How Can Cutout Be Leveraged to Improve Out-of-Distribution Generalization?

Kuo Shi
Independent Researcher
kuo.shi@foxmail.com

Abstract

Cutout, a widely adopted data augmentation that masks random square regions of training images, has shown remarkable success in standard supervised learning. Yet its potential for out-of-distribution (OOD) generalization remains largely untapped: we observe that simply combining Cutout with existing OOD methods yields only marginal benefits, and on several PACS and Office-Home domains even degrades performance. Motivated by this counterintuitive finding, we seek a principled way to make OOD generalization benefit from Cutout. We first reinterpret the state-of-the-art hyperspherical prototype learning as a mutual information (MI) maximization framework. Within this framework, naively applying Cutout reduces the objective to estimating MI from a single masked subview, leaving much of Cutout's potential untapped. To address this, we apply the MI chain rule to decompose the original objective into two complementary terms: one encourages the Cutout subview to align closely with the label, while the other drives the full image to recover the residual information unavailable in that subview. We further simplify hyperspherical learning by fixing class prototypes to a regular simplex in closed form, eliminating prototype updates and an extra hyperparameter. Extensive experiments across a broad range of OOD benchmarks demonstrate that our method outperforms competitive baselines and achieves superior performance.

1 Introduction

In recent years, out-of-distribution (OOD) generalization has received much attention [Zhou et al., 2023]. As shown in Fig. 1, the goal of OOD generalization is to learn a model from one or several different but related domains (i.e., diverse training datasets) that will generalize well on unseen testing domains. For instance, given a training set consisting of images coming from paintings, cartoon images and sketches, OOD generalization requires training a good model that has minimum prediction error in classifying images coming from natural images or photos, which clearly has distinguished distributions from the images in the training set [Wang et al., 2023a].

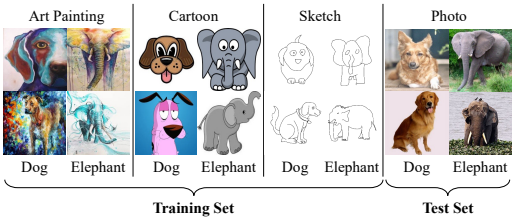


Figure 1: Examples from PACS [Li et al., 2017]. The model is trained on three source domains and evaluated on the held-out unseen target domain.

OOD generalization is critical in real-world applications such as medical diagnosis [Yoon et al., 2024], autonomous driving [Wang et al., 2024b, Zhu and Peng, 2022], and natural language processing [Yang et al., 2023], where distributional shifts inevitably arise and necessitate models that learn domain-invariant representations for robust generalization beyond the training domains [Shen et al., 2021].

A variety of algorithms have been proposed to address OOD generalization, broadly including feature alignment methods aimed at learning domain-invariant representations [Arjovsky et al., 2019,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

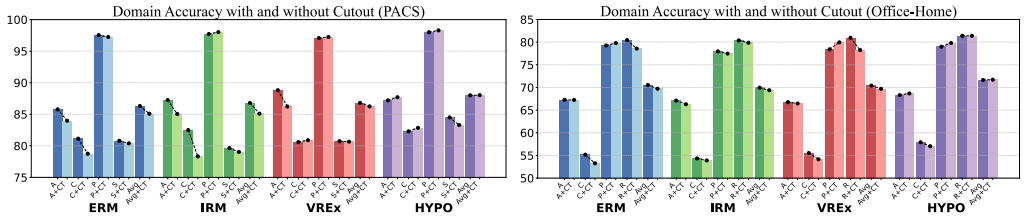


Figure 2: Domain accuracy with and without Cutout on PACS (left) and Office-Home (right). We compare four representative OOD baselines (ERM, IRM, VREx, HYPO) with each domain held out in turn for OOD evaluation; "X+CT" denotes that the model is trained on the remaining domains with Cutout augmentation and evaluated on domain X. Naively adding Cutout yields only marginal or even negative gains, motivating a more principled integration.

