

FOCUS: Dependence-Regularized Disentanglement for Domain Generalization

Noaman Mehmood

noaman@udel.edu

University of delaware

Kenneth Barner

University of delaware

Research Article

Keywords: Domain Generalization, Mutual Information, Disentangled Representation Learning, Domain Shift, Representation Learning

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11028594/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

FOCUS: Dependence-Regularized Disentanglement for Domain Generalization

Noaman Mehmood^{1*} and Kenneth E. Barner^{2†}

^{1,2*}Electrical and Computer Engineering Department, University of Delaware, Street, Newark, 19716, Delaware, United States.

*Corresponding author(s). E-mail(s): noaman@udel.edu;

Contributing authors: barner@udel.edu;

[†]These authors contributed equally to this work.

Abstract

Deep neural networks (DNNs) often experience substantial performance degradation when deployed in environments that differ from their training conditions, a challenge commonly referred to as domain shift. Domain Generalization (DG) aims to address this problem by learning representations from multiple source domains that remain robust to distributional changes and generalize effectively to unseen target domains. While many existing DG approaches attempt to suppress or align domain-specific variations within a shared feature space, disentangled representation learning provides an alternative strategy by explicitly separating task-relevant semantic information from domain-dependent factors.

This work investigates Feature Orthogonalization via Contrastive Upper-bound Separation (FOCUS), a mutual information-based disentanglement framework that explicitly models domain-agnostic and domain-specific representations. FOCUS employs separate content and domain branches and reduces statistical dependence between their corresponding embeddings using the Contrastive Log-ratio Upper Bound (CLUB). In this extended study, we further investigate whether the effectiveness of mutual information-based disentanglement depends strongly on the choice of mutual information estimator by introducing a lower-bound formulation based on Mutual Information Neural Estimation (MINE) and systematically comparing it with the original CLUB-based formulation.

Extensive experiments on PACS, Office-Home, VLCS, and TerraIncognita demonstrate that FOCUS achieves average accuracy improvements of 1.26%, 0.70%, 0.71%, and 2.70%, respectively, over strong existing DG methods. The extended comparison further shows that both lower-bound and upper-bound mutual information regularization substantially improve over ERM and achieve

broadly comparable generalization performance. CLUB provides a modest performance advantage when appropriately weighted, whereas MINE produces a more naturally balanced regularization signal during optimization. These findings suggest that the primary benefit stems from the mutual information-based separation of domain-specific and task-relevant representations, whereas the choice of estimator primarily influences the optimization characteristics and the scaling of the regularization objective.

Keywords: Domain Generalization, Mutual Information, Disentangled Representation Learning, Domain Shift, Representation Learning..

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in computer vision by learning highly discriminative representations. However, their performance often degrades substantially when they are deployed on data distributions that differ from those observed during training. This problem, commonly referred to as domain shift [1, 2], limits the reliability of DNNs in real-world applications where environmental conditions, image styles, backgrounds, viewpoints, and acquisition settings may vary significantly. A central premise in Domain Generalization (DG) [3] is that the underlying semantic relationship between inputs and labels remains relatively stable across environments, while the observed input distribution changes due to domain-specific factors. DG therefore aims to train models on multiple source domains so that they can learn representations that remain robust and transferable to unseen target domains.

A wide range of DG methods has been proposed to improve generalization under domain shift. These include adversarial learning approaches [4–7], which encourage domain-invariant representations by reducing the distinguishability of source domains and data augmentation strategies [8–12], which expose models to a broader range of visual variations. Other approaches seek to learn robust domain-agnostic features by minimizing distributional discrepancies among latent representations across source domains [13, 14]. Although these strategies have shown promising results, many of them primarily focus on suppressing or aligning domain-specific variations within a shared feature space.

A key limitation of standard empirical risk minimization (ERM) [15] and many alignment-based DG strategies is that they may exploit spurious correlations that are predictive in the training domains but fail under unseen domain shifts. This issue is especially pronounced in visual recognition, where deep networks can entangle task-relevant semantic content with domain-specific cues such as texture, illumination, background, or style. As a result, the learned classifier may inadvertently rely on nuisance variations, making predictions sensitive to changes in the target environment. Moreover, simply suppressing domain-specific information can discard useful factors that could instead be modeled separately to improve representation learning. Motivated by disentangled representation learning [16–19], recent methods attempt

to address this limitation by separating latent representations into domain-invariant and domain-specific components. The key intuition is that if semantic content and domain/style information are represented in distinct latent subspaces, the classifier can base its decisions primarily on the domain-invariant content representation while reducing reliance on nuisance variations. For example, Polarization of Embeddings (POEM) [17] learns separate representation branches and encourages their separation by minimizing cosine similarity between domain-specific and domain-agnostic embeddings. This promotes geometric decorrelation and encourages linear independence between the two feature spaces.

Inspired by this direction, our previous conference work [20] introduced Feature Orthogonalization via Contrastive Upper-bound Separation (FOCUS), a disentanglement-based framework for Domain Generalization (DG) that explicitly leverages domain-specific information to improve the learning of domain-agnostic representations. FOCUS employs two encoders to learn category-related and domain-specific representations, denoted as Z_c and Z_d , respectively. Two corresponding classifiers predict class and domain labels from these representations. The domain-specific branch is encouraged to capture domain-related factors, whereas the task-specific branch focuses on semantic information expected to transfer across environments. By explicitly modeling these complementary factors, FOCUS seeks to reduce the entanglement of semantic information with domain-dependent variations.

A central motivation behind FOCUS is that purely geometric separation [17] may be insufficient for robust disentanglement. Although cosine similarity or orthogonality-based constraints can reduce geometric overlap or linear correlation between embeddings, they do not necessarily eliminate higher-order or nonlinear statistical dependencies. Consequently, domain-related information may remain encoded within the task-relevant representation and provide spurious predictive cues that fail under unseen domain shifts. FOCUS therefore regularizes the mutual information between Z_c and Z_d , providing a more general dependence criterion that can capture both linear and nonlinear relationships between the two representation spaces.

The original FOCUS formulation employs the Contrastive Log-ratio Upper Bound (CLUB) estimator [21] to obtain a tractable upper-bound estimate of mutual information. Because the objective of disentanglement is to reduce dependence between Z_c and Z_d , an upper-bound formulation is conceptually well aligned with this goal: reducing the estimated upper bound directly constrains an estimate lying above the underlying mutual information. In practice, however, the effectiveness of representation disentanglement may depend not only on the theoretical form of the estimator but also on its numerical scale, optimization behavior, and interaction with the supervised learning objectives.

Motivated by this observation, the present version extends our conference study by systematically investigating the role of mutual information estimation in disentangled DG. In addition to the original CLUB-based upper-bound formulation, we introduce an alternative formulation based on Mutual Information Neural Estimation (MINE) [22], which provides a lower-bound estimate of mutual information. By evaluating both estimators within the same dual-branch representation-learning framework, we examine whether the generalization benefit of mutual information regularization

depends strongly on the specific estimator or instead arises primarily from encouraging separation between task-relevant and domain-specific representations.

The extended analysis reveals that both lower-bound and upper-bound mutual information regularization substantially improve over the ERM baseline and provide broadly comparable generalization performance. The CLUB-based formulation achieves a modest accuracy advantage when appropriately scaled, while the MINE-based formulation produces a mutual-information signal whose magnitude is more naturally balanced with the supervised losses. These observations indicate that the two estimators differ more noticeably in their optimization characteristics and required regularization strength than in their resulting DG accuracy. Thus, the extended study provides a broader perspective on mutual information-based disentanglement by distinguishing the benefit of dependence regularization itself from the effect of the particular estimator used to implement it.

The main contributions of this extended work can therefore be summarized as follows:

- We build upon the previously introduced FOCUS framework [20], which learns complementary domain-agnostic and domain-specific representations and reduces their statistical dependence using mutual information regularization.
- We extend the original formulation with a lower-bound mutual information objective based on MINE and provide a systematic comparison with the CLUB-based upper-bound formulation under a common disentangled representation-learning framework.
- We analyze not only downstream DG accuracy but also the optimization behavior and numerical scaling of the two mutual information objectives, showing that comparable generalization performance can arise despite substantial differences in the magnitude and weighting requirements of their regularization signals.
- We conduct extensive evaluations on PACS [23], Office-Home [24], VLCS [25], and TerraIncognita [26], demonstrating the effectiveness of mutual information-based disentanglement across diverse forms of domain shift.

The remainder of this paper is structured as follows. Section 2 reviews relevant work in domain generalization and disentangled representation learning. Section 3 describes the FOCUS framework and the upper- and lower-bound mutual information formulations considered in this study. Section 4 presents the experimental settings and evaluation results. Section 4.5 provides the ablation studies and comparative analysis of the mutual information estimators. Finally, Section 5 concludes the paper.

2 Related Work

Domain Generalization (DG) [10] aims to learn models from multiple labeled source domains that can generalize effectively to unseen target domains. Unlike domain adaptation, DG assumes that no target-domain data are available during training, making the problem particularly challenging under distributional shifts. A central objective in DG is to learn representations that preserve task-relevant semantic information while

reducing sensitivity to domain-specific variations such as illumination, texture, background, viewpoint, and acquisition conditions. Existing DG approaches can be broadly categorized into domain-invariant representation learning, adversarial learning, data augmentation, and disentangled embedding learning. In the following subsections, we review these directions and position our proposed framework within the broader literature.

