly outlined as follows:

1. A **Co**llaborative **D**ual-**E**vidence **R**epresentation Learning framework (CoDER) is proposed to enhance prediction reliability across unseen domains.
2. CoDER introduces a context-aware pathway for sample-conditioned modulation, complementing the aligned representation with a distinct discriminative perspective.
3. Extensive experiments on five benchmarks demonstrate that CoDER consistently outperforms strong competitors.
4. Further analysis indicates that prediction errors are concentrated at smaller modulation magnitudes, supporting the role of context-aware modulation.

2 Related Work

In this section, we review existing DG methods in two categories, that is, invariant representation learning, and data diversity and robust optimization.

2.1 Invariant Representation Learning for DG

Learning representations that remain stable across domains is a central direction in DG. Early methods mainly reduce distribution discrepancies among source domains. MMD Li et al. (2018b) matches distributions through kernel mean embeddings, while CORAL Sun & Saenko (2016) aligns their second-order statistics. DANN Ganin et al. (2016) instead uses an adversarial domain classifier to suppress domain-identifiable information in the learned representation. Subsequent studies extend alignment to other levels. For example, RDM Nguyen et al. (2024) aligns domain-wise risk distributions rather than high-dimensional features. Other methods promote cross-domain consistency through gradient matching Rame et al. (2022); Wang et al. (2023b) or worst-domain optimization Sagawa et al. (2020). To better preserve class discrimination, several approaches further align class-conditional or joint distributions Chen et al. (2023); Zhou et al. (2025).

Another line of work studies invariance from a causal perspective. IRM Arjovsky et al. (2019) seeks a representation on which the same predictor is simultaneously optimal across training environments. Related methods exploit causal structures or conditional independence assumptions to separate stable task factors from environment-dependent variations Li et al. (2022); Salaudeen & Koyejo (2024). Compared with marginal distribution matching, causal invariance provides a stronger view of cross-environment stability. In practice, however, the underlying causal variables and intervention mechanisms are usually unavailable, and invariance must still be estimated from a finite set of observed source domains.

Our CoDER does not replace statistical alignment or claim to identify causal mechanisms. Instead, it addresses the limited guidance that a single alignment pathway provides for retaining predictive information. By introducing a context-aware modulation pathway over a shared encoder, it adds an inductive bias and integrates the resulting dual predictive evidence for more reliable prediction.

2.2 Data Diversity and Robust Optimization for DG

Beyond invariant representation learning, many DG methods improve generalization by introducing diversity into data, or optimization. Mixup Yan et al. (2020) interpolates training samples to encourage smoother decision functions, while MixStyle Zhou et al. (2024) mixes feature statistics to synthesize style variations. Re2SL Cao & Chen (2025) preserves domain diversity through relative representations, thereby improving generalization. MLDG Li et al. (2018a) constructs meta-train and meta-test splits from source domains to simulate domain shifts during training. Optimization-based methods seek solutions that are less sensitive to domain variation. Fish Shi et al. (2022) and PGrad Wang et al. (2023b) encourage agreement among domain-wise optimization directions, whereas SAGM Wang et al. (2023a) improves robustness by optimizing model performance in the neighborhood of perturbed parameters. DA-SAM He et al. (2026) accounts for domain heterogeneity by adaptively adjusting both the direction and magnitude of sharpness-aware perturbations.

Unlike these data diversity and robust optimization methods, our CoDER focuses on the predictive ambiguity left by a single aligned representation. It complements statistical alignment with a sample-conditioned representation pathway and integrates their predictive evidence, thereby reducing reliance on any single discriminative perspective.

3 Methodology

In this section, we present the CoDER framework together with its inference procedure in detail.

3.1 Preliminaries

In DG, each domain can be characterized by a joint distribution $P(\boldsymbol{X}, \boldsymbol{Y})$ over the input space $\mathcal{X}$ and corresponding label space $\mathcal{Y}$ Cao & Chen (2024). Domain shift arises when these joint

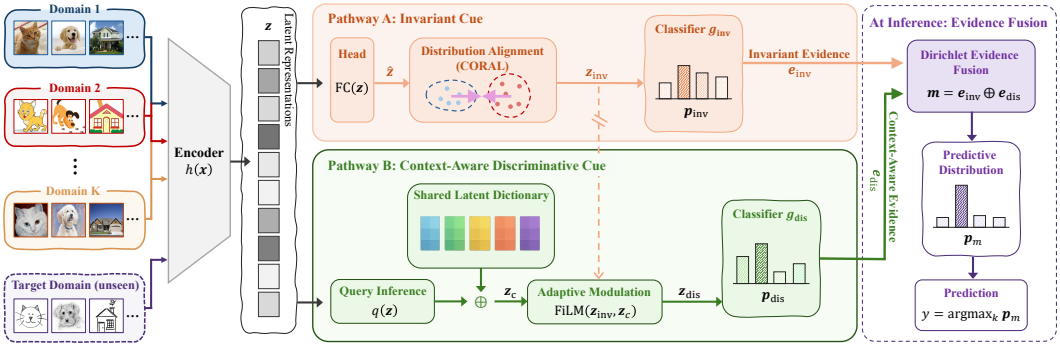

Figure 2: A pipeline of the proposed **Co**llaborative **D**ual-**E**vidence **R**epresentation Learning (CoDER), which consists of two pathways built upon a shared representation $\boldsymbol{z}$. The arrow $\dashrightarrow$ denotes that no gradient is propagated backward, while $\oplus$ denotes a combination operation over its two inputs. During inference, the evidence produced by g_{inv} and g_{dis} is fused for final prediction.

distributions differ across domains, i.e., $P^d(\boldsymbol{X}, \boldsymbol{Y}) \neq P^{d'}(\boldsymbol{X}, \boldsymbol{Y}), \forall d \neq d'$. Given N source domains, the dataset from the d-th domain can be denoted by $\mathcal{D}_{\mathcal{S}}^d = \left\{\left(\boldsymbol{x}_i^d, y_i^d\right)\right\}_{i=1}^{N_d}$, where N_d is the number of sample pairs drawn from $P^d(\boldsymbol{X}, \boldsymbol{Y})$. Meanwhile, the target domain differs in distribution from the source domains and remains unseen during training. Following the standard DG setting, all domains share a common label space $\mathcal{Y} = \{1, \ldots, K\}$.

Given an input sample $\boldsymbol{x}$, a shared feature extractor $h : \mathcal{X} \to \mathbb{R}^M$ maps it to a latent representation $\boldsymbol{z} = h(\boldsymbol{x})$, where M denotes the dimension of the latent representation.

3.2 CoDER

As illustrated in Fig. 2, CoDER builds two collaborative pathways upon a shared representation $\boldsymbol{z}$ to capture invariant and context-aware discriminative cues, respectively. The predictive evidence produced by these two pathways is subsequently fused to obtain the final prediction. Accordingly, we first detail the two pathways, then present the overall training objective, and finally describe the evidence-fusion procedure employed during inference.

Pathway A: Invariant Cue Given the shared representation $\boldsymbol{z}$, the invariant pathway first employs a pathway-specific projection head $\text{FC}(\cdot)$ to obtain

$$\hat{\boldsymbol{z}} = \text{FC}(\boldsymbol{z}). \tag{1}$$

This projection head provides a dedicated representation space for statistical alignment while retaining the shared encoder as the common basis of the two pathways.

To suppress domain-related variations, we apply CORAL regularization Sun & Saenko (2016) to match the second-order statistics of $\hat{\boldsymbol{z}}$ across source domains. Specifically, let $\boldsymbol{C}^d$ denote the covariance matrix[1] of the projected representations from the d-th source domain within the current mini-batch. The alignment objective is formulated as

$$\mathcal{L}_{\text{coral}} = \frac{1}{N(N-1)} \sum_{d=1}^{N} \sum_{d \neq d'} \frac{1}{4M^2} \left\| \boldsymbol{C}^d - \boldsymbol{C}^{d'} \right\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Under this regularization, the projected representation $\hat{\boldsymbol{z}}$ is learned as the statistically aligned representation $\boldsymbol{z}_{\text{inv}}$, which captures invariant cues and is subsequently fed into the pathway-specific classifier g_{inv}.

[1] $\boldsymbol{C}^d = \frac{1}{B-1}\left(\boldsymbol{Z}^d - \mathbf{1}\boldsymbol{\mu}^{d\top}\right)^\top \left(\boldsymbol{Z}^d - \mathbf{1}\boldsymbol{\mu}^{d\top}\right)$, where $\boldsymbol{Z}^d \in \mathbb{R}^{B \times M}$ collects the projected representations and $\boldsymbol{\mu}^d$ is their mean.