Krueger et al., 2021, Wald et al., 2021, Li et al., 2021b], particularly hyperspherical learning guided by intra-class variation and inter-class separation principles [Ye et al., 2021, Bai et al., 2024], and data augmentation approaches [Wang et al., 2024a, Zhou et al., 2024, Zhu et al., 2021], ranging from image generation [Zhou et al., 2020a,b] and transformation [Xu et al., 2021, Wang et al., 2022] to feature-level augmentation [Yoo et al., 2019, Zhou et al., 2021].

Cutout [Devries and Taylor, 2017], a widely recognized data augmentation technique, improves model generalization by randomly masking out square regions of inputs during training and has been highly successful in standard supervised learning [Cubuk et al., 2018, 2020]. Given that OOD generalization fundamentally requires models to remain robust under partial or shifted observations, one might naturally expect Cutout to be a strong ally. Surprisingly, this is not the case. As illustrated in Fig. 2, on both PACS [Li et al., 2017] and Office-Home [Venkateswara et al., 2017] datasets, four representative OOD baselines (ERM [Vapnik, 1999], IRM [Arjovsky et al., 2019], VREx [Krueger et al., 2021], and HYPO [Bai et al., 2024]) see negligible or even negative gains when Cutout is naively added. This counterintuitive observation suggests that Cutout's potential is misaligned with how current methods exploit augmented samples, motivating us to seek a principled integration that makes OOD generalization truly benefit from Cutout.

We approach this problem from an information-theoretic perspective. We first show that the state-of-the-art hyperspherical prototype learning method (HYPO) can be reinterpreted as maximizing the mutual information (MI) between the input image X and its label Y, i.e. $I(X;Y)$. Under this view, when Cutout is inserted as a common augmentation, each training sample is first masked to produce a subview V, and the model then learns from V alone, as shown in Fig. 3. From the MI perspective, this procedure optimizes only $I(V;Y)$, treating Cutout as an ordinary augmentation and silently discarding the complementary cues hidden in the occluded region of X. This insight reveals precisely why naive integration fails: substantial label-relevant information is left unexploited.

Rather than treating MI as a single, monolithic objective, we adopt a divide-and-conquer strategy that invokes the MI chain rule to split it into two complementary terms, $I(X;Y) = I(V;Y) + I(X;Y|V)$. The first term, $I(V;Y)$, forces the network to predict the label Y from a partially occluded image, thereby emphasizing the label-relevant structures which survive after masking and are more likely to generalize across style shifts. The second term, $I(X;Y|V)$, drives the model to mine the full image X for discriminative cues that the subview V cannot provide, thus recovering the residual information essential for accurate classification. Crucially, these two terms operate in concert: the gradient of $I(X;Y|V)$ flows back through the encoder shared with V, implicitly teaching the subview representation to also encode information about what was masked out.

We evaluate our method against strong OOD baselines on several benchmarks. Experimental results show consistent and significant gains over both standard hyperspherical learning and naive applications of Cutout. Our approach achieves state-of-the-art performance on these benchmarks, thereby validating that an MI-driven, subview-based integration of Cutout, grounded in chain-rule decomposition, provides a practical and theoretically principled benefit to OOD generalization. A comprehensive discussion of related work is provided in Appendix A.

2 Preliminaries

We denote vectors in boldface lowercase (e.g., $\mathbf{x}$), random variables in uppercase (e.g., X, Y), and their realizations in the corresponding lowercase. $\mathbb{P}_{XY}$ denotes the joint distribution of X and Y over spaces $\mathcal{X}$ and $\mathcal{Y}$, and p_{XY}, p_X denote the corresponding density functions. The indicator function is written as $\mathbb{1}\{\cdot\}$. Considering K-way classification as a case study, we use $\mathcal{X}$ and $\mathcal{Y} = \{1, \ldots, K\}$ to denote the input and label spaces. In practice, data points are sampled from a set of environmental conditions (also known as domains) $\mathcal{E}$, of which only a finite subset $\mathcal{E}_{tr} \subset \mathcal{E}$ is accessible during training. Our goal is to find an optimal classifier $f^* : \mathcal{X} \to \mathbb{R}^K$ that, given $\mathcal{E}_{tr}$, generalizes best across all environmental conditions in $\mathcal{E}$, formally stated as follows:

Definition 1 (**OOD Generalization**). *For each training domain $e \in \mathcal{E}_{tr}$, we are given a labeled training dataset $\mathcal{D}_e = \{(\mathbf{x}_i^e, y_i^e)\}_{i=1}^{|\mathcal{D}_e|}$, where $(\mathbf{x}_i^e, y_i^e)$ is drawn i.i.d. from the domain-conditioned joint distribution $\mathbb{P}_{XY|E=e}$. The goal of OOD generalization is to learn an optimal classifier $f^* : \mathcal{X} \to \mathbb{R}^K$ from $\mathcal{D} = \bigcup_{e \in \mathcal{E}_{tr}} \mathcal{D}_e$ that minimizes the worst-case risk over all domains $\mathcal{E}$:*

$$f^* = \arg\min_{f \in \mathcal{F}} \max_{e \in \mathcal{E}} \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY|\mathcal{E}=e}} \left[\ell(f(\mathbf{x}), y)\right], \tag{1}$$

where $\mathcal{F}$ is a hypothesis space (e.g., parameterized by a neural network), and ℓ is a loss function.

OOD generalization is challenging due to the unavailability of data from test environments $\mathcal{E}_{te} = \mathcal{E} \setminus \mathcal{E}_{tr}$. Ye et al. [2021] prove that the OOD generalization error can be bounded in terms of intra-class variation and inter-class separation across domains. Building on this insight, the recent HYPO [Bai et al., 2024] pulls features toward class prototypes (reducing intra-class variation) while pushing prototypes apart on the unit hypersphere (increasing inter-class separation), implemented by minimizing the following empirical risk:

$$\begin{aligned}\mathcal{L}_{\text{HYPO}} \triangleq &- \frac{1}{|\mathcal{D}|} \sum_{e \in \mathcal{E}_{tr}} \sum_{(\mathbf{x},y) \in \mathcal{D}_e} \log \frac{\exp(g(\mathbf{x})^\top \boldsymbol{\mu}_y / \kappa)}{\sum_{j=1}^K \exp(g(\mathbf{x})^\top \boldsymbol{\mu}_j / \kappa)} \\ &+ \frac{1}{K} \sum_{i \in \mathcal{Y}} \log \frac{1}{K-1} \sum_{j \in \mathcal{Y} \setminus \{i\}} \exp(\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j / \kappa),\end{aligned} \tag{2}$$

where $\kappa > 0$ is a temperature hyperparameter, $g(\mathbf{x}) \in \mathcal{S}^{d-1}$ is the representation produced by the neural network $g(\cdot)$ on the unit hypersphere $\mathcal{S}^{d-1} \triangleq \{\mathbf{r} \in \mathbb{R}^d \mid \|\mathbf{r}\|_2 = 1\}$, and $\boldsymbol{\mu}_y \in \mathcal{S}^{d-1}$ denotes the prototype of class y.

Cutout [Devries and Taylor, 2017] is a simple yet powerful augmentation that randomly masks out a square region of an image, forcing the model to learn from the remaining pixels.

Definition 2 (**Cutout**). *For an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ and a side length s, sample a top-left corner (i_0, j_0) uniformly at random from $\{0, \ldots, H - s\} \times \{0, \ldots, W - s\}$, and define the masked region as $\Omega = \{(i, j) : i_0 \le i < i_0 + s,\ j_0 \le j < j_0 + s\}$. The Cutout transform $T_{\text{cutout}}(\mathbf{x}; \Omega)$ replaces all pixels in Ω with 0 and leaves the rest of $\mathbf{x}$ unchanged.*

While Cutout has proven highly effective for standard supervised learning, our empirical study (Fig. 2) shows that simply combining Cutout with existing OOD methods often yields marginal or even negative gains. This observation motivates a principled framework that fully harnesses Cutout for OOD generalization, which we develop next.