2.1 Domain-Invariant Representation Learning

Learning domain-invariant representations has been one of the most widely studied strategies for domain generalization. The main objective is to extract features that remain stable across source domains and are therefore more likely to generalize to unseen environments. Early approaches focused on explicitly reducing distributional discrepancies across domains while preserving class-discriminative information. For example, Domain-Invariant Component Analysis (DICA) [27] learns a feature transformation that suppresses domain-specific variations while retaining label-relevant structure. Similarly, Scatter Component Analysis (SCA) [28] projects data into a latent space where inter-domain discrepancy is reduced and class separability is enhanced.

Another important line of work relies on distribution matching objectives, such as Maximum Mean Discrepancy (MMD), to align latent representations across source domains. These methods encourage samples from different domains to share similar feature distributions, thereby reducing domain-dependent variation. Domain-Adversarial Neural Networks (DANN) [5] extend this idea by introducing an adversarial loss that encourages the feature extractor to learn representations that are difficult for a domain discriminator to distinguish. By reducing domain-specific cues in the latent space, these methods promote domain-invariant feature learning.

More recent approaches have extended invariance learning beyond direct feature-distribution alignment. Fishr [29] encourages invariance by matching domain-level gradient variances, thereby promoting similar optimization behavior across source domains, while DomainDrop [30] identifies and suppresses domain-sensitive feature channels. Other approaches impose structural constraints on the representation space; for example, HYPO [31] promotes domain-invariant representations through hyperspherical feature learning with intra-class compactness and inter-class separation. Collectively, these methods demonstrate a shift from explicit distribution matching toward enforcing invariance through optimization dynamics, feature structure, and representation geometry.

2.2 Adversarial Learning

Adversarial learning provides another mechanism for learning domain-invariant representations. In this setting, a feature extractor is trained to support the main classification task while simultaneously confusing a domain discriminator. The discriminator attempts to predict the domain label from the learned representation, whereas the feature extractor is optimized to make the domain labels indistinguishable. This adversarial interaction encourages the latent representation to remove domain-specific cues and retain features that are shared across domains.

Li *et al.* [5] proposed an adversarial autoencoder-based framework that learns domain-invariant representations through an adversarial objective. Later, Li *et al.* [6] introduced a conditional invariant adversarial network that aligns representations conditioned on class labels, thereby preserving discriminative structure while reducing domain discrepancy. Adversarial methods have shown strong potential for DG, particularly when domain distributions differ significantly. However, adversarial training can be unstable and often requires careful balancing between classification and domain-confusion objectives. Moreover, aggressively suppressing domain information may inadvertently remove useful semantic cues if domain and class information are entangled.

More recently, Dayal *et al.* proposed Margin-based Adversarial Learning for Domain Generalization (MADG) [32], which formulates adversarial domain alignment using a margin-based discrepancy measure. Rather than relying solely on conventional domain discrimination, MADG seeks to reduce source-domain discrepancies through a margin-aware adversarial objective and provides a theoretical connection to unseen-domain generalization.

2.3 Data Augmentation

Data augmentation improves DG by increasing the diversity of training data and exposing the model to variations that may occur in unseen domains. Classical augmentation strategies apply label-preserving transformations to input samples, such as color jittering, cropping, flipping, or geometric transformations. DG-specific augmentation methods go further by attempting to synthesize domain-shifted examples that preserve semantic content while altering domain-specific appearance.

Volpi *et al.* [9] generated adversarial perturbations that simulate challenging unseen domains by maximizing the model’s loss within a constrained neighborhood. Shankar *et al.* [33] proposed CrossGrad, which perturbs samples using gradients from a domain classifier to encourage robustness to domain variations. Zhou *et al.* [11] introduced Learning to Augment by Optimal Transport (L2A-OT), which learns a generator to synthesize pseudo-novel domain samples. They later proposed Deep Domain-Adversarial Image Generation (DDAIG) [10], which generates domain-shifted variants of input images using an adversarial generator.

In addition to image-level augmentation, recent DG methods increasingly manipulate intermediate feature representations. Uncertainty Modeling for Out-of-Distribution Generalization (DSU) [34] models feature statistics as uncertain quantities and synthesizes alternative feature statistics during training to improve robustness to potential domain shifts. Exact Feature Distribution Matching (EFDM) [35] further moves beyond manipulating only feature means and variances by matching empirical feature distributions, enabling richer feature-level style transformations. Cross Contrasting Feature Perturbation [12] generates novel latent representations while explicitly seeking to preserve semantic consistency, whereas Randomized Adversarial Style Perturbation (RASP) [4] perturbs style-related feature statistics in adversarial directions to improve robustness to unexpected styles. These approaches demonstrate that feature-space augmentation can expose models to a substantially broader range of domain variations without requiring explicit image generation.

Although augmentation-based methods can enrich the training distribution, their effectiveness depends on whether the generated samples faithfully represent realistic domain shifts without introducing semantic distortions or artifacts.

2.4 Disentangled Embedding Learning

Disentangled embedding learning [16–19] offers a principled way to address domain shift by explicitly separating task-relevant semantic content from domain-specific factors. In the DG setting, the goal is to decompose the latent representation into a domain-invariant component Z_c , which captures class-relevant information, and a domain-specific component Z_d , which captures style or domain-related variations. This separation is motivated by the assumption that the semantic relationship between inputs and labels remains stable across domains, while nuisance factors such as background, texture, illumination, and acquisition conditions vary across environments.

Several DG methods have investigated this separation using different forms of supervision and regularization. Bui *et al.* [36] exploit both domain-invariant and domain-specific representations and encourage their separation through covariance-based constraints. Zhang *et al.* [16] provide a more principled treatment of disentanglement for DG by explicitly modeling latent factors associated with transferable and domain-dependent information. POEM [17] adopts separate domain-agnostic and domain-specific branches and promotes separation between their embeddings using cosine similarity, encouraging the two representations to occupy distinct directions in the latent space.

More recent work has extended this idea beyond purely geometric separation. INSURE [18] introduces an information-theoretic perspective for learning complementary representations and emphasizes that domain-specific information need not be entirely discarded, provided that it is appropriately separated from task-relevant information. DSDRNet [19] further explores dual-stream disentanglement through reconstruction and feature-style manipulation, reinforcing the usefulness of explicitly modeling domain-dependent factors rather than simply suppressing them.

Despite recent advances in disentangled DG, many existing approaches rely on covariance, cosine similarity, orthogonality, or reconstruction-based objectives to separate domain-specific and task-relevant information. Such constraints can effectively reduce geometric overlap or low-order statistical dependence between representations, but they do not necessarily eliminate more complex nonlinear relationships. Consequently, residual domain-related information may remain within the task-relevant representation and contribute to spurious correlations under unseen domain shifts.

FOCUS [20] addresses this issue by regularizing the statistical dependence between domain-agnostic and domain-specific embeddings using mutual information. Rather than simply encouraging geometric separation, the framework seeks to reduce a more general measure of dependence capable of capturing nonlinear relationships between latent variables. The original FOCUS formulation employs the Contrastive Log-ratio Upper Bound (CLUB) [21] as the mutual information regularizer.

Mutual information has also been explored in DG from different perspectives. For example, MIRO [37] employs mutual-information-based regularization to preserve useful information from pretrained representations during domain-generalized learning. In contrast, FOCUS applies mutual information directly between learned task-relevant and domain-specific embeddings with the objective of reducing their dependence. Thus, the role of mutual information in FOCUS is specifically associated with disentangled representation learning rather than representation preservation.

In this extended study, we additionally examine whether this disentanglement mechanism depends strongly on the particular mutual information estimator. Specifically, we compare the original CLUB-based upper-bound formulation with a lower-bound alternative based on MINE [22]. This comparison provides insight into how the mathematical form, numerical scale, and optimization behavior of different mutual information estimators influence representation separation and downstream domain generalization.

3 FOCUS Framework

3.1 Problem Setting

Let $\mathcal{X} \subset \mathbb{R}^D$ denote the input space, $\mathcal{Y} = \{1, \dots, C\}$ denote the label space, and $\mathcal{D} = \{1, \dots, K\}$ denote the set of source domain indices. A domain is characterized by a joint probability distribution $P_{XY}^{(k)} \in \mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$, where $\mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$ denotes the set of all joint distributions over $\mathcal{X} \times \mathcal{Y}$. In the multi-source Domain Generalization (DG) setting, we are given K labeled source domains:

$$S^{(k)} = \left\{ \left(x_i^{(k)}, y_i^{(k)} \right) \right\}_{i=1}^{N^{(k)}}, \quad k \in \mathcal{D}, \quad (1)$$

where each sample is drawn i.i.d. as

$$\left(x_i^{(k)}, y_i^{(k)} \right) \sim P_{XY}^{(k)}. \quad (2)$$

Although all domains share the same label space $\mathcal{Y}$, their marginal input distributions may differ:

$$P_X^{(k)} \neq P_X^{(j)}, \quad k \neq j, \quad (3)$$

The goal of DG is to learn a predictor

$$f : \mathcal{X} \rightarrow \mathcal{Y}, \quad (4)$$

using only the source domains $\{S^{(k)}\}_{k=1}^K$, such that it generalizes well to an unseen target domain τ characterized by a distribution $P_{XY}^{(\tau)}$. During training, no samples from the target domain are available, *i.e.*, $S^{(\tau)} = \emptyset$.

A central premise of DG is that the semantic relationship between inputs and labels remains stable across domains, while the observed input distribution changes due to domain-specific factors such as illumination, background, texture, viewpoint,

and acquisition conditions. Therefore, a representation learning approach to DG seeks to learn a content representation

$$Z_c = E_{c,\theta_c}(X), \quad (5)$$

that is both sufficient for predicting the task label and invariant to domain-specific variations. This principle can be expressed in information-theoretic form as

$$\max_{E_c} I(Z_c; Y) \quad \text{subject to} \quad I(Z_c; D) \approx 0, \quad (6)$$

where $I(\cdot; \cdot)$ denotes mutual information. Maximizing $I(Z_c; Y)$ encourages the representation to retain task-relevant semantic information, while minimizing $I(Z_c; D)$ encourages the representation to discard domain-dependent information.