Pathway B: Context-Aware Discriminative Cue While Pathway A learns $\boldsymbol{z}_{\text{inv}}$ by aligning representations *across source domains*, the discriminative information it retains may vary in relevance across individual samples. We therefore turn to the *sample itself* and mine the sample-conditioned context from its shared representation $\boldsymbol{z}$, providing discriminative information complementary to the invariant cue.

Specifically, we introduce a shared latent dictionary $\boldsymbol{D} = \{\boldsymbol{b}_r\}_{r=1}^{R}$, which serves as a latent knowledge base for sample-level context mining. Given the shared representation $\boldsymbol{z}$, a query function $q(\cdot)$ first produces a query vector $\boldsymbol{q} = q(\boldsymbol{z})$. We then measure the similarity between $\boldsymbol{q}$ and each latent basis $\boldsymbol{b}_r$, and normalize the resulting scores via a softmax function to obtain the relevance weights $\{\omega_r\}_{r=1}^{R}$:

$$\omega_r = \frac{\exp\left(\text{sim}(\boldsymbol{q}, \boldsymbol{b}_r)/\tau\right)}{\sum_{j=1}^{R} \exp\left(\text{sim}(\boldsymbol{q}, \boldsymbol{b}_j)/\tau\right)}, \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a learnable temperature parameter that controls the sharpness of the weight distribution. Accordingly, the sample-conditioned context can be constructed as a weighted combination of the latent bases Vaswani et al. (2017):

$$\boldsymbol{z}_{\text{c}} = \sum_{r=1}^{R} \omega_r \boldsymbol{b}_r. \tag{4}$$

To prevent the dictionary bases from becoming redundant and encourage them to span diverse latent directions, we herein impose an orthogonality regularizer on $\boldsymbol{D}$:

$$\mathcal{L}_{\text{orth}} = \frac{1}{R(R-1)} \left\| \bar{\boldsymbol{D}} \bar{\boldsymbol{D}}^{\top} - \mathbb{I}_R \right\|_1, \tag{5}$$

where $\bar{\boldsymbol{D}} \in \mathbb{R}^{R \times M}$ denotes the row-wise ℓ_2-normalized dictionary, $\mathbb{I}_R$ denotes the identity matrix, and $\|\cdot\|_1$ denotes the element-wise ℓ_1 norm. This regularizer encourages orthogonality by minimizing the pairwise cosine similarities between normalized bases.

Subsequently, to obtain a context-aware discriminative representation, we utilize the retrieved context $\boldsymbol{z}_{\text{c}}$ to modulate the aligned representation $\boldsymbol{z}_{\text{inv}}$. As $\boldsymbol{z}_{\text{c}}$ provides a compact summary of the context associated with the current sample, using it alone for prediction may overlook discriminative details retained in $\boldsymbol{z}_{\text{inv}}$. We therefore preserve $\boldsymbol{z}_{\text{inv}}$ as the representational basis and treat $\boldsymbol{z}_{\text{c}}$ as a conditioning signal for its sample-conditioned adaptation. Accordingly, this transformation can be implemented through Feature-wise Linear Modulation (FiLM) Perez et al. (2018), yielding

$$\boldsymbol{z}_{\text{dis}} = \text{FiLM}\left(\boldsymbol{z}_{\text{inv}}, \boldsymbol{z}_{\text{c}}\right) = \left[\mathbf{1} + \boldsymbol{\gamma}(\boldsymbol{z}_{\text{c}})\right] \odot \boldsymbol{z}_{\text{inv}} + \boldsymbol{\beta}(\boldsymbol{z}_{\text{c}}), \tag{6}$$

where $\boldsymbol{\gamma}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ are learnable mappings that generate feature-wise scale and shift parameters, respectively, and $\odot$ denotes element-wise multiplication. In this way, the retrieved context adaptively transforms the aligned representation, conditioning its feature responses on the current sample.

Since $\boldsymbol{z}_{\text{c}}$ and $\boldsymbol{z}_{\text{inv}}$ are both derived from the shared representation $\boldsymbol{z}$, they may encode redundant information. To this end, we employ class-conditional HSIC Ma et al. (2020) to penalize their dependence within each class:

$$\mathcal{L}_{\text{cHSIC}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{(n_k - 1)^2} \text{Tr}\left(\boldsymbol{K}_{\text{c}}^{(k)} \boldsymbol{H}^{(k)} \boldsymbol{K}_{\text{inv}}^{(k)} \boldsymbol{H}^{(k)}\right), \tag{7}$$

where n_k is the number of samples from class k in the current mini-batch, $\text{Tr}(\cdot)$ denotes the trace operator, $\boldsymbol{K}_{\text{c}}^{(k)}$ ($\boldsymbol{K}_{\text{inv}}^{(k)}$) is the Gram matrices constructed from $\boldsymbol{z}_{\text{c}}$ ($\boldsymbol{z}_{\text{inv}}$), and $\boldsymbol{H}^{(k)} = \boldsymbol{I}_{n_k} - \frac{1}{n_k}\mathbf{1}\mathbf{1}^{\top}$ is the centering matrix.

Together with pathway-specific classification supervision, this objective encourages $\boldsymbol{z}_{\text{c}}$ to capture class-relevant contextual variation beyond that encoded in $\boldsymbol{z}_{\text{inv}}$, thereby *supporting a complementary context-aware discriminative* cue.

Finally, $\boldsymbol{z}_{\text{dis}}$ is fed into the pathway-specific classifier g_{dis}.

Algorithm 1 CoDER

Input: Training set, batch size B, number of source domains N, number of class K, number of dictionary bases R, maximum iteration T, and the trade-offs λ_1, λ_2, and λ_3
Parameters: An encoder h, two classifiers g_{inv} and g_{dis}, a projection head FC, a query function q, a dictionary $\boldsymbol{D}$, and two learnable mappings $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$

for $t \leftarrow 1$ to T **do**
  Randomly sample a batch $S = \{(\boldsymbol{x}_b, y_b)\}_{b=1}^{B}$
  # Extract the latent representation
  $\boldsymbol{z} \leftarrow h(\boldsymbol{x})$
  # Pathway A: Invariant Cue
  $\boldsymbol{z}_{\text{inv}} \leftarrow$ Eq. (1) with FC
  $\mathcal{L}_{\text{CORAL}} \leftarrow$ Eq. (2)
  # Pathway B: Context-Aware Discriminative Cue
  $\boldsymbol{\omega} \leftarrow$ Eq. (3) with q and $\boldsymbol{D}$
  $\boldsymbol{z}_{\text{c}} \leftarrow$ Eq. (4)
  $\boldsymbol{z}_{\text{dis}} \leftarrow$ Eq. (6) with $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$
  $\mathcal{L}_{\text{orth}} \leftarrow$ Eq. (5) with $\boldsymbol{D}$
  $\mathcal{L}_{\text{cHSIC}} \leftarrow$ Eq. (7)
  # Overall Objective Function
  $\mathcal{L} \leftarrow$ Eq. (8) with g_{inv} and g_{dis}
  Update network parameters
end for

Overall Training Objective Algorithm 1 summarizes the overall training procedure. During training, both pathways are supervised by the ground-truth labels to ensure that their representations remain predictive. To this end, the total objective function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{CORAL}} + \lambda_2 \mathcal{L}_{\text{orth}} + \lambda_3 \mathcal{L}_{\text{cHSIC}}, \tag{8}$$

where λ_1, λ_2, and λ_3 are trade-off coefficients. The classification objective is defined as $\mathcal{L}_{\text{CE}} = \frac{1}{2}\left(\mathcal{L}_{\text{CE}}^{\text{inv}} + \mathcal{L}_{\text{CE}}^{\text{dis}}\right)$, where $\mathcal{L}_{\text{CE}}^{\rho} = \frac{1}{B}\sum_i \ell_{\text{CE}}\left(g_\rho\left(\boldsymbol{z}_\rho^i\right), y_i\right)$ for $\rho \in \{\text{inv}, \text{dis}\}$.

Evidence-Fusion Inference Procedure At inference, each pathway produces a non-negative evidence vector by applying the $\text{Softplus}(\cdot)$ function to its classifier output:

$$\boldsymbol{e}_\rho = \text{Softplus}\left(g_\rho(\boldsymbol{z}_\rho)\right), \quad \rho \in \{\text{inv}, \text{dis}\}, \tag{9}$$

where $\text{Softplus}(\cdot)$ ensures non-negative evidence.