3 Methodology

3.1 Motivation

Our framework is motivated by rethinking hyperspherical prototype learning from the perspective of the InfoMax principle [Bell and Sejnowski, 1995], where the goal is to maximize the MI between input X and label Y. We begin with the formal definition of MI:

Definition 3 (**Mutual Information (MI)** [Shannon, 2001]). *Given two random variables X and Y, the MI between X and Y is the Kullback–Leibler (KL) divergence between the joint distribution*

$\mathbb{P}_{XY}$ *and the product of marginal distributions* $\mathbb{P}_X\mathbb{P}_Y$*:*

$$\begin{aligned} I(X;Y) &\triangleq D_{\mathrm{KL}}(\mathbb{P}_{XY} \,\|\, \mathbb{P}_X\mathbb{P}_Y) \\ &= \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}\right]. \end{aligned} \tag{3}$$

Direct computation of Eq. (3) requires the conditional distribution $\mathbb{P}_{Y|X}$ and the marginal $\mathbb{P}_Y$, both of which are typically unknown in practice [van den Oord et al., 2018, Poole et al., 2019]. A standard remedy is to replace $p_{Y|X}(y|\mathbf{x})$ with a learnable estimator $\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})$ (parameterized by $\boldsymbol{\theta}$, with $\sum_{k=1}^{K}\hat{p}_{Y|X}(y_k|\mathbf{x};\boldsymbol{\theta}) = 1$), yielding the following tractable variational lower bound:

Definition 4 (**Variational MI Lower Bound** [Barber and Agakov, 2004])**.** *For two random variables* X, Y*, the mutual information* $I(X;Y)$ *satisfies:*

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}\right] \\ &\geq I_{VAR}(X;Y) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log \frac{\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})}{p_Y(y)}\right] \end{aligned} \tag{4}$$

To tractably derive $\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})$, motivated by van den Oord et al. [2018], we assume it belongs to the energy-based variational family that uses a critic $\phi_{\boldsymbol{\theta}}(\mathbf{x}, y)$ parameterized by $\boldsymbol{\theta}$ and is scaled by $p_Y(y)$:

$$\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta}) = \frac{p_Y(y)\exp\phi_{\boldsymbol{\theta}}(\mathbf{x},y)}{\sum_{k=1}^{K} p_Y(y_k)\exp\phi_{\boldsymbol{\theta}}(\mathbf{x},y_k)}. \tag{5}$$

Combining Eq. (5) with Eq. (4) and assuming a uniform class prior $p_Y(y_k) = 1/K$, we obtain:

$$I_{\mathrm{VAR}}(X;Y) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log \frac{\exp\phi_{\boldsymbol{\theta}}(\mathbf{x},y)}{\sum_{k=1}^{K}\exp\phi_{\boldsymbol{\theta}}(\mathbf{x},y_k)}\right] + \log K. \tag{6}$$

Setting $\phi_{\boldsymbol{\theta}}(\mathbf{x}, y) = g(\mathbf{x})^\top \boldsymbol{\mu}_y/\kappa$, we recognize the first term of $\mathcal{L}_{\mathrm{HYPO}}$ in Eq. (2) as an estimator of $-I_{\mathrm{VAR}}(X;Y)$ (up to a constant). This provides an MI-based interpretation of HYPO's mechanism.

A straightforward way to incorporate Cutout into this MI framework is to replace the input $\mathbf{x}$ in Eq. (2) with its Cutout subview $\mathbf{v} = T_{\mathrm{cutout}}(\mathbf{x};\Omega)$, as illustrated in Fig. 3. From the MI perspective, this naive variant optimizes only $I(V;Y)$, leaving the conditional information $I(X;Y\,|\,V)$ hidden in the occluded region completely unexploited. As empirically shown in Fig. 2, this explains why naively combining Cutout with HYPO yields only marginal or even negative gains.