In practice, directly enforcing this objective is difficult because semantic and domain-specific factors are often entangled in the learned feature space. To address this issue, we explicitly learn two complementary representations: a domain-invariant content embedding Z_c and a domain-specific embedding Z_d . The objective is to make Z_c predictive of the class label Y , make Z_d predictive of the domain label D , and reduce statistical dependence between the two embeddings.

3.2 Dual-Branch Disentangled Representation Learning

The proposed framework employs two encoders: a content encoder $E_{c,\theta_c} : \mathcal{X} \rightarrow \mathbb{R}^m$, parameterized by θ_c , and a domain encoder $E_{d,\theta_d} : \mathcal{X} \rightarrow \mathbb{R}^n$, parameterized by θ_d . Given an input sample X , the two encoders produce

$$Z_c = E_{c,\theta_c}(X), \quad Z_d = E_{d,\theta_d}(X). \quad (7)$$

The content embedding Z_c is intended to capture category-relevant semantic information that generalizes across domains, while the domain embedding Z_d is intended to capture domain-specific factors such as style, background, and acquisition-related variations.

Two classifiers are used to supervise these embeddings. The class classifier $C_{c,\delta_c} : \mathbb{R}^m \rightarrow \mathcal{Y}$, parameterized by δ_c , predicts class labels from Z_c . The domain classifier $C_{d,\delta_d} : \mathbb{R}^n \rightarrow \mathcal{D}$, parameterized by δ_d , predicts domain labels from Z_d . The corresponding cross-entropy losses are defined as

$$\mathcal{L}_c = \mathcal{L}_{\text{CE}}(C_{c,\delta_c}(E_{c,\theta_c}(X)), Y), \quad (8)$$

$$\mathcal{L}_d = \mathcal{L}_{\text{CE}}(C_{d,\delta_d}(E_{d,\theta_d}(X)), D), \quad (9)$$

where $\mathcal{L}_{\text{CE}}$ denotes the multi-class cross-entropy loss.

The domain supervision term encourages Z_d to capture domain-related information, while the class supervision term encourages Z_c to capture task-relevant information. However, these two objectives alone do not guarantee that Z_c and Z_d are statistically independent. Domain-specific cues may still leak into the content representation, especially when semantic and stylistic factors are correlated in the training

domains. Therefore, we introduce a mutual information-based regularization term to reduce dependence between the two embeddings:

$$I(Z_c; Z_d) \rightarrow 0. \quad (10)$$

Unlike covariance or cosine-similarity regularization, mutual information can capture both linear and nonlinear statistical dependencies, making it a stronger criterion for disentangling domain-invariant and domain-specific representations.

3.3 Mutual Information Estimation

The mutual information between Z_c and Z_d is defined as

$$I(Z_c; Z_d) = \iint p(z_c, z_d) \log \left(\frac{p(z_c, z_d)}{p(z_c)p(z_d)} \right) dz_c dz_d, \quad (11)$$

where $p(z_c, z_d)$ denotes the joint density of (Z_c, Z_d) , and $p(z_c)$ and $p(z_d)$ denote the corresponding marginal densities. Direct computation of mutual information is generally intractable for high-dimensional neural representations because the true joint and marginal distributions are unknown.

The original FOCUS framework [20] uses CLUB to obtain an upper-bound estimate of $I(Z_c; Z_d)$ and minimizes this estimate during representation learning. In the present extended study, we additionally consider MINE as a lower-bound estimator in order to investigate whether the effectiveness of mutual information-based disentanglement depends strongly on the estimator used to construct the regularization objective.

These two formulations provide complementary perspectives. CLUB produces an upper-bound estimate and is therefore naturally aligned with an objective that seeks to reduce mutual dependence. MINE instead provides a lower-bound estimate whose quality depends on optimization of a neural critic. Although reducing a lower bound does not provide the same direct interpretation as reducing an upper bound, it can nevertheless provide a useful dependence-sensitive regularization signal in practice. Comparing these alternatives under the same architecture allows us to separate the effect of mutual information-based regularization from the effect of the particular estimator.

3.4 Lower-Bound Mutual Information Estimation with MINE

MINE [22] estimates a lower bound on mutual information using the Donsker–Varadhan representation of the Kullback–Leibler divergence. For two probability distributions P and Q , the KL divergence is given by

$$D_{\text{KL}}(P||Q) = \sup_{T: \Omega \rightarrow \mathbb{R}} \mathbb{E}_P[T] - \log \mathbb{E}_Q[e^T], \quad (12)$$

where the supremum is taken over all functions T for which both expectations remain finite.

The mutual information $I(X; Y)$ between two random variables X and Y can be defined through KL divergence as

$$I(X; Y) = D_{\text{KL}}(p_{XY} \parallel p_X p_Y), \quad (13)$$

where p_{XY} is the joint distribution and p_X and p_Y are the marginal distributions. Following MINE, we use a neural network N_θ to estimate the lower bound of $I(Z_c; Z_d)$:

$$I(Z_c; Z_d) \geq \sup_{\theta \in \Theta} \underbrace{\mathbb{E}_{p_{Z_c Z_d}}[N_\theta(Z_c, Z_d)] - \log \mathbb{E}_{p_{Z_c}} \mathbb{E}_{p_{Z_d}}[\exp(N_\theta(Z_c, Z_d))]}_{\hat{I}_\theta(Z_c, Z_d)}. \quad (14)$$

Using this lower-bound estimate, the MINE-based objective is written as:

$$\begin{aligned} \min_{\theta_c, \theta_d, \delta_c, \delta_d} \max_{\theta \in \Theta} \mathbb{E}_{(X, Y, D) \sim P_{\text{src}}} & \left[\mathcal{L}_c(C_{c, \delta_c}(E_{c, \theta_c}(X)), Y) \right. \\ & + \lambda_1 \mathcal{L}_d(C_{d, \delta_d}(E_{d, \theta_d}(X)), D) \\ & \left. + \lambda_2 \hat{I}_\theta(E_{c, \theta_c}(X); E_{d, \theta_d}(X)) \right]. \end{aligned} \quad (15)$$

Here, $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, and $D \in \mathcal{D}$ denote the input sample, class label, and domain label, respectively. P_{src} denotes the empirical mixture distribution over the available source domains $\{P_{XY}^{(k)}\}_{k \in \mathcal{D}}$. The content encoder E_{c, θ_c} produces the content embedding Z_c , while the domain encoder E_{d, θ_d} produces the domain embedding Z_d . The classifiers C_{c, δ_c} and C_{d, δ_d} predict class and domain labels, respectively. The scalars $\lambda_1, \lambda_2 \geq 0$ balance domain supervision and mutual information regularization.

Optimization is performed alternately. The MINE critic is first updated while the encoders are fixed in order to improve the lower-bound estimate for the current representations. The critic is subsequently held fixed while the encoders and classifiers are optimized using the estimated mutual information as a regularization signal. This alternating procedure allows the dependence estimator and representation model to evolve jointly during training.

Because MINE provides a lower bound, reducing its estimated value should not be interpreted as guaranteeing a corresponding reduction in the exact mutual information when the bound is loose. Nevertheless, the estimator can still supply a useful dependence-sensitive learning signal. We therefore use the MINE formulation in this extended study as an alternative mechanism for MI-based representation separation and empirically compare its behavior with the original CLUB-based formulation.

3.5 Upper-Bound Mutual Information Estimation with CLUB

The original FOCUS framework employs the Contrastive Log-ratio Upper Bound (CLUB) [21] to regularize the dependence between Z_c and Z_d . CLUB expresses an upper bound on mutual information as

$$I(Z_c; Z_d) \leq \mathbb{E}_{p_{Z_c Z_d}}[\log p(Z_d | Z_c)] - \mathbb{E}_{p_{Z_c}} \mathbb{E}_{p_{Z_d}}[\log p(Z_d | Z_c)]. \quad (16)$$

Since the true conditional distribution $p(Z_d | Z_c)$ is unavailable, variational CLUB introduces a variational approximation $q_\psi(Z_d | Z_c)$, parameterized by ψ . The mutual information estimate is then given by

$$I(Z_c; Z_d) \leq \hat{I}_\psi(Z_c; Z_d) = \mathbb{E}_{p_{Z_c Z_d}} [\log q_\psi(Z_d | Z_c)] - \mathbb{E}_{p_{Z_c}} \mathbb{E}_{p_{Z_d}} [\log q_\psi(Z_d | Z_c)]. \quad (17)$$

Given N sample pairs $\{(z_{c,i}, z_{d,i})\}_{i=1}^N$, the unbiased mini-batch estimate is given by

$$\hat{I}_\psi = \frac{1}{N} \sum_{i=1}^N \left[\log q_\psi(z_{d,i} | z_{c,i}) - \frac{1}{N} \sum_{j=1}^N \log q_\psi(z_{d,j} | z_{c,i}) \right]. \quad (18)$$

The first term assigns high likelihood to matched pairs $(z_{c,i}, z_{d,i})$, while the second term penalizes high likelihood assigned to mismatched pairs $(z_{c,i}, z_{d,j})$ within the mini-batch. Minimizing the CLUB upper bound promotes statistical independence between Z_c and Z_d , leading to more robust and disentangled representation learning.