Following Sensoy et al. (2018), each evidence vector parameterizes a Dirichlet distribution as $\boldsymbol{\alpha}_\rho = \boldsymbol{e}_\rho + \mathbf{1}$. The corresponding belief mass and uncertainty are then defined as

$$m_\rho^k = \frac{e_\rho^k}{S_\rho}, \quad u_\rho = \frac{K}{S_\rho}, \tag{10}$$

where $S_\rho = \sum_{k=1}^{K} \alpha_\rho^k$ denotes the total Dirichlet strength, with a larger value indicating greater overall evidence and lower uncertainty for pathway ρ.

Accordingly, the fused belief mass and uncertainty can be formulated as

$$m_{\text{f}}^k = \frac{m_{\text{inv}}^k u_{\text{dis}} + m_{\text{dis}}^k u_{\text{inv}}}{\mathcal{C}}, \quad u_{\text{f}} = \frac{u_{\text{inv}} u_{\text{dis}}}{\mathcal{C}}, \tag{11}$$

where $\mathcal{C} = u_{\text{inv}} u_{\text{dis}} + u_{\text{inv}}(1 - u_{\text{dis}}) + u_{\text{dis}}(1 - u_{\text{inv}})$ is the normalization factor.

The final prediction is given by $\hat{y} = \arg\max_k m_{\text{f}}^k$.

4 Experiments

In this section, extensive experiments have been constructed to comprehensively evaluate the effectiveness of CoDER on five Benchmarks.

Table 1: Overall accuracy results on five Benchmarks. The **bold** and underline items are the best and the second-best results, respectively. † denotes our reproduced results.

Methods	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Avg
ERM (1998)	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.3	63.3
Mixup (2020)	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	39.2 ± 0.4	63.4
DANN (2016)	78.6 ± 0.7	83.6 ± 1.1	65.9 ± 0.6	46.7 ± 1.6	38.3 ± 0.3	62.6
MLDG (2018a)	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.3	63.6
GroupDRO (2020)	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	60.7
IRM (2019)	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 1.9	61.6
ARM (2021)	77.6 ± 0.7	85.1 ± 0.7	64.8 ± 0.4	45.5 ± 1.3	35.5 ± 0.5	61.7
VREx (2021)	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 3.0	61.9
EQRM (2022)	77.8 ± 0.6	86.5 ± 0.2	67.5 ± 0.3	47.8 ± 0.6	41.0 ± 0.6	64.1
Fish (2022)	77.8 ± 0.3	85.5 ± 0.3	68.6 ± 0.4	45.1 ± 1.3	42.7 ± 0.2	63.9
Fishr (2022)	77.8 ± 0.1	85.5 ± 0.4	67.8 ± 0.1	47.4 ± 1.6	41.7 ± 0.3	64.0
CORAL (2016)	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.2	64.6
MMD (2018b)	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	58.8
SagNet (2021)	77.8 ± 0.7	86.3 ± 0.5	68.1 ± 0.3	48.6 ± 1.8	40.3 ± 0.3	64.2
Mixstyle (2024)	77.6 ± 0.9	85.2 ± 1.1	60.4 ± 0.5	44.0 ± 1.3	34.0 ± 0.3	60.2
RDM (2024)	78.4 ± 0.4	87.2 ± 0.7	67.3 ± 0.4	47.5 ± 1.0	43.4 ± 0.4	64.8
SAGM (2023a)	$\underline{80.0 \pm 0.4}$	86.6 ± 0.3	70.1 ± 0.2	48.8 ± 0.8	$\underline{45.0 \pm 0.2}$	66.1
PGrad (2023b)	78.9 ± 0.5	86.2 ± 0.6	69.8 ± 0.3	50.7 ± 1.0	41.0 ± 0.2	65.3
TCRI† (2024)	79.9 ± 0.6	87.0 ± 0.5	67.9 ± 0.4	$\underline{51.2 \pm 0.8}$	42.2 ± 0.1	65.6
I^3C (2025)	79.6 ± 0.9	87.1 ± 0.9	$\underline{70.2 \pm 0.5}$	49.6 ± 1.9	42.6 ± 0.6	65.8
GGA (2025)	79.9 ± 0.4	87.3 ± 0.4	68.5 ± 0.2	50.6 ± 0.1	$\mathbf{45.2 \pm 0.2}$	$\underline{66.3}$
DA-SAM (2026)	79.6 ± 0.8	$\underline{87.5 \pm 0.4}$	68.4 ± 0.3	50.0 ± 0.2	44.5 ± 0.2	66.0
CoDER (*ours*)	$\mathbf{80.4 \pm 0.2}$	$\mathbf{88.2 \pm 0.2}$	$\mathbf{70.4 \pm 0.4}$	$\mathbf{53.5 \pm 0.4}$	44.4 ± 0.1	**67.4**

4.1 Datasets and Settings

For the architecture, our CoDER aligns with the settings outlined in DomainBed Gulrajani & Lopez-Paz (2021), wherein ResNet-50 has been selected as the backbone and the training-domain validation method has been employed for model selection. Extensive experiments are constructed on five Benchmarks, i.e., VLCS, PACS, OfficeHome, TerraIncognita, and DomainNet. The average results on these dataset are reported in Tab. 1.

Twenty recent strong comparison methods and two other representative methods are deployed to compare with our CoDER. These methods can be divided into six categories: 1) *distribution robust method* (GroupDRO Sagawa et al. (2020)), 2) *causal methods learning invariance* (IRM Arjovsky et al. (2019), VREx Krueger et al. (2021), EQRM Eastwood et al. (2022), TCRI Salaudeen & Koyejo (2024)), 3) *gradient matching* (Fish Shi et al. (2022), Fishr Rame et al. (2022), SAGM Wang et al. (2023a), PGrad Wang et al. (2023b), GGA Ballas & Diou (2025), DA-SAM He et al. (2026)), 4) *representation distribution matching* (MMD Li et al. (2018b), CORAL Sun & Saenko (2016), DANNGanin et al. (2016), RDM Nguyen et al. (2024), I^3C Zhou et al. (2025), 5) *meta-learning* (MLDG Li et al. (2018a), ARM Zhang et al. (2021)) 6) and *other variants* (Mixup Yan et al. (2020), SagNet Nam et al. (2021), Mixstyle Zhou et al. (2024)). For more details on datasets, comparison methods, and experimental settings, please refer to our supplementary material.

4.2 Results and Analysis

As reported in Tab. 1, we can make the following key observations: *1)* Overall, our CoDER consistently achieves top-tier performance across all benchmarks and delivers the best average result. Notably, it surpasses the strongest competing results by *2.3% on TerraIncognita* and by *1.1% on average*, demonstrating a clear advantage under diverse domain shifts. These gains validate the effectiveness of our proposed CoDER. *2)* Compared with CORAL, which serves as the baseline alignment technique in our statistical-invariance pathway, CoDER achieves consistent improvements across all five benchmarks, yielding *an average gain of 2.8%*. This improvement is particularly pronounced

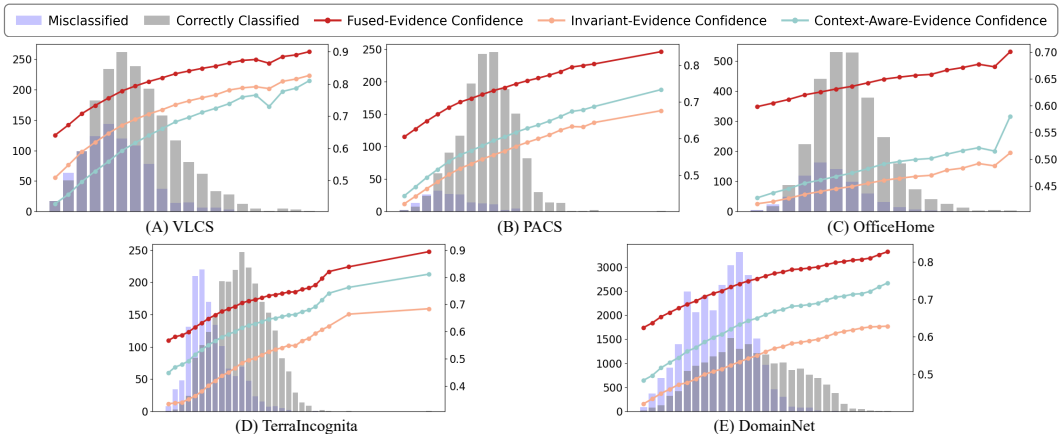

Figure 3: Distributions of classification outcomes and trends in predictive confidence with respect to the modulation magnitude $\Delta = \|\boldsymbol{z}_{\text{dis}} - \boldsymbol{z}_{\text{inv}}\|_2$ on five Benchmarks. The bars show the numbers of correctly and incorrectly classified samples within each Δ bin, while the curves report the corresponding mean confidence of the fused, invariant, and context-aware evidence.