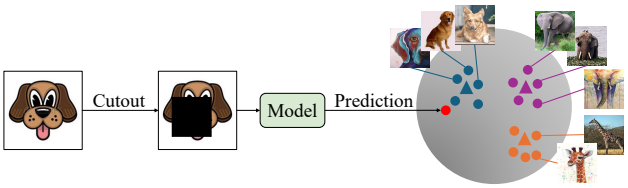


Figure 3: Illustration of the naive HYPO variant using Cutout. The image is masked to obtain a subview, which is aligned to class prototypes. Information in occluded region is discarded.

3.2 Decomposing Mutual Information

An overview of our framework is shown in Fig. 4. We propose to fully exploit the benefits of Cutout by decomposing MI estimation into more manageable subtasks via the chain rule of MI. Inspired by the divide-and-conquer philosophy [Peng et al., 2025b], rather than directly estimating the entire MI between the input X and the label Y, we break it down into two components: (i) $I(V;Y)$ measures how much label information survives in the partially occluded view V; and (ii) $I(X;Y\,|\,V)$ captures the residual label information that the full image X contributes beyond what V already provides. This decomposition assigns complementary roles to the two terms: the first encourages each subview to align tightly with its class label, while the second drives the model to fill in the gaps by mining discriminative cues absent from the subview. Formally, we begin with the chain rule of MI:

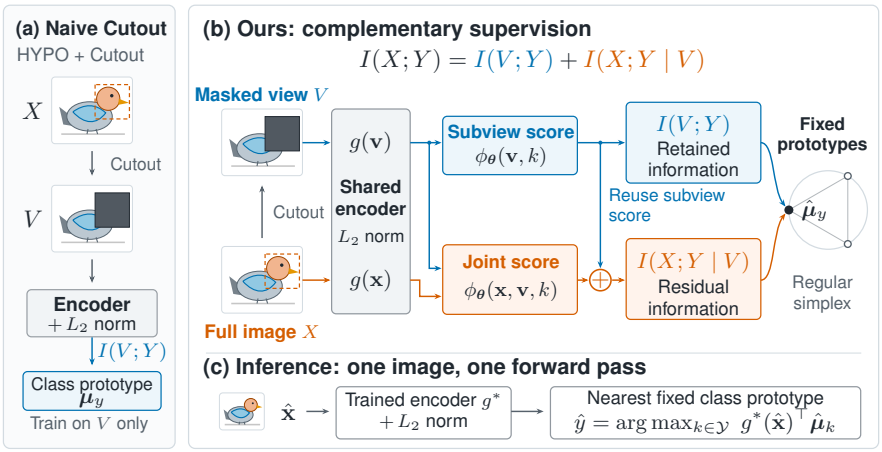


Figure 4: Overview of our framework. (a) Naive Cutout uses V alone. (b) Our method jointly optimizes $I(V;Y)$ and $I(X;Y \mid V)$ to exploit retained cues and additional label information from X, using a shared encoder and fixed prototypes. (c) During inference, we use only a single image.

Definition 5 (Chain Rule for MI). *Given three random variables X, Y, and V, the mutual information $I(X,V;Y)$ can be decomposed as:*

$$I(X,V;Y) = I(X;Y) + I(V;Y \mid X) = I(V;Y) + I(X;Y \mid V). \tag{7}$$

In this way, we can rewrite $I(X;Y)$ as a linear combination of three independent MI terms:

$$\begin{aligned} I(X;Y) &= I(V;Y) + I(X;Y \mid V) - I(V;Y \mid X) && (8)\\ &\leq I(V;Y) + I(X;Y \mid V). && (9) \end{aligned}$$

The inequality in Eq. (9) becomes an equality when $I(V;Y \mid X) = 0$. We now show that this condition holds exactly in our Cutout-based setup.