An upper-bound formulation is conceptually attractive for disentanglement because the learning objective is explicitly concerned with suppressing dependence. However, the empirical effectiveness of a particular estimator also depends on its numerical scale and optimization behavior. As demonstrated in the extended experiments, the CLUB estimate can have a substantially larger magnitude than the supervised objectives and therefore requires appropriate weighting during training.

3.6 FOCUS: Feature Orthogonalization via Contrastive Upper-bound Separation

FOCUS, originally introduced in our conference work [20], adopts the CLUB-based upper-bound formulation as its mutual information regularization mechanism. The framework is designed around the principle that domain-specific information need not be completely discarded; instead, it can be explicitly modeled in a separate representation space and statistically decoupled from the representation used for semantic prediction.

FOCUS (Figure 1) employs two dedicated encoders: E_{c,θ_c} to capture category-relevant characteristics Z_c , and E_{d,θ_d} to extract domain-specific characteristics Z_d . The class classifier C_{c,δ_c} predicts class labels from Z_c , while the domain classifier C_{d,δ_d} predicts domain labels from Z_d . The CLUB estimator then regularizes the two latent spaces by minimizing their mutual dependence.

In the general CLUB formulation, the variational conditional distribution $q_\psi(Z_d | Z_c)$ can be modeled as a Gaussian distribution with a learned mean and variance:

$$q_\psi(Z_d | Z_c) = \mathcal{N}(Z_d; \mu_\psi(Z_c), \text{diag}(\sigma_\psi^2(Z_c))), \quad (19)$$

where $\mu_\psi(Z_c)$ denotes the predicted conditional mean of Z_d given Z_c , and $\sigma_\psi^2(Z_c)$ denotes the predicted conditional variance.

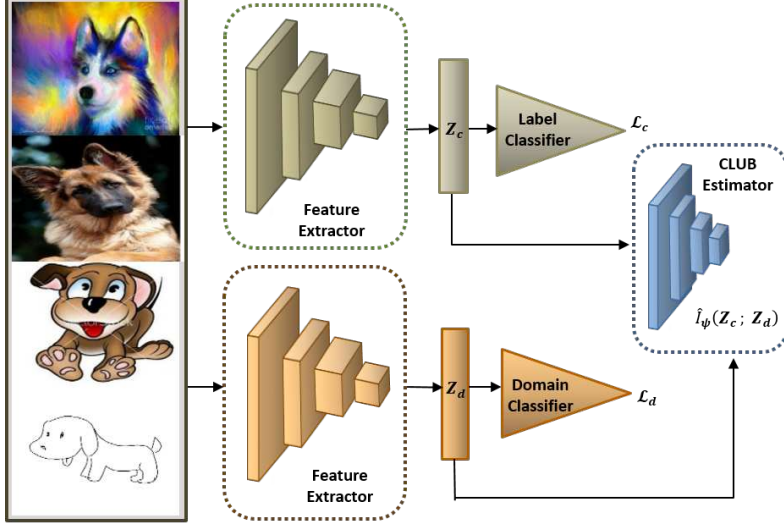


Fig. 1 Overview of the FOCUS framework. FOCUS (Feature Orthogonalization via Contrastive Upper-bound Separation) employs parallel encoders to extract domain-invariant and domain-specific features, and uses the CLUB mutual information upper bound to minimize statistical dependence between the two embedding spaces. This promotes effective disentanglement and enhances generalization to unseen domains.

Using this general Gaussian formulation, the CLUB estimate can be written as:

$$\begin{aligned} \hat{I}_{CLUB} = & -\frac{1}{2} \left\{ \frac{1}{N} \sum_{i=1}^N (z_{d,i} - \mu_{\psi}(z_{c,i}))^T \text{Diag}[\sigma_{\psi}^{-2}(z_{c,i})] (z_{d,i} - \mu_{\psi}(z_{c,i})) \right\} \\ & + \frac{1}{2} \left\{ \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (z_{d,j} - \mu_{\psi}(z_{c,i}))^T \text{Diag}[\sigma_{\psi}^{-2}(z_{c,i})] (z_{d,j} - \mu_{\psi}(z_{c,i})) \right\}. \end{aligned} \quad (20)$$

In practice, we adopt the CLUBMean variant by fixing the covariance to the identity matrix:

$$q_{\psi}(Z_d | Z_c) = \mathcal{N}(Z_d; \mu_{\psi}(Z_c), I). \quad (21)$$

Equivalently, this assumes

$$\sigma_{\psi}^2(Z_c) = \mathbf{1}, \quad \text{Diag}[\sigma_{\psi}^{-2}(Z_c)] = I.$$

With this unit-covariance assumption, the CLUB estimate simplifies to:

$$\hat{I}_{CLUBMean} = -\frac{1}{2} \left\{ \frac{1}{N} \sum_{i=1}^N \|z_{d,i} - \mu_{\psi}(z_{c,i})\|^2 \right\} + \frac{1}{2} \left\{ \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \|z_{d,j} - \mu_{\psi}(z_{c,i})\|^2 \right\}. \quad (22)$$

This choice is motivated by stability considerations. In preliminary experiments, learning the conditional variance introduced unstable mutual-information estimates and, in some cases, produced negative values, which is undesirable because CLUB is intended to act as an upper-bound estimator. Fixing the covariance to the identity matrix avoids unreliable variance estimation and reduces the estimator to a distance-based objective: the first term measures the distance between the predicted mean $\mu_\psi(z_{c,i})$ and the true paired embedding $z_{d,i}$, while the second term measures the distance between the same predicted mean and independently sampled embeddings $z_{d,j}$. Thus, CLUB-Mean still captures dependence through how well Z_d can be predicted from Z_c , while providing a more stable and computationally efficient regularization signal.

The optimization objective function in FOCUS for learning a domain-agnostic feature space is defined as:

$$\min_{\theta_c, \theta_d, \delta_c, \delta_d} \mathbb{E}_{(X,Y,D) \sim P_{\text{src}}} \left[\mathcal{L}_c(C_{c,\delta_c}(E_{c,\theta_c}(X)), Y) \right. \\ \left. + \lambda_1 \mathcal{L}_d(C_{d,\delta_d}(E_{d,\theta_d}(X)), D) \right. \\ \left. + \lambda_2 \mathcal{J}(E_{c,\theta_c}(X), E_{d,\theta_d}(X), \psi^*(E_{c,\theta_c}(X), E_{d,\theta_d}(X))) \right], \quad (23)$$

$$\text{where } \mathcal{J}(Z_c, Z_d, \psi) = \hat{I}_\psi(Z_c; Z_d), \\ \psi^*(Z_c, Z_d) = \arg \max_{\psi} \mathbb{E}_{p_{Z_c Z_d}} [\log q_\psi(Z_d | Z_c)].$$

Here, the first term preserves the class-discriminative capacity of Z_c , the second term encourages Z_d to capture domain-specific information, and the third term minimizes the CLUB upper-bound estimate of mutual information between Z_c and Z_d . This encourages the two embeddings to become statistically independent, reducing the leakage of domain-specific information into the content branch.

FOCUS is optimized using an alternating training strategy. At each iteration, the CLUB estimator parameters are first updated by maximizing the conditional likelihood of matched representation pairs while the encoders are fixed. The CLUB parameters are then held fixed while the encoders and classifiers are optimized using the complete FOCUS objective.

The MINE formulation introduced in this extended study does not redefine FOCUS; rather, it serves as an alternative dependence regularizer used to investigate the role of the mutual information estimator. This distinction allows the original CLUB-based FOCUS design to be evaluated against a lower-bound alternative while keeping the underlying dual-branch architecture and disentanglement objective conceptually consistent.

4 Experimental Setup and Evaluation

This section presents the experimental setup used to evaluate the proposed FOCUS framework for domain generalization. We first describe the benchmark datasets and evaluation protocol, followed by the implementation details, network architecture, and

Algorithm 1 Training procedure for FOCUS

```
1: Input: Source datasets  $\{S^{(k)}\}_{k=1}^K$ ; encoders  $E_{c,\theta_c}, E_{d,\theta_d}$ ; classifiers  $C_{c,\delta_c}, C_{d,\delta_d}$ ;
   Gaussian CLUB model  $q_\psi(Z_d | Z_c) = \mathcal{N}(Z_d; \mu_\psi(Z_c), I)$ ; weights  $\lambda_1, \lambda_2$ ; minibatch
   size  $B$ ; epochs  $T$ ; estimator steps  $n_{\text{est}}$ 
2: Output: Trained  $E_c$  and  $C$ .
3: for  $t = 1$  to  $T$  do
4:   Sample a mini-batch  $\{(x_i^{(k)}, y_i^{(k)}, d_i^{(k)})\}_{i=1}^B$  from the source domains
   // (i) Update CLUB estimator (encoders fixed)
5:    $z_{c,i} \leftarrow E_{c,\theta_c}(x_i^{(k)}), z_{d,i} \leftarrow E_{d,\theta_d}(x_i^{(k)})$ 
6:   for  $s = 1$  to  $n$  do
7:     Log-likelihood  $= \mathcal{L}(\psi) = \frac{1}{B} \sum_{i=1}^B \log q_\psi(z_{d,i} | z_{c,i})$ 
8:     Update  $\psi$  by maximizing  $\mathcal{L}(\psi)$ 
9:   end for
   // (ii) Update encoders/classifiers (CLUB fixed)
10:  Compute  $\hat{I}_\psi(Z_c; Z_d)$  using Eq. (18)
11:  Compute  $\mathcal{L}_{\text{FOCUS}}$  using Eq. (23) with  $\psi$  fixed
12:  Update  $\{\theta_c, \theta_d, \delta_c, \delta_d\}$  by minimizing  $\mathcal{L}_{\text{FOCUS}}$ 
13: end for
14: Return  $E_c$  and  $C$ 
```

quantitative results. The experiments are conducted on four widely used DG benchmarks: PACS, VLCS, Office-Home, and TerraIncognita. These datasets cover a broad range of distribution shifts, including artistic style changes, dataset-specific acquisition differences, object appearance variations, and environmental shifts caused by different camera-trap locations. Together, they provide a rigorous evaluation setting for assessing the robustness and transferability of the learned representations.