Table 2: Ablation study on five Benchmarks. ✓and ✗ indicate whether the corresponding component is included or excluded.

	$\boldsymbol{D}$	FiLM	$\mathcal{L}_{\text{CORAL}}$	$\mathcal{L}_{\text{orth}}$	$\mathcal{L}_{\text{cHSIC}}$	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet
Baseline	✗	✗	✗	✗	✗	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.3
CORAL	✗	✗	✓	✗	✗	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.2
No/FiLM	✓	✗	✓	✓	✓	79.6 ± 0.7	87.2 ± 0.5	69.5 ± 0.4	51.6 ± 0.9	43.6 ± 0.6
No/D	✗	✓	✓	✗	✓	79.3 ± 0.5	87.9 ± 0.4	69.6 ± 0.2	51.8 ± 0.7	43.0 ± 0.2
No/DFiLM	✗	✗	✓	✗	✓	79.0 ± 1.0	86.7 ± 0.4	69.2 ± 0.5	50.9 ± 0.3	42.5 ± 0.2
No/orth	✓	✓	✓	✗	✓	79.5 ± 0.2	87.1 ± 0.6	69.7 ± 0.1	52.0 ± 0.8	43.3 ± 0.3
No/cHSIC	✓	✓	✓	✓	✗	78.6 ± 0.3	85.6 ± 0.8	68.5 ± 0.2	49.8 ± 0.6	41.5 ± 0.4
Ours	✓	✓	✓	✓	✓	80.4 ± 0.2	88.2 ± 0.2	70.4 ± 0.4	53.5 ± 0.4	44.4 ± 0.1

on TerraIncognita, where *the accuracy increases by 5.9%*, from 47.6% to 53.5%. These consistent gains support the benefit of complementing statistical alignment with context-aware modulation and evidence fusion for improved DG. *3)* CoDER exhibits consistently small standard deviations across all five benchmarks, indicating its remarkable stability over multiple experimental runs. Furthermore, although CoDER slightly trails the best-performing method on the challenging DomainNet benchmark, it remains competitive with recent methods and still achieves the highest average accuracy overall. *4)* Compared with I^3C, which further distills cross-domain invariance through class-level operations, our CoDER consistently outperforms it across all benchmark, with a 1.6% gain in average accuracy. These consistent gains highlight the advantage of integrating diverse and complementary predictive cues over further strengthening domain invariance alone. *5)* Compared with recent optimization-based methods, including GGA and DA-SAM, our CoDER improves the average accuracy by 1.1% and 1.4%, respectively. Coupled with the training cost analysis presented below, these results further support the effectiveness and efficiency of CoDER.

Analysis of Modulation Magnitude Δ Fig. 3 analyzes the effect of context-aware modulation by measuring $\Delta = \|\boldsymbol{z}_{\text{dis}} - \boldsymbol{z}_{\text{inv}}\|_2$, which quantifies the magnitude of the sample-conditioned transformation. For each benchmark, samples are binned into intervals based on Δ to evaluate the corresponding classification outcomes and predictive confidence. In this way, we can observe two key insights. *First*, misclassified samples are mainly concentrated in the small-Δ region, whereas correctly classified samples extend over a broader range of modulation magnitudes. This pattern suggests that a more pronounced sample-conditioned transformation is often associated with successful predictions, indicating that the retrieved context can meaningfully adjust the aligned representation when additional adaptation is needed. *Second*, the fused evidence consistently exhibits higher con-

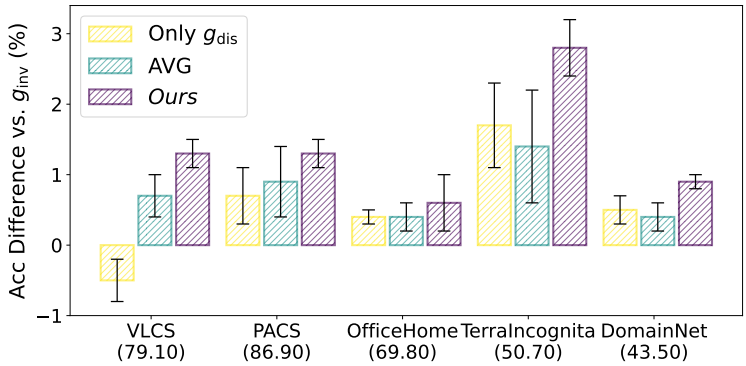

Figure 4: Performance comparison of different prediction strategies relative to the invariant pathway g_{inv} on five DG benchmarks. The bars report accuracy differences, and the values in parentheses denote the absolute accuracy of g_{inv}.

Table 3: The time cost(s) per batch compared to recent studies on representative Benchmarks.

	TCRI	I^{3}C	GGA	DA-SAM	*Ours*
PACS	0.432	<u>0.217</u>	0.230	0.443	**0.209**
OfficeHome	0.445	<u>0.283</u>	0.296	0.454	**0.237**
DomainNet	0.623	<u>0.406</u>	0.461	0.666	**0.392**

fidence than either individual pathway, followed by the invariant and context-aware evidence. The invariant and context-aware pathways generally maintain comparable confidence levels, where the context-aware pathway is relatively higher confident. The consistently increased confidence after fusion indicates that the two pathways provide mutually supportive evidence for prediction.

Ablation Study Tab. 2 reports the ablation results of CoDER. The complete model consistently achieves the best performance across all benchmarks. Omitting either the latent dictionary $\boldsymbol{D}$ or FiLM leads to clear performance degradation, while removing both causes a larger average drop, confirming the complementary roles of context mining and sample-conditioned modulation. The two auxiliary regularization terms are also beneficial. Notably, removing $\mathcal{L}_{\text{cHSIC}}$ results in the largest decrease among all ablations. Overall, these results support the effectiveness of jointly learning statistically aligned and context-aware representations.

Fig. 4 compares the two individual pathways and different fusion strategies. The individual pathways exhibit different strengths across benchmarks, indicating that they capture complementary predictive cues. Simple AVG outperforms both individual pathways on partial benchmarks and remains inferior to evidence fusion. Coupled with the consistently increased fused confidence in Fig. 3, these results support the effectiveness of our evidence fusion strategy. Notably, the independently evaluated g_{inv} learned via CoDER consistently outperforms the original CORAL, suggesting that collaborative learning with the context-aware pathway also *improves the quality of the statistically aligned representation*.

Time Cost Tab. 3 compares the per-batch training cost of CoDER with representative DG methods. Our CoDER consistently achieves the lowest time cost across all three benchmarks, indicating that its performance gains do not incur substantial computational overhead.

5 Conclusion

We develop CoDER, a collaborative dual-evidence representation learning framework for DG to addresses the potential unreliability of a single representation, particularly an invariant one, across diverse unseen target domains. Extensive experiments demonstrate that CoDER can **improve generalization** while **maintaining high computational efficiency**.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with the presentation and writing of this paper, including refining the paper structure, drafting and editing text for clarity and concision. We have not used generative AI tools for generating experimental data, implement the evaluated methods, or execute experiment. We have reviewed all AI-assisted work. We manually ensure the citations are correct. LLM-generated code was verified and tested for correctness by the authors. The core research idea and scientific contributions were developed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code is available in the supplementary materials.

REFERENCES

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

Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20558–20568, 2025.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Meng Cao and Songcan Chen. Guidance not obstruction: A conjugate consistent enhanced strategy for domain generalization. *arXiv e-prints*, pp. arXiv–2412, 2024.

Meng Cao and Songcan Chen. A simple yet lightweight module for enhancing domain generalization through relative representation. *Pattern Recognition*, pp. 112423, 2025.

Sentao Chen, Lei Wang, Zijie Hong, and Xiaowei Yang. Domain generalization by joint-product distribution alignment. *Pattern Recognition*, 134:109086, 2023.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE international conference on computer vision*, pp. 1657–1664, 2013.

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.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, pp. 11856–11884, 2021.