Following the standard leave-one-domain-out DG protocol, one domain is selected as the unseen target domain for testing, while the remaining domains are used as source domains for training and validation. No samples from the target domain are used during training. This protocol reflects the multi-source DG setting considered in this work, where the goal is to learn representations that remain discriminative for semantic prediction while being robust to domain-specific variations.

4.1 Datasets

4.1.1 PACS

PACS [23] is a widely used object recognition benchmark for domain generalization. It contains four domains: Photo, Art-Painting, Cartoon, and Sketch. These domains exhibit substantial visual differences in texture, color distribution, abstraction level, and image detail, making PACS particularly useful for evaluating robustness to style-driven domain shifts. The dataset contains 9,991 images across seven object categories: dog, elephant, giraffe, guitar, horse, house, and person. In this work, we follow the official training-validation split commonly used in DG evaluation.

4.1.2 VLCS

VLCS is another standard DG benchmark, consisting of images collected from four different datasets: VOC2007/PASCAL, LabelMe, Caltech, and SUN. It contains 10,729 images across five semantic categories: bird, car, chair, dog, and person. Unlike PACS, which primarily captures artistic and stylistic variation, VLCS reflects dataset-induced shifts caused by differences in scene composition, background clutter, lighting, object scale, and acquisition conditions. These characteristics make VLCS useful for evaluating whether a model can learn representations that remain stable across dataset-specific distribution shifts. We follow the training and validation protocol described in [5].

4.1.3 Office-Home

Office-Home [24] is a challenging domain generalization benchmark designed to evaluate robustness across object appearance and rendering-style variations. It contains 15,500 images from 65 object categories commonly found in office and home environments. The dataset includes four domains: Art, Clipart, Product, and Real-World. The Art domain contains artistic depictions of objects, Clipart consists of abstract illustrations, Product includes high-resolution commercial product images, and Real-World contains natural photographs. Due to its larger number of categories and diverse visual styles, Office-Home presents a more complex recognition setting than PACS and VLCS. We use the domain splits and training-validation protocol adopted in [38].

4.1.4 TerraIncognita

TerraIncognita [26] is a wildlife recognition benchmark designed to evaluate DG under severe environmental and geographic distribution shifts. It contains 24,788 images from 10 animal categories collected using camera traps at four locations: L38, L43, L100, and L46. Each location corresponds to a distinct domain and introduces variations in background, illumination, viewpoint, camera placement, and ecological context. Unlike style-based benchmarks, TerraIncognita contains naturally occurring shifts where domain-specific environmental cues may become spuriously correlated with class labels. This makes it particularly challenging for evaluating whether a method can suppress domain-dependent information while preserving semantic discriminability. Following the protocol in [37], we use 80% of each source domain for training and reserve the remaining 20% for validation.

4.2 Implementation Details

We follow the standard multi-source DG evaluation protocol, where one domain is held out as the target domain and the remaining domains are used as source domains. From the source domains, the training split is used to optimize the model, while the validation split is used to select the best-performing configuration. The held-out target domain is used only for final testing.

All experiments are conducted using mini-batch stochastic gradient descent (SGD). We use a batch size of 32 for PACS, VLCS, Office-Home, and TerraIncognita. The

training process is carried out for 50 epochs with a weight decay of 5×10^{-4} and an initial learning rate of 0.001. The learning rate is reduced by a factor of 0.1 after epoch 40. Standard image augmentation techniques are applied during training, including color jittering, random horizontal flipping, and random resized cropping. All input images are resized to 224×224 to ensure consistency across datasets.

Based on our empirical analysis, for the MINE-based lower-bound formulation, we set $(\lambda_1 = 1, 0.5, 1, 0.5)$ and $(\lambda_2 = 1, 1, 1, 0.5)$ for PACS, VLCS, Office-Home, and TerraIncognita, respectively. For the CLUB-based FOCUS formulation, we use $(\lambda_1 = 1, 1, 1, 0.5)$ for the same four benchmarks, while (λ_2) is fixed at (0.01) across all datasets. In particular, our experiments showed that the CLUB-based upper-bound estimate operates on a substantially larger numerical scale and can dominate the supervised objectives if assigned a larger weight, whereas the MINE-based lower-bound estimate remains more comparable in magnitude to the classification losses and therefore supports larger regularization coefficients. Both estimators are updated over 5 steps per iteration, providing a stable mutual information regularization signal without over-optimizing the estimators for each mini-batch.

4.3 Network Architecture

4.3.1 Feature Extractors

FOCUS uses a dual-encoder architecture to separately model domain-invariant and domain-specific information. Both encoders are implemented using ResNet-50 [39] pretrained on ImageNet. The use of a pretrained backbone provides strong initial feature representations and stabilizes training across all datasets.

The content encoder $E_{c,\theta_c} : \mathcal{X} \rightarrow \mathbb{R}^{2048}$ maps an input image X to a domain-invariant embedding:

$$Z_c = E_{c,\theta_c}(X). \quad (24)$$

This embedding is optimized to retain semantic information relevant to the classification task while suppressing domain-dependent variations such as texture, style, illumination, and background.

The domain encoder $E_{d,\theta_d} : \mathcal{X} \rightarrow \mathbb{R}^{2048}$ uses the same ResNet-50 architecture and maps the input image to a domain-specific embedding:

$$Z_d = E_{d,\theta_d}(X). \quad (25)$$

This branch is supervised using domain labels and is encouraged to capture domain-related factors, including stylistic attributes, low-level appearance patterns, and imaging conditions. The use of two separate encoders enables FOCUS to explicitly disentangle task-relevant semantic information from domain-dependent nuisance variations.

4.3.2 Predictors

FOCUS employs two prediction heads to supervise the disentangled latent spaces generated by the dual encoders. The class predictor C_{c,δ_c} is implemented as a single

fully connected layer that maps the 2048-dimensional content embedding Z_c to class logits:

$$C_{c,\delta_c} : \mathbb{R}^{2048} \rightarrow \mathbb{R}^{|\mathcal{Y}|}.$$

The output dimension corresponds to the number of object categories in each dataset.

The domain predictor C_{d,δ_d} is implemented as a multi-layer perceptron that maps the 2048-dimensional domain-specific embedding Z_d to domain logits:

$$C_{d,\delta_d} : \mathbb{R}^{2048} \rightarrow \mathbb{R}^K,$$

where K denotes the number of source domains. The MLP consists of an input layer of size 2048, two hidden layers of size 2048 each, and an output layer of size K . Each hidden layer is followed by a ReLU activation:

$$\text{ReLU}(x) = \max(0, x). \quad (26)$$

Training this predictor with domain labels encourages E_{d,θ_d} to emphasize domain-specific variations such as style, texture, illumination, and background statistics. Together, the class and domain predictors provide complementary supervision signals that guide the formation of disentangled latent spaces.

4.3.3 MINE Estimator

For the lower-bound mutual information formulation, the MINE critic N_θ is implemented as a two-layer multilayer perceptron (MLP). The critic takes the concatenated representations (Z_c, Z_d) as input and maps them to a 512-dimensional hidden representation followed by a ReLU activation. A final fully connected layer produces a scalar critic score, which is used to estimate the Donsker–Varadhan lower bound on mutual information.

4.3.4 CLUB Estimator

To estimate and minimize the dependence between the domain-invariant and domain-specific feature spaces, FOCUS uses the CLUBMean estimator [21]. The estimator models the conditional distribution $q_\psi(Z_d | Z_c)$ as a Gaussian with fixed identity covariance:

$$q_\psi(Z_d | Z_c) = \mathcal{N}(Z_d; \mu_\psi(Z_c), I), \quad (27)$$

where $\mu_\psi(Z_c)$ is predicted by a neural network and I denotes the identity covariance matrix.

The neural network used to model $\mu_\psi(Z_c)$ consists of an input layer, a fully connected hidden layer with 512 units, and an output layer matching the dimensionality of Z_d . A ReLU activation follows the hidden layer:

$$h = \text{ReLU}(W_1 Z_c + b_1), \quad h \in \mathbb{R}^{512}, \quad (28)$$

$$\mu_\psi(Z_c) = W_2 h + b_2. \quad (29)$$

The CLUB upper-bound estimate is minimized during training to reduce statistical dependence between Z_c and Z_d , thereby promoting stronger disentanglement between semantic and domain-specific representations.

4.4 Results and Discussion

We compare FOCUS against Empirical Risk Minimization (ERM) [40] and several state-of-the-art DG methods across PACS, VLCS, Office-Home, and TerraIncognita. ERM serves as the baseline and trains a single model across all source domains by minimizing the average classification loss. All reported results follow the leave-one-domain-out evaluation protocol, where each domain is used once as the unseen target domain.

The results demonstrate that FOCUS consistently improves average accuracy across all four DG benchmarks. On PACS, FOCUS achieves an average accuracy of 88.99%, outperforming the strongest competing method by +1.26%. On VLCS, FOCUS reaches 80.51%, providing a +0.71% improvement over the strongest reported baseline. On Office-Home, which contains 65 object categories and substantial stylistic diversity, FOCUS achieves 71.60%, improving over the best competing method by +0.70%. On TerraIncognita, which presents severe environmental and geographic shifts, FOCUS achieves 53.10%, yielding a notable +2.70% gain over the strongest prior method.