Junwen He, Yang He, Lebing Zheng, Zirui Yin, Hong-Yu Zhang, and Yulong Wang. Navigating the flatlands: Dual adaptive sharpness-aware minimization for domain generalization. In *Forty-third International Conference on Machine Learning*, 2026.

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*, pp. 5815–5826. PMLR, 2021.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022.

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*, pp. 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, pp. 3490–3497, 2018a.

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*, pp. 5400–5409, 2018b.

Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5085–5092, 2020.

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

Toan Nguyen, Kien Do, Bao Duong, and Thin Nguyen. Domain generalisation via risk distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2790–2799, 2024.

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

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, pp. 2661–2679, 2020.

Olawale Salaudeen and Sanmi Koyejo. Causally inspired regularization enables domain general representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3124–3132. PMLR, 2024.

Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Domain generalization of 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18077–18087, 2023.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, pp. 15803–15830, 2022.

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

Vladimir Naumovich Vapnik, Vlamimir Vapnik, et al. Statistical learning theory. 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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*, pp. 5018–5027, 2017.

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3769–3778, 2023a.

Zhe Wang, Jake Grigsby, and Yanjun Qi. PGrad: Learning principal gradients for domain generalization. In *The Eleventh International Conference on Learning Representations*, pp. 17807–17827, 2023b.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

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

Chaoyang Zhou, Zengmao Wang, and Bo Du. Learning intrinsic invariance within intra-class for domain generalization. *IEEE Transactions on Multimedia*, pp. 3807–3820, 2025.

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.

Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7108–7118, 2022.

A Additional Empirical Studies for Collaborative Representation

In the main manuscript, Figs. 3 and Fig. 4 have provided indirect evidence that the two pathways involved learn complementary representations. In Fig. 3, larger modulation magnitudes are generally associated with higher predictive confidence, while in Fig. 4 classifiers built on $\boldsymbol{z}_{\text{inv}}$ and $\boldsymbol{z}_{\text{dis}}$ exhibit different strengths across benchmarks. To further examine this complementarity, the joint correctness patterns of the two pathway are reported in Tab. 4.

Table 4: Joint correctness patterns of the two pathways across five benchmarks. B.C. and B.Inc. denote the proportion of samples correctly and incorrectly classified by both pathways, respectively, while F.B.C. denotes the proportion of B.C. samples whose correct predictions are retained after evidence fusion. $\text{Com}_{\text{inv}\leftarrow\text{dis}}$ denotes the proportion of samples misclassified by g_{inv} but correctly classified by g_{dis} among all samples misclassified by g_{inv}, while $\text{Com}_{\text{dis}\leftarrow\text{inv}}$ is defined analogously with the roles of the two pathways reversed. F.R. denotes the proportion of samples correctly classified by exactly one pathway whose correct predictions are retained after evidence fusion.

Accuracy			Correctness Patterns					
Only g_{inv} (↑)	Only g_{dis} (↑)	*Ours* (↑)	B.C. (↑)	B.Inc. (↓)	F.B.C. (↑)	$\text{Com}_{\text{inv}\leftarrow\text{dis}}$ (↑)	$\text{Com}_{\text{dis}\leftarrow\text{inv}}$ (↑)	F.R. (↑)
VLCS								
79.1	78.6	80.4	75.4	17.7	100	15.3	17.3	72.5
PACS								
86.9	87.6	88.2	84.9	10.4	100	20.6	16.1	70.2
OfficeHome								
69.8	70.2	70.4	69.3	29.3	100	3.0	1.7	78.6
TerraIncognita								
50.7	52.4	53.5	48.8	45.7	100	7.3	4.0	85.6
DomainNet								
43.5	43.9	44.4	41.7	54.3	100	3.9	3.2	67.5

For clarity, we first formally define the correctness patterns and statistics utilized in the subsequent analysis. Let $\mathcal{S}$ denote the complete target-domain test set, and let $\mathcal{S}_{\text{inv}}^{C}$, $\mathcal{S}_{\text{dis}}^{C}$, and $\mathcal{S}_{\text{f}}^{C}$ denote the subsets of target-domain samples correctly classified by g_{inv}, g_{dis}, and evidence fusion, respectively. Correspondingly, $\mathcal{S}_{\text{inv}}^{\text{Inc}} = \mathcal{S} \setminus \mathcal{S}_{\text{inv}}^{C}$ and $\mathcal{S}_{\text{dis}}^{\text{Inc}} = \mathcal{S} \setminus \mathcal{S}_{\text{dis}}^{C}$ denote the subsets misclassified by the respective pathways. These sets allow us to characterize the agreement and disagreement between the two pathway classifiers, quantify their directional complementarity, and evaluate how effectively evidence fusion utilizes their distinct correct predictions. Based on these sets, the proportion of samples correctly classified by both pathways is defined as

$$\text{B.C.} = \frac{\left|\mathcal{S}_{\text{inv}}^{C} \cap \mathcal{S}_{\text{dis}}^{C}\right|}{|\mathcal{S}|} \times 100\%, \tag{12}$$

whereas the proportion of samples misclassified by both pathways is defined as

$$\text{B.Inc.} = \frac{\left|\mathcal{S}_{\text{inv}}^{\text{Inc}} \cap \mathcal{S}_{\text{dis}}^{\text{Inc}}\right|}{|\mathcal{S}|} \times 100\%. \tag{13}$$

To evaluate whether evidence fusion preserves predictions on which both pathways agree correctly, we further define

$$\text{F.B.C.} = \frac{\left|\mathcal{S}_{\text{f}}^{C} \cap \mathcal{S}_{\text{inv}}^{C} \cap \mathcal{S}_{\text{dis}}^{C}\right|}{\left|\mathcal{S}_{\text{inv}}^{C} \cap \mathcal{S}_{\text{dis}}^{C}\right|} \times 100\%. \tag{14}$$

We then quantify the directional complementarity between the two pathways. Specifically,

$$\text{Com}_{\text{inv}\leftarrow\text{dis}} = \frac{\left|\mathcal{S}_{\text{inv}}^{\text{Inc}} \cap \mathcal{S}_{\text{dis}}^{C}\right|}{\left|\mathcal{S}_{\text{inv}}^{\text{Inc}}\right|} \times 100\% \tag{15}$$

measures the proportion of errors made by g_{inv} that are correctly covered by g_{dis}. Likewise,

$$\text{Com}_{\text{dis}\leftarrow\text{inv}} = \frac{\left|\mathcal{S}_{\text{dis}}^{\text{Inc}} \cap \mathcal{S}_{\text{inv}}^{C}\right|}{\left|\mathcal{S}_{\text{dis}}^{\text{Inc}}\right|} \times 100\% \tag{16}$$

is defined analogously with the roles of the two pathways reversed. Finally, let

$$\mathcal{S}^{\mathrm{Exc}} = \left(\mathcal{S}_{\mathrm{inv}}^{C} \cap \mathcal{S}_{\mathrm{dis}}^{\mathrm{Inc}}\right) \cup \left(\mathcal{S}_{\mathrm{inv}}^{\mathrm{Inc}} \cap \mathcal{S}_{\mathrm{dis}}^{C}\right) \times 100\% \tag{17}$$

denote the set of samples correctly classified by exactly one pathway. The fusion retention rate is then defined as

$$\mathrm{F.R.} = \frac{\left|\mathcal{S}_{\mathrm{f}}^{C} \cap \mathcal{S}^{\mathrm{Exc}}\right|}{\left|\mathcal{S}^{\mathrm{Exc}}\right|} \times 100\%, \tag{18}$$

which measures the extent to which evidence fusion retains the correct predictions provided exclusively by either pathway.

As shown in Tab. 4, the two pathways exhibit both shared predictive competence and complementary predictive behavior. The substantial proportion of samples correctly classified by both pathways (B.C.) indicates that $\boldsymbol{z}_{\mathrm{inv}}$ and $\boldsymbol{z}_{\mathrm{dis}}$ preserve a common discriminative foundation. More importantly, the nonzero complementarity in both directions across all benchmarks shows that each pathway correctly recognizes a subset of samples misclassified by the other. The relative contributions of the two directions vary across benchmarks, suggesting that the context-aware pathway does not merely duplicate or consistently improve upon the invariant pathway. Instead, the two representations provide distinct predictive cues whose relative usefulness varies with the encountered domain shift.

The proposed evidence fusion further translates this complementarity into improved overall predictions. It retains the correct predictions for all samples jointly classified correctly by the two pathways, as indicated by the consistent $\mathrm{F.B.C.} = 100\%$, while also preserving a large proportion of the correct predictions provided exclusively by either pathway, as reflected by the F.R. scores. Consequently, CoDER consistently outperforms both individual classifiers across all five benchmarks. These observations provide prediction-level evidence supporting the view that $\boldsymbol{z}_{\mathrm{inv}}$ and $\boldsymbol{z}_{\mathrm{dis}}$ encode complementary discriminative information, which is effectively integrated by the evidence fusion.

B Further Analysis of the Ablation Study

In the main manuscript, we report the ablation results in Tab. 2 and Fig. 4, and provide a brief discussion. Here, we present a more detailed analysis of the contribution of each component.

To clarify the ablation variants reported in Tab. 2, No/FiLM removes the FiLM module and directly feeds the representation $\boldsymbol{z}_{\mathrm{c}}$ into g_{dis}, while retaining the remaining objectives. No/D removes the latent dictionary $\boldsymbol{D}$ and utilizes the query projection $q(\boldsymbol{z})$ as the conditioning signal for modulating $\boldsymbol{z}_{\mathrm{inv}}$. Accordingly, the orthogonality objective $\mathcal{L}_{\mathrm{orth}}$ is discarded. No/DFiLM removes both the dictionary $\boldsymbol{D}$ and the FiLM module, and directly feeds the query projection $q(\boldsymbol{z})$ into g_{dis}. Consequently, the two pathways become structurally identical, with each learning a distinct representation through its own objective function. In Fig. 4, AVG denotes the arithmetic mean of the predictions from the two pathways, assigning equal weight to each pathway.

Table 5: Ablation results of the proposed framework. This table is reproduced from Tab. 2 in the main manuscript for convenience.

	$\boldsymbol{D}$	FiLM	$\mathcal{L}_{\mathrm{CORAL}}$	$\mathcal{L}_{\mathrm{orth}}$	$\mathcal{L}_{\mathrm{cHSIC}}$	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet
Baseline	✗	✗	✗	✗	✗	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.3
CORAL	✗	✗	✓	✗	✗	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.2
No/FiLM	✓	✗	✓	✓	✓	79.6 ± 0.7	87.2 ± 0.5	69.5 ± 0.4	51.6 ± 0.9	43.6 ± 0.6
No/D	✗	✓	✓	✗	✓	79.3 ± 0.5	87.9 ± 0.4	69.6 ± 0.2	51.8 ± 0.7	43.0 ± 0.2
No/DFiLM	✗	✗	✓	✗	✓	79.0 ± 1.0	86.7 ± 0.4	69.2 ± 0.5	50.9 ± 0.3	42.5 ± 0.2
No/orth	✓	✓	✓	✗	✓	79.5 ± 0.2	87.1 ± 0.6	69.7 ± 0.1	52.0 ± 0.8	43.3 ± 0.3
No/cHSIC	✓	✓	✓	✓	✗	78.6 ± 0.3	85.6 ± 0.8	68.5 ± 0.2	49.8 ± 0.6	41.5 ± 0.4
Ours	✓	✓	✓	✓	✓	80.4 ± 0.2	88.2 ± 0.2	70.4 ± 0.4	53.5 ± 0.4	44.4 ± 0.1

According to the ablation results in Tab. 2, several observations can be provided. *First*, the complete CoDER consistently outperforms CORAL across all five benchmarks. This result indicates that statistical alignment alone is insufficient and that the additional context-aware representation provides predictive information beyond that retained by the invariant pathway.

Second, both the latent dictionary $\boldsymbol{D}$ and FiLM contribute consistently to the final performance. Removing either component results in an average degradation of approximately 1.1%, while removing both produces a larger drop of 1.7% and consistently underperforms the two single-removal variants. These results support their complementary roles in constructing the context-aware representation. Specifically, the latent dictionary organizes a set of shared bases from which sample-relevant context can be retrieved, whereas FiLM translates the retrieved context into a sample-conditioned transformation of $\boldsymbol{z}_{\mathrm{inv}}$. Without the dictionary, the conditioning signal lacks an explicit mechanism for selecting and combining shared contextual bases. Without FiLM, the retrieved context cannot directly reorganize the aligned representation according to the current sample. Although No/DFiLM still improves over CORAL, its smaller gain suggests that a parallel prediction pathway alone cannot fully exploit the contextual information.

Third, the two auxiliary objectives play distinct roles in maintaining the quality and complementarity of the learned representations. Removing $\mathcal{L}_{\mathrm{orth}}$ causes a consistent performance drop, which is consistent with its role in encouraging diversity among the dictionary bases and reducing redundant context retrieval. More notably, removing $\mathcal{L}_{\mathrm{cHSIC}}$ leads to the largest degradation among the ablated variants, reducing the average performance to a level close to that of CORAL. Since $\boldsymbol{z}_{\mathrm{c}}$ and $\boldsymbol{z}_{\mathrm{inv}}$ are both derived from the shared representation $\boldsymbol{z}$, the context-aware pathway may otherwise reproduce information already retained by the invariant pathway. By reducing their within-class dependence, $\mathcal{L}_{\mathrm{cHSIC}}$ encourages $\boldsymbol{z}_{\mathrm{c}}$ to absorb context-dependent variations beyond $\boldsymbol{z}_{\mathrm{inv}}$, thereby providing a more informative conditioning signal for FiLM. This result further indicates that the benefit of the second pathway does not arise merely from adding another classifier, but from learning non-redundant predictive cues.

Overall, the complete model achieves the best performance on every benchmark. The results show that invariant alignment, context retrieval, sample-conditioned modulation, and redundancy control operate collaboratively rather than independently. Together, they enable the two pathways to preserve a common discriminative foundation while capturing complementary predictive information, which is subsequently integrated through evidence fusion to improve domain generalization.

C Alternative Implementations of the Invariant Pathway

In this section, alternative implementations of the invariant learning objective are adopted to examine the compatibility of our CoDER. Specifically, we replace the CORAL objective in CoDER with several representative methods, including the kernel-based MMD, the adversarial DANN, and the recent RDM, while keeping all other components unchanged. This comparison evaluates whether the collaborative representation learning and evidence fusion mechanisms remain effective under different forms of invariant learning. The results on PACS benchmark are reported in Tab. 6.

Table 6: Performance on PACS using different invariant representation objectives in the statistical-invariance pathway. Vanilla denotes the corresponding invariant representation learning method trained independently. Values in parentheses with upward arrows indicate improvements over the corresponding Vanilla baselines.

	CORAL	MMD	DANN	RDM
Vanilla	86.2 ± 0.3	84.6 ± 0.3	83.6 ± 1.1	87.2 ± 0.7
Only g_{inv}	86.9 ± 0.5 (+0.7 ↑)	86.4 ± 0.2 (+1.8 ↑)	86.0 ± 0.5 (+2.4 ↑)	87.7 ± 0.3 (+0.5 ↑)
Ours	88.2 ± 0.2 (+2.0 ↑)	87.9 ± 0.2 (+3.3 ↑)	86.8 ± 0.5 (+3.2 ↑)	88.6 ± 0.2 (+1.4 ↑)

As shown in Tab. 6, collaborative training consistently improves the invariant classifier under all four implementations. In particular, Only g_{inv} outperforms its corresponding Vanilla baseline regardless of the invariant learning method employed. This indicates that the context-aware pathway not only provides additional predictive evidence at inference time but also refines the representation learned by the invariant pathway through joint optimization. This improvement primarily stems from the fact that $\boldsymbol{z}_{\mathrm{c}}$ captures context-dependent variations that would otherwise be encoded in $\boldsymbol{z}_{\mathrm{inv}}$, thereby enabling the latter to focus more strictly on statistically invariant and class-discriminative cues.