These improvements indicate that explicitly separating domain-invariant and domain-specific representations, while minimizing their mutual dependence using the CLUB upper bound, improves robustness to unseen domain shifts. The gains are especially pronounced on TerraIncognita, suggesting that FOCUS is particularly effective when spurious correlations between domain-specific environmental cues and class labels are strong. Overall, the results confirm that upper-bound mutual information regularization provides an effective mechanism for learning transferable and domain-agnostic embeddings.

Figures 2–4 present t-SNE [41] visualizations of the embeddings learned by the FOCUS framework on the VLCS benchmark. The first visualization illustrates class separation, where all samples are colored by semantic class label to assess how well the learned embeddings preserve category-level structure. The second visualization shows domain alignment, where the same embeddings are colored by domain to evaluate the degree of source–target mixing in the latent space. The third visualization provides a class-conditional domain-alignment view for a selected class, where samples belonging to a single semantic category are colored by domain. This finer-grained plot is particularly useful for examining whether samples of the same class but from different domains are mapped into a shared feature region, thereby offering more direct evidence of class-wise domain invariance.

Overall, the visualizations show that FOCUS learn embeddings that are semantically structured and substantially aligned across domains. In the class-separation plot, samples within the same category tend to form relatively compact and distinct clusters, indicating that the framework preserves meaningful semantic information in its learned representations. In the domain-alignment plot, samples from different source domains and the target domain exhibit considerable overlap, suggesting that

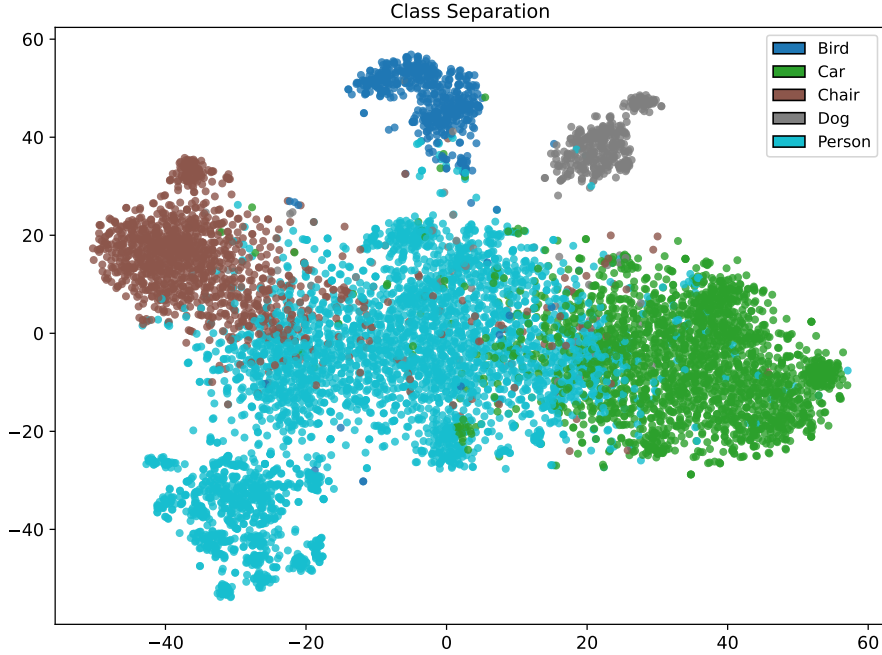


Fig. 2 t-SNE visualization of embeddings learned by FOCUS on the VLCS benchmark, with points colored by class label. The plot illustrates the semantic separability of the learned representation, showing that samples from the same category are mapped into relatively compact and distinguishable regions.

the learned embeddings suppress a substantial portion of domain-specific variation. These visual results are consistent with the quantitative findings and support the conclusion that FOCUS improves domain generalization by learning embeddings that are both semantically discriminative and more domain-agnostic.

The main benchmark results evaluate the original CLUB-based FOCUS formulation. The extended estimator analysis presented in Section 4.5 further examines whether these gains depend strongly on using an upper-bound estimator specifically. As discussed there, the lower-bound MINE formulation also achieves substantial improvement over ERM, suggesting that an important component of the observed benefit arises from mutual information-based dependence reduction itself rather than exclusively from the mathematical form of the CLUB estimator.

4.5 Ablation Study

To better understand the factors that contribute to the framework’s performance, we conduct ablation studies on the PACS benchmark. The analysis examines two related questions. First, we evaluate the contributions of domain-specific supervision and mutual information regularization to the learned representations. Second, as an extension of our previous conference study, we investigate whether the effectiveness of mutual information-based disentanglement depends strongly on the estimator used to construct the regularization objective.

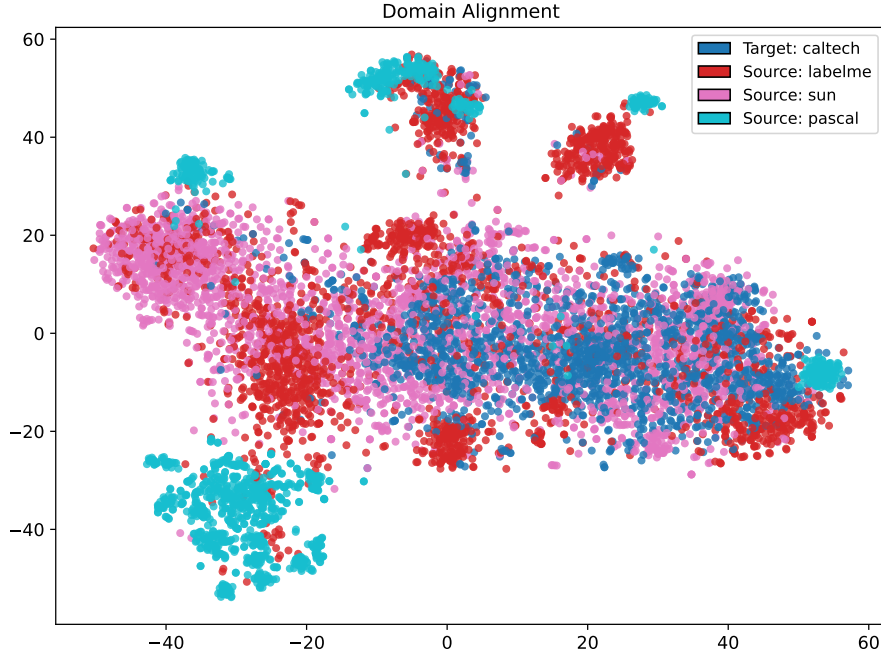


Fig. 3 t-SNE visualization of embeddings learned by FOCUS on the VLCS benchmark, with points colored by domain. The substantial overlap among source and target domains suggests that the learned features reduce domain-specific variation while preserving semantic structure.

Table 1 Leave-one-domain-out results on PACS. The best and second-best results are bolded and underlined, respectively.

Methods	Art	Cartoon	Photo	Sketch	Avg.
ResNet-50					
ERM	86.18	78.58	97.24	73.55	83.88
SagNet (CVPR 21)	87.40	80.70	97.10	80.00	86.30
MatchDG (ICML 21)	85.61	82.12	<u>97.94</u>	78.76	86.11
mDSOI (NeurIPS 21)	<u>88.10</u>	81.10	98.40	79.60	86.80
MIRO (ECCV 22)	-	-	-	-	85.04
POEM (AAAI 23)	-	-	-	-	86.90
CCFP (ICCV 23)	-	-	-	-	86.60
CMCL (TNNLS 25)	87.57	<u>83.60</u>	96.03	83.73	<u>87.73</u>
FOCUS (Ours)	90.03	84.68	97.90	<u>83.35</u>	88.99

For the second analysis, we compare the original CLUB-based FOCUS formulation with an alternative lower-bound formulation based on MINE. The MINE critic is implemented as a two-layer multilayer perceptron operating on concatenated pairs (Z_c, Z_d) , whereas FOCUS employs CLUBMean with a learned conditional mean and fixed unit covariance. Both estimators are integrated into the same dual-branch representation-learning setting, enabling us to examine differences in generalization performance and optimization behavior.

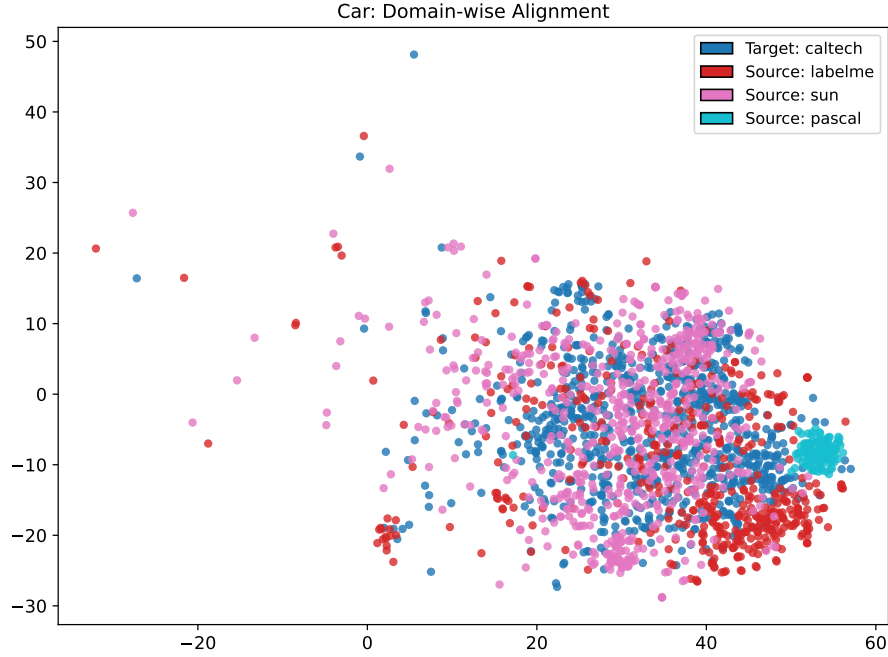


Fig. 4 Class-conditional t-SNE visualization of embeddings learned by FOCUS for a selected class on the VLCS benchmark, with points colored by domain. This figure highlights class-wise domain alignment and provides a more direct view of the extent to which embeddings of the same semantic category are shared across domains.

Table 2 Leave-one-domain-out results on VLCS. The best and second-best results are bolded and underlined, respectively.

Methods	Caltech	LabelMe	Pascal	Sun	Avg.
ResNet-50					
ERM	98.58	64.53	75.23	71.51	77.46
SagNet (CVPR 21)	97.90	64.50	77.50	71.40	77.80
mDSDI (NeurIPS 21)	97.60	<u>66.40</u>	<u>77.80</u>	<u>74.00</u>	79.00
MIRO (ECCV 22)	-	-	-	-	79.00
POEM (AAAI 23)	-	-	-	-	<u>79.80</u>
CCFP (ICCV 23)	-	-	-	-	79.20
INSURE (TIP 24)	<u>98.80</u>	63.80	81.04	72.20	79.01
FOCUS (Ours)	98.86	69.20	77.60	76.38	80.51

4.5.1 Impact of Domain-Specific Supervision and Mutual Information Regularization

The coefficients λ_1 and λ_2 control the contributions of domain supervision and mutual information regularization, respectively. Domain supervision encourages (Z_d)

Table 3 Leave-one-domain-out results on Office-Home. The best and second-best results are bolded and underlined, respectively.

Methods	Art	Clipart	Product	Real	Avg.
ResNet-50					
ERM	67.30	55.80	<u>78.98</u>	79.06	70.28
SagNet (CVPR 21)	63.40	54.80	<u>75.80</u>	78.30	68.10
mDSDI (NeurIPS 21)	68.40	52.50	76.20	80.60	69.42
IIB (AAAI 22)	-	-	-	-	68.60
MIRO (ECCV 22)	-	-	-	-	<u>70.90</u>
POEM (AAAI 23)	-	-	-	-	68.20
CCFP (ICCV 23)	-	-	-	-	68.90
CMCL (TNNLS 25)	67.22	<u>57.88</u>	78.47	79.79	70.84
FOCUS (Ours)	<u>68.25</u>	59.33	79.59	<u>80.53</u>	71.60

Table 4 Leave-one-domain-out results on TerraIncognita. The best and second-best results are bolded and underlined, respectively.

Methods	L100	L38	L43	L46	Avg.
ResNet-50					
ERM	55.93	44.56	54.02	38.66	48.29
SagNet (CVPR 21)	53.00	43.00	57.90	<u>40.40</u>	48.60
mDSDI (NeurIPS 21)	53.20	43.30	56.70	39.20	48.10
IIB (AAAI 22)	-	-	-	-	47.20
MIRO (ECCV 22)	-	-	-	-	<u>50.40</u>
POEM (AAAI 23)	-	-	-	-	50.10
CCFP (ICCV 23)	-	-	-	-	49.00
FOCUS (Ours)	60.48	51.21	<u>56.81</u>	43.93	53.10

to encode information that is predictive of the source domain, while the mutual information term discourages statistical dependence between (Z_d) and the task-relevant representation (Z_c).

For the MINE-based formulation, the best configuration uses ($\lambda_1=1$) and ($\lambda_2=1$), achieving an average PACS accuracy of (88.11%), compared with (83.88%) for ERM. Importantly, setting ($\lambda_1=0$) while retaining the mutual information regularizer results in (87.98%), whereas jointly using domain supervision and dependence regularization increases performance to (88.11%). These results indicate that both explicit modeling of domain-related factors and separation between the two representation spaces contribute to improved DG performance.

For the CLUB-based FOCUS formulation, the best configuration uses ($\lambda_1=1$) and ($\lambda_2=0.01$), achieving an average accuracy of (88.99%). Removing domain supervision while retaining CLUB regularization yields (88.12%), whereas including domain supervision increases performance to (88.99%). These results again support the complementary roles of domain-specific modeling and mutual information-based dependence reduction.

A notable difference between the two estimators is the appropriate scale of (λ_2). MINE performs well with ($\lambda_2=1$), whereas CLUB requires a substantially smaller

Table 5 Impact of domain-specific supervision and mutual information regularization using the MINE-based lower-bound formulation on PACS.

Methods	Art	Cartoon	Photo	Sketch	Avg.
ResNet-50					
ERM	86.18	78.58	97.24	73.55	83.88
$(\lambda_1=0.5, \lambda_2=0.5)$	87.69	83.48	97.42	82.10	87.67
$(\lambda_1=0.5, \lambda_2=1)$	86.86	83.19	97.00	83.66	87.68
$(\lambda_1=0, \lambda_2=1)$	88.42	82.84	97.84	82.82	87.98
$(\lambda_1=1, \lambda_2=1)$	88.05	84.29	97.83	83.05	88.11

Table 6 Impact of domain-specific supervision and mutual information regularization using the CLUB-based upper-bound formulation in FOCUS on PACS.

Methods	Art	Cartoon	Photo	Sketch	Avg.
ResNet-50					
ERM	86.18	78.58	97.24	73.55	83.88
$(\lambda_1=1, \lambda_2=0.1)$	85.25	83.40	97.36	82.69	87.17
$(\lambda_1=0, \lambda_2=0.01)$	88.57	83.06	97.78	83.07	88.12
$(\lambda_1=0.5, \lambda_2=0.01)$	88.72	83.44	97.60	83.66	88.35
$(\lambda_1=1, \lambda_2=0.01)$	90.03	84.68	97.90	83.35	88.99

value of (0.01). Increasing the CLUB weight to (0.1) reduces average accuracy to (87.17%). This behavior indicates that the raw magnitudes of the two MI objectives differ considerably and that the regularization coefficient should not be interpreted or transferred directly across estimators.

4.5.2 Comparison of Lower-Bound and Upper-Bound Mutual Information Regularization

We next directly compare the lower-bound MINE formulation with the upper-bound CLUB formulation used in FOCUS. Both approaches attempt to reduce statistical dependence between the domain-agnostic representation (Z_c) and the domain-specific representation (Z_d), but they construct the corresponding dependence signal differently.

The MINE-based formulation achieves an average PACS accuracy of (88.11%), while the CLUB-based formulation achieves (88.99%). Thus, both formulations substantially outperform the ERM baseline of (83.88%), while the difference between the two MI estimators is comparatively modest. These results indicate that the generalization benefit of the proposed disentanglement strategy is not restricted to a single mutual information estimator. Instead, both upper- and lower-bound objectives can provide useful signals for separating domain-specific and task-relevant representations.

The two formulations nevertheless differ in their theoretical interpretation. MINE provides a lower bound on mutual information, and therefore minimizing its estimated

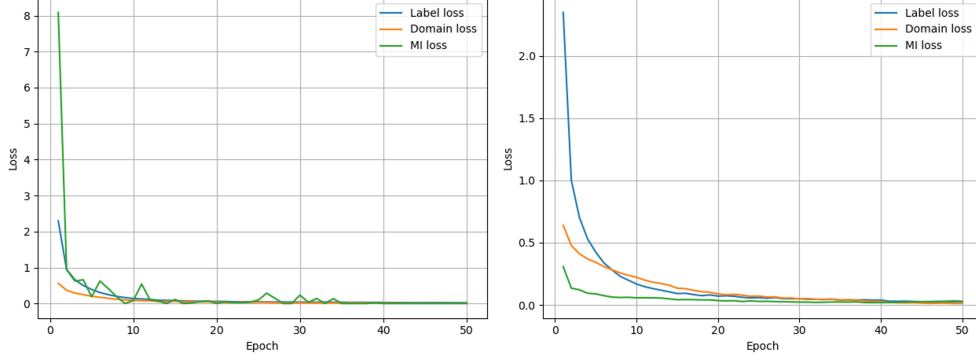


Fig. 5 Comparison of training dynamics for MINE- and CLUB-based mutual information regularization on Office-Home, with Clipart, Product, and Real-World used as source domains. The MINE-based formulation produces an MI regularization term whose magnitude remains relatively comparable to the supervised losses, whereas the CLUB-based formulation operates on a larger numerical scale and requires stronger down-weighting to maintain balanced optimization. Despite these differences in loss scale and training dynamics, the two formulations achieve broadly comparable generalization performance, indicating that estimator choice primarily affects the optimization characteristics of the dependence regularizer rather than producing a substantial difference in downstream accuracy.

value does not necessarily imply a corresponding decrease in the exact mutual information when the bound is loose. CLUB provides an upper-bound formulation and is consequently more directly aligned with the objective of limiting dependence. This theoretical distinction motivates the original FOCUS design, but the empirical results show that it does not translate into a large difference in DG accuracy in the evaluated setting.

A more pronounced difference appears in the optimization behavior of the two objectives. As illustrated in Figure 5, the MINE-based regularization term has a magnitude that is relatively comparable to the supervised objectives and can be used with ($\lambda_2=1$). The CLUB term operates on a substantially larger numerical scale and requires stronger down-weighting, with the best performance obtained using ($\lambda_2=0.01$). Therefore, estimator choice affects the scale and balancing of the multi-objective optimization problem even when the final generalization performance remains similar.