Evidence fusion yields further improvements over the independently evaluated g_{inv} in every setting. The consistent gains obtained with involved invariant objectives suggest that the complementary

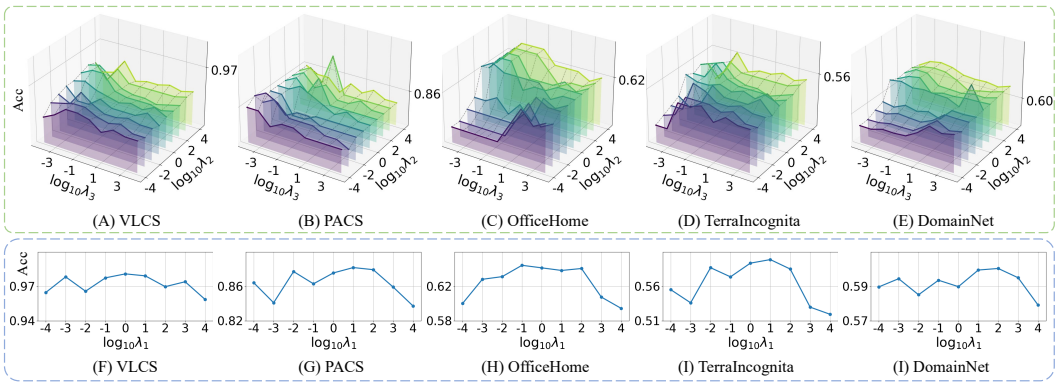

Figure 5: Hyperparameter sensitivity analysis of CoDER. The top panel shows the joint sensitivity to λ_2 and λ_3, while the bottom panel shows the sensitivity to λ_1. All hyperparameter values are displayed on a $\log_{10}$ scale, e.g., $0 = \log_{10} 1$.

information introduced by the context-aware pathway can be effectively integrated with different forms of invariant cues. Notably, improvements are observed for both relatively weak and strong Vanilla baselines, indicating that the benefit is not limited to compensating for a particular invariant learning method. These results reveal that CoDER provides a general collaborative representation learning and evidence fusion framework, rather than an extension specifically tailored to CORAL.

D Hyperparameter Sensitivity Analysis

Fig. 5 presents the sensitivity to λ_1, as well as the joint impact of λ_2 and λ_3. The accuracy remains relatively stable over a broad range of (λ_2, λ_3) combinations, with noticeable degradation occurring mainly at extreme values. Similarly, the model maintains consistent performance for $\lambda_1 \in [10^{-3}, 10^{2}]$, indicating limited sensitivity to hyperparameter selection.

E Limitations

Although CoDER is designed as a general framework discussed in Appendix C, Pathway B involves a relatively complex set of components. Exploring a more streamlined design without sacrificing performance is a promising direction for future work.

F Datasets and Settings

Datasets Five commonly used DG benchmarks have been included in our main manuscript. VLCS Fang et al. (2013) consists of images from 4 different domains: Voc2007, LabelMe, Caltech101, Sun09. It contains around 10,729 images from 5 different classes. PACS Li et al. (2017) consists of images from 4 different domains: Art, Cartoon, Photo, Sketch. It contains around 9,991 images from 7 different classes. Office-Home (Venkateswara et al., 2017) consists of images from 4 different domains: Art, Clipart, Product and Real-World. It contains around 15,500 images from 65 different categories. TerraIncognita (Beery et al., 2018) consists of images from 4 different domains: L100, L38, L43, L46. It contains around 24,778 images from 10 different classes. DomainNet (Peng et al., 2019), which is the largest and most complex dataset currently available, consists of images from 6 domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. It contains over 600,000 images from 345 different classes.

Settings Following the settings on DomainBed Gulrajani & Lopez-Paz (2021), we adopt ResNet-50 architecture as the backbone with the latent representation dimension M set to 1,024, and each input image is resized and cropped to 224×224. Adam is utilized as the optimizer. The training

Algorithm	C	L	S	V	Avg
ERM	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
Mixup	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4
DANN	99.0 ± 0.3	65.1 ± 1.4	73.1 ± 0.3	77.2 ± 0.6	78.6
MLDG	97.4 ± 0.2	65.2 ± 0.7	71.0 ± 1.4	75.3 ± 1.0	77.2
GroupDRO	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7
IRM	98.6 ± 0.1	64.9 ± 0.9	73.4 ± 0.6	77.3 ± 0.9	78.6
ARM	98.7 ± 0.2	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
VREx	98.4 ± 0.3	64.4 ± 1.4	74.1 ± 0.4	76.2 ± 1.3	78.3
EQRM	98.3 ± 0.0	63.7 ± 0.8	72.6 ± 1.0	76.7 ± 1.1	77.8
Fish	-	-	-	-	77.8
Fishr	98.9 ± 0.3	64.0 ± 0.5	71.5 ± 0.2	76.8 ± 0.7	77.8
CORAL	98.3 ± 0.1	66.1 ± 1.2	73.4 ± 0.3	77.5 ± 1.2	78.8
MMD	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.4
SagNet	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8
Mixstyle	98.6 ± 0.3	64.5 ± 1.1	72.6 ± 0.5	75.7 ± 1.7	77.8
RDM	98.1 ± 0.2	64.9 ± 0.7	72.6 ± 0.5	77.9 ± 1.2	78.4
SAGM	99.0 ± 0.2	65.2 ± 0.4	75.1 ± 0.2	**80.7 ± 0.8**	80.0
PGrad	**99.1 ± 0.1**	63.8 ± 0.7	73.5 ± 0.5	79.0 ± 0.5	78.8
TCRI	**99.1 ± 0.4**	67.2 ± 0.6	74.4 ± 0.6	79.0 ± 0.6	79.9
I^3C	98.8 ± 0.2	**67.5 ± 0.6**	72.1 ± 2.0	80.2 ± 0.8	79.6
GGA	**99.1 ± 0.2**	**67.5 ± 0.6**	75.1 ± 0.3	78.0 ± 0.1	79.9
DA-SAM	98.7 ± 0.2	65.9 ± 1.3	75.2 ± 0.8	78.8 ± 1.3	79.6
CoDER	98.7 ± 0.3	**67.5 ± 0.2**	**76.2 ± 0.4**	79.1 ± 0.4	**80.4**

Table 7: Overall results on VLCS Benchmark. The bold and underline items are the best and the second-best results, respectively.

data are randomly split into two parts: 80% for training and 20% for validation. The best model on the validation split is selected to evaluate the target domain. The training batch size is set to 32 per source domain for all benchmarks except DomainNet, where it is set to 64. The test batch size is set to 64 for all benchmarks. Maximum iteration T is set to 5,000. The default learning rate is set to $5e$-5. Three independent trials are conducted, each with 20 randomly sampled hyperparameter configurations. In CoDER, the number of dictionary bases R is set to $2N$, where N denotes the number of source domains. For OfficeHome and DomainNet, λ_1 is sampled log-uniformly from $[10^0, 10^2]$, with a default value of 10; for the remaining benchmarks, it is sampled from $[10^{-1}, 10^1]$, with a default value of 1. Both λ_2 and λ_3 are sampled log-uniformly from $[10^{-1}, 10^1]$, with their default values set to 1. The codes are run on Python 3.10 and PyTorch 2.7 on Arch Linux with NVIDIA GeForce RTX 4090 and RTX 5090 GPUs.

G Full Experimental Results

In this section, we provide the complete results on all benchmarks in Tab. 7-11. Our CoDER achieves the best performance in most cases, further demonstrating its effectiveness across diverse domain shifts.

Algorithm	A	C	P	S	Avg
ERM	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5
Mixup	86.1 ± 0.5	78.9 ± 0.8	97.6 ± 0.1	75.8 ± 1.8	84.6
DANN	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
MLDG	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
GroupDRO	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
IRM	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
ARM	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
VREx	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
EQRM	86.5 ± 0.4	82.1 ± 0.7	96.6 ± 0.2	80.8 ± 0.2	86.5
Fish	-	-	-	-	85.5
Fishr	88.4 ± 0.2	78.7 ± 0.7	97.0 ± 0.1	77.8 ± 2.0	85.5
CORAL	88.3 ± 0.2	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	86.2
MMD	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
SagNet	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	80.0 ± 0.4	86.3
Mixstyle	86.8 ± 0.5	79.0 ± 1.4	96.9 ± 0.1	78.5 ± 2.3	85.3
RDM	88.4 ± 0.2	81.3 ± 1.6	97.1 ± 0.1	81.8 ± 1.1	87.1
SAGM	87.4 ± 0.2	80.2 ± 0.3	98.0 ± 0.2	80.8 ± 0.6	86.6
PGrad	**89.6 ± 0.3**	78.9 ± 0.6	97.7 ± 0.3	78.8 ± 1.0	86.2
TCRI	87.8 ± 0.7	81.8 ± 0.6	**98.1 ± 0.4**	80.3 ± 0.4	87.0
I^3C	89.2 ± 1.1	80.6 ± 1.3	97.7 ± 0.6	81.4 ± 0.6	87.2
GGA	88.8 ± 0.2	80.1 ± 0.3	97.3 ± 0.2	81.2 ± 0.5	86.8
DA-SAM	88.6 ± 0.8	82.9 ± 0.4	96.7 ± 0.3	**81.9 ± 1.2**	87.5
CoDER	88.7 ± 0.2	**83.6 ± 0.6**	**98.1 ± 0.2**	**81.9 ± 1.2**	**88.1**

Table 8: Overall results on PACS Benchmark.

Algorithm	A	C	P	R	Avg
ERM	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
Mixup	62.4 ± 0.8	54.8 ± 0.6	76.9 ± 0.3	78.3 ± 0.2	68.1
DANN	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.8
MLDG	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
GroupDRO	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
IRM	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3
ARM	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
VREx	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
EQRM	60.5 ± 0.1	56.0 ± 0.2	76.1 ± 0.4	77.4 ± 0.3	67.5
Fish	-	-	-	-	68.6
Fishr	62.4 ± 0.5	54.4 ± 0.4	76.2 ± 0.5	78.3 ± 0.1	67.8
CORAL	65.3 ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.6
MMD	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.4
SagNet	63.4 ± 0.2	54.8 ± 0.4	75.8 ± 0.4	78.3 ± 0.3	68.1
Mixstyle	51.1 ± 0.3	53.2 ± 0.4	68.2 ± 0.7	69.2 ± 0.6	60.4
RDM	61.1 ± 0.4	55.1 ± 0.3	75.7 ± 0.5	77.3 ± 0.3	67.3
SAGM	65.4 ± 0.4	**57.0 ± 0.3**	78.0 ± 0.3	80.0 ± 0.2	70.1
PGrad	65.8 ± 0.2	55.4 ± 0.4	78.0 ± 0.1	80.0 ± 0.4	69.8
TCRI	62.9 ± 0.6	54.0 ± 0.0	76.5 ± 0.5	78.0 ± 0.4	67.8
I^3C	**67.5 ± 0.8**	56.1 ± 0.4	77.5 ± 0.4	79.8 ± 0.4	70.2
GGA	64.3 ± 0.1	54.4 ± 0.2	76.5 ± 0.3	78.9 ± 0.2	68.5
DA-SAM	64.2 ± 0.7	53.8 ± 1.0	77.4 ± 0.2	78.1 ± 0.7	68.4
CoDER	66.9 ± 0.5	56.5 ± 0.4	**78.2 ± 0.8**	**80.2 ± 0.3**	**70.4**

Table 9: Overall results on OfficeHome Benchmark.

Algorithm	L100	L38	L43	L46	Avg
ERM	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
Mixup	59.6 ± 2.0	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
DANN	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
MLDG	54.2 ± 3.0	44.3 ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.8
GroupDRO	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
IRM	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
ARM	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
VREx	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.3
EQRM	47.9 ± 1.9	45.2 ± 0.3	59.1 ± 0.3	38.8 ± 0.6	47.8
Fish	-	-	-	-	45.1
Fishr	50.2 ± 3.9	43.9 ± 0.8	55.7 ± 2.2	39.8 ± 1.0	47.4
CORAL	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.7
MMD	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2
SagNet	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	48.6
Mixstyle	54.3 ± 1.1	34.1 ± 1.1	55.9 ± 1.1	31.7 ± 2.1	44.0
RDM	52.9 ± 1.2	43.1 ± 1.0	58.1 ± 1.3	36.1 ± 2.9	47.6
SAGM	54.8 ± 1.3	41.4 ± 0.8	57.7 ± 0.6	41.3 ± 0.4	48.8
PGrad	**61.2 ± 2.2**	45.7 ± 0.9	58.2 ± 0.3	37.9 ± 0.7	50.8
TCRI	58.9 ± 1.6	45.0 ± 1.0	59.3 ± 0.2	41.5 ± 0.5	51.2
I^3C	57.0 ± 3.5	43.2 ± 2.4	56.4 ± 1.7	41.9 ± 1.3	49.6
GGA	55.9 ± 0.1	45.5 ± 0.1	**59.7 ± 0.1**	41.5 ± 0.1	50.7
DA-SAM	59.0 ± 0.6	39.2 ± 1.1	58.5 ± 1.9	43.3 ± 1.5	50.0
CoDER	60.8 ± 3.7	**49.7 ± 0.3**	59.5 ± 0.9	**44.1 ± 1.2**	**53.5**

Table 10: Overall results on TerraIncognita Benchmark.

Algorithm	C	I	P	Q	R	S	Avg
ERM	58.1 ± 0.3	18.8 ± 0.3	46.7 ± 0.3	12.2 ± 0.4	59.6 ± 0.1	49.8 ± 0.4	40.9
Mixup	55.7 ± 0.3	18.5 ± 0.5	44.3 ± 0.5	12.5 ± 0.4	55.8 ± 0.3	48.2 ± 0.5	39.2
DANN	53.1 ± 0.2	18.3 ± 0.1	44.2 ± 0.7	11.8 ± 0.1	55.5 ± 0.4	46.8 ± 0.6	38.3
MLDG	59.1 ± 0.2	19.1 ± 0.3	45.8 ± 0.7	13.4 ± 0.3	59.6 ± 0.1	50.2 ± 0.4	41.2
GroupDRO	47.2 ± 0.5	17.5 ± 0.4	33.8 ± 0.5	9.3 ± 0.3	51.6 ± 0.4	40.1 ± 0.6	33.2
IRM	48.5 ± 2.8	15.0 ± 1.5	38.3 ± 4.3	10.9 ± 0.5	48.2 ± 5.2	42.3 ± 3.1	33.9
ARM	49.7 ± 0.3	16.3 ± 0.5	40.9 ± 1.1	9.4 ± 0.1	53.4 ± 0.4	43.5 ± 0.4	35.5
VREx	47.3 ± 3.5	16.0 ± 1.5	35.8 ± 4.6	10.9 ± 0.3	49.6 ± 4.9	42.0 ± 3.0	33.6
EQRM	56.1 ± 1.3	19.6 ± 0.1	46.3 ± 1.5	12.9 ± 0.3	61.1 ± 0.0	50.3 ± 0.1	41.1
Fish	-	-	-	-	-	-	42.7
Fishr	58.2 ± 0.5	20.2 ± 0.2	47.7 ± 0.3	12.7 ± 0.2	60.3 ± 0.2	50.8 ± 0.1	41.7
CORAL	59.2 ± 0.1	19.7 ± 0.2	46.6 ± 0.3	13.4 ± 0.4	59.8 ± 0.2	50.1 ± 0.6	41.5
MMD	32.1 ± 13.3	11.0 ± 4.6	26.8 ± 11.3	8.7 ± 2.1	32.7 ± 13.8	28.9 ± 11.9	23.4
SagNet	57.7 ± 0.3	19.0 ± 0.2	45.3 ± 0.3	12.7 ± 0.5	58.1 ± 0.5	48.8 ± 0.2	40.3
Mixstyle	51.9 ± 0.4	13.3 ± 0.2	37.0 ± 0.5	12.3 ± 0.1	46.1 ± 0.3	43.4 ± 0.4	34.0
RDM	62.1 ± 0.2	20.7 ± 0.1	49.2 ± 0.4	14.1 ± 0.2	63.0 ± 1.3	51.4 ± 0.1	43.4
SAGM	**64.9 ± 0.2**	21.1 ± 0.3	51.5 ± 0.2	**14.8 ± 0.2**	64.1 ± 0.2	53.6 ± 0.2	45.0
PGrad	57.0 ± 0.5	18.2 ± 0.2	48.4 ± 0.2	13.0 ± 0.1	60.9 ± 0.1	48.8 ± 0.1	41.1
TCRI	58.6 ± 0.3	18.6 ± 0.1	49.4 ± 0.2	12.9 ± 0.1	61.5 ± 0.2	50.8 ± 0.1	42.0
I^3C	59.4 ± 0.7	20.8 ± 0.4	49.1 ± 1.1	13.2 ± 0.4	61.8 ± 0.4	51.1 ± 0.8	42.6
GGA	64.0 ± 0.2	**22.2 ± 0.3**	**51.7 ± 0.1**	14.3 ± 0.2	64.1 ± 0.4	**54.3 ± 0.3**	**45.1**
DA-SAM	63.4 ± 0.2	21.9 ± 0.2	50.6 ± 0.6	14.2 ± 0.6	64.4 ± 0.3	52.7 ± 0.7	44.5
CoDER	62.9 ± 0.2	21.5 ± 0.2	50.3 ± 0.4	14.4 ± 0.1	**64.6 ± 0.3**	52.4 ± 0.2	44.4

Table 11: Overall results on DomainNet Benchmark.