Taken together, these findings suggest that mutual information-based dependence regularization itself is a major contributor to the improvement over ERM, whereas the specific choice between MINE and CLUB has a smaller effect on downstream accuracy. CLUB retains a modest performance advantage in our PACS experiments and remains the estimator used in FOCUS, while MINE demonstrates that comparable improvements can also be obtained using a lower-bound formulation. The extended analysis therefore provides evidence that the proposed disentanglement principle is relatively robust to the particular MI estimator used to implement it.

Overall, the ablation studies demonstrate that both explicit domain supervision and mutual information-based dependence regularization contribute to the effectiveness of the disentangled representation-learning framework. Domain supervision

encourages Z_d to represent domain-related factors, while the MI objective reduces statistical dependence between Z_d and the task-relevant representation Z_c . The extended estimator comparison further shows that this benefit is not exclusive to the CLUB formulation: MINE and CLUB both yield substantial gains over ERM and achieve broadly comparable performance. CLUB provides a modest accuracy advantage and offers an upper-bound formulation that is conceptually aligned with dependence minimization, whereas MINE exhibits a more naturally balanced objective scale. FOCUS therefore retains the original CLUB-based formulation while the comparative analysis demonstrates the broader effectiveness of mutual information-based disentanglement.

5 Conclusion

This work investigated mutual information-based disentangled representation learning for Domain Generalization. Building upon our previously introduced Feature Orthogonalization via Contrastive Upper-bound Separation (FOCUS) framework, the study considers a dual-branch architecture in which task-relevant and domain-specific information are explicitly modeled in separate latent representations. Class supervision encourages the content representation Z_c to preserve semantic information, domain supervision encourages Z_d to capture domain-related characteristics, and mutual information regularization reduces statistical dependence between the two representation spaces.

The original FOCUS formulation uses the CLUB upper-bound estimator as its dependence regularizer. Extensive evaluations on PACS, VLCS, Office-Home, and TerraIncognita demonstrate that this formulation provides strong performance across diverse domain shifts, supporting the effectiveness of explicitly modeling domain-specific information while reducing its dependence on task-relevant representations.

The present study extends the original work by additionally investigating whether the observed benefit depends strongly on the choice of mutual information estimator. To this end, we introduced a lower-bound formulation based on MINE and compared it with the original CLUB-based approach. The results show that both formulations substantially outperform ERM and achieve broadly comparable generalization performance. On PACS, MINE achieves an average accuracy of (88.11%), whereas the CLUB-based FOCUS formulation achieves (88.99%). Thus, although CLUB provides a modest performance advantage, the relatively small difference between the two estimators suggests that the broader benefit lies in the mutual information-based separation of domain-specific and task-relevant representations rather than with a particular estimator.

The extended analysis also reveals important differences in optimization behavior. MINE produces a mutual-information signal whose magnitude is relatively well balanced with the supervised objectives, whereas CLUB requires substantially stronger down-weighting to prevent the MI term from dominating training. These findings highlight that different mutual information estimators may lead to similar generalization performance while exhibiting different numerical scales, hyperparameter requirements, and optimization characteristics.

Overall, the results suggest that explicitly modeling domain-specific factors and reducing their statistical dependence on task-relevant representations are effective strategies for improving robustness to domain shift. They further demonstrate that mutual information-based disentanglement is relatively robust to the choice of estimator, while emphasizing the importance of appropriate objective scaling when integrating neural mutual information estimators into domain-generalized representation learning.

5.1 Limitations

Despite the effectiveness of the proposed framework, several limitations remain. First, FOCUS uses the CLUBMean formulation, in which the conditional distribution ($q_\psi(Z_d | Z_c)$) is modeled with a learned mean and a fixed identity covariance. Although this choice improves optimization stability and avoids unreliable conditional-variance estimates, it limits the expressiveness of the conditional model and may not fully capture complex variance-dependent or multimodal relationships among the learned representations.

Second, mutual information-based regularization is sensitive to the estimator’s numerical scale. In particular, the CLUB objective can have a substantially larger magnitude than the supervised losses and therefore requires careful weighting through (λ_2). The comparison with MINE further demonstrates that suitable regularization coefficients can differ substantially across MI estimators. Developing adaptive balancing mechanisms that automatically normalize or weight the dependence objective relative to the supervised losses represents an important direction for future work.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve experiments with human participants or animals.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used in this study are publicly available benchmark datasets, including PACS, VLCS, Office-Home, and TerraIncognita, and can be obtained from their respective original sources. No new datasets were generated as part of this study.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors declare that no specific funding was received for this study.

Authors' contributions

N.M. conceived the main idea, developed the methodology, carried out the experiments, analyzed the results, and prepared the initial draft of the manuscript. K.B. contributed to refining and finalizing the research concept, provided supervision and technical guidance, reviewed the manuscript, and provided critical feedback and final suggestions that improved the presentation and overall quality of the work. Both authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

- [1] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. *Machine learning* **79**, 151–175 (2010)
- [2] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: *Dataset Shift in Machine Learning*. Mit Press, ??? (2022)
- [3] Blanchard, G., Lee, G., Scott, C.: Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems* **24** (2011)
- [4] Kim, T., Han, B.: Randomized adversarial style perturbations for domain generalization. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2306–2314 (2024). IEEE
- [5] Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409 (2018)
- [6] Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639 (2018)
- [7] Sinha, A., Namkoong, H., Volpi, R., Duchi, J.: Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571* (2017)
- [8] Mehmood, N., Barner, K.: Augmentation, mixing, and consistency regularization for domain generalization. In: 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), pp. 1–6 (2024). IEEE

- [9] Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems* **31** (2018)
- [10] Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13025–13032 (2020)
- [11] Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 561–578 (2020). Springer
- [12] Li, C., Zhang, D., Huang, W., Zhang, J.: Cross contrasting feature perturbation for domain generalization. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1327–1337 (2023). IEEE
- [13] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(4), 4396–4415 (2022)
- [14] Lu, W., Wang, J., Li, H., Chen, Y., Xie, X.: Domain-invariant feature exploration for domain generalization. *Transactions on Machine Learning Research (TMLR)* (2022)
- [15] Vapnik, V.: *Statistical learning theory*, new york: Wiley (1998)
- [16] Zhang, H., Zhang, Y.-F., Liu, W., Weller, A., Schölkopf, B., Xing, E.P.: Towards principled disentanglement for domain generalization. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8014–8024 (2022). IEEE
- [17] Jo, S.-Y., Yoon, S.W.: Poem: polarization of embeddings for domain-invariant representations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 8150–8158 (2023)
- [18] Yu, X., Tseng, H.-H., Yoo, S., Ling, H., Lin, Y.: Insure: an information theory inspired disentanglement and purification model for domain generalization. *IEEE Transactions on Image Processing* **33**, 3508–3519 (2024)
- [19] Yang, J., Li, Q., Xie, S., Yu, W., Li, S.: Dsdnet: Disentangling representation and reconstruct network for domain generalization. In: *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2024). IEEE
- [20] Mehmood, N., Barner, K.: Feature orthogonalization via contrastive upper-bound separation for domain generalization. In: *2025 International Conference on Machine Learning and Applications (ICMLA)*, pp. 123–129 (2025). IEEE

- [21] Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., Carin, L.: Club: A contrastive log-ratio upper bound of mutual information. In: International Conference on Machine Learning, pp. 1779–1788 (2020). PMLR
- [22] Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: International Conference on Machine Learning, pp. 531–540 (2018). PMLR
- [23] Li, D., Yang, Y., Song, Y.-Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5542–5550 (2017)
- [24] Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5018–5027 (2017)
- [25] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011, pp. 1521–1528 (2011). IEEE
- [26] Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 456–473 (2018)
- [27] Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International Conference on Machine Learning, pp. 10–18 (2013). PMLR
- [28] Ghifary, M., Balduzzi, D., Kleijn, W.B., Zhang, M.: Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence* **39**(7), 1414–1430 (2016)
- [29] Rame, A., Dancette, C., Cord, M.: Fishr: Invariant gradient variances for out-of-distribution generalization. In: International Conference on Machine Learning, pp. 18347–18377 (2022). PMLR
- [30] Guo, J., Qi, L., Shi, Y.: Domaindrop: Suppressing domain-sensitive channels for domain generalization. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 19057–19067 (2023). IEEE
- [31] Bai, H., Ming, Y., Katz-Samuels, J., Li, Y.: Hypo: Hyperspherical out-of-distribution generalization. In: International Conference on Learning Representations, vol. 2024, pp. 49188–49214 (2024)
- [32] Dayal, A., KB, V., Cenkeramaddi, L.R., Mohan, C., Kumar, A., N Balasubramanian, V.: Madg: margin-based adversarial learning for domain generalization. *Advances in Neural Information Processing Systems* **36**, 58938–58952 (2023)

- [33] Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745 (2018)
- [34] Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., Duan, L.-Y.: Uncertainty modeling for out-of-distribution generalization. arXiv preprint arXiv:2202.03958 (2022)
- [35] Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8025–8035 (2022). IEEE
- [36] Bui, M.-H., Tran, T., Tran, A., Phung, D.: Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems* **34**, 21189–21201 (2021)
- [37] Cha, J., Lee, K., Park, S., Chun, S.: Domain generalization by mutual-information regularization with pre-trained models. In: European Conference on Computer Vision, pp. 440–457 (2022). Springer
- [38] Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14383–14392 (2021)
- [39] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
- [40] Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, ??? (2013)
- [41] Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)