Remark 1 (Validity of $I(V;Y \mid X) = 0$). *In our setting, the Cutout subview is generated as $V = T_{\text{cutout}}(X;\Omega)$, where the random mask Ω is sampled independently of (X,Y), i.e., $\Omega \perp\!\!\!\perp (X,Y)$. Conditioned on X, the subview V is a deterministic function of X and Ω, and the only randomness in V comes from Ω, which is independent of Y. Therefore V provides no additional information about Y once X is known, i.e., $V \perp\!\!\!\perp Y \mid X$, which gives $I(V;Y \mid X) = 0$. Consequently, the equality $I(X;Y) = I(V;Y) + I(X;Y \mid V)$ holds exactly.*

Under Remark 1, we obtain a new variational MI lower-bound estimator:

$$I_{\text{VAR}}(X;Y) \triangleq I_{\text{VAR}}(V;Y) + I_{\text{VAR}}(X;Y \mid V) \leq I(X;Y), \tag{10}$$

where $V = T_{\text{cutout}}(X;\Omega)$, and similar to Eq. (6), we have:

$$I_{\text{VAR}}(V;Y) = \mathbb{E}_{(\mathbf{v},y)\sim\mathbb{P}_{VY}}\left[\log \frac{\exp \phi_{\boldsymbol{\theta}}(\mathbf{v},y)}{\sum_{k=1}^{K} \exp \phi_{\boldsymbol{\theta}}(\mathbf{v},y_k)}\right] + \log K. \tag{11}$$

Furthermore, the conditional MI $I(X;Y \mid V)$ is defined as:

$$\begin{aligned} I(X;Y \mid V) &= \mathbb{E}_{(\mathbf{x},\mathbf{v},y)\sim\mathbb{P}_{XVY}}\left[\log \frac{p_{XY|V}(\mathbf{x},y \mid \mathbf{v})}{p_{X|V}(\mathbf{x} \mid \mathbf{v}) p_{Y|V}(y \mid \mathbf{v})}\right] \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{v},y)\sim\mathbb{P}_{XVY}}\left[\log \frac{p_{Y|XV}(y \mid \mathbf{x},\mathbf{v})}{p_{Y|V}(y \mid \mathbf{v})}\right]. \end{aligned} \tag{12}$$

Similar to Eq. (4), let $\hat{p}_{Y|XV}(y|\mathbf{x},\mathbf{v};\boldsymbol{\theta})$ and $\hat{p}_{Y|V}(y|\mathbf{v};\boldsymbol{\theta})$ denote the estimators of $p_{Y|XV}$ and $p_{Y|V}$ with learnable parameters $\boldsymbol{\theta}$. We can derive $I_{\text{VAR}}(X;Y \mid V)$ as:

$$\begin{aligned} I(X;Y \mid V) &\geq I_{\text{VAR}}(X;Y \mid V) \\ &\triangleq \mathbb{E}_{(\mathbf{x},\mathbf{v},y)\sim\mathbb{P}_{XVY}}\left[\log \frac{\hat{p}_{Y|XV}(y \mid \mathbf{x},\mathbf{v};\boldsymbol{\theta})}{\hat{p}_{Y|V}(y \mid \mathbf{v};\boldsymbol{\theta})}\right]. \end{aligned} \tag{13}$$

Discussion. A natural baseline for using Cutout as augmentation is to train the network on the masked view V alone, effectively maximizing $I(V;Y)$. Even if one tries to include both X and V but treats them as two independent views, the objective collapses to $I(X,V;Y) = I(V;Y) + I(X;Y)$, valid only when X and V are statistically independent—generally not the case. Hence, the baseline either ignores the occluded region or double-counts redundant information, leaving the complementary cues hidden behind the mask untapped. By contrast, our chain-rule decomposition $I(X,V;Y) = I(V;Y) + I(X;Y \mid V)$ is always valid. The conditional term $I(X;Y \mid V)$ explicitly targets the residual label-relevant information hidden in the occluded region, compelling the model to recover discriminative cues that the masked view alone cannot provide. This principled decomposition avoids redundancy and yields richer, more robust representations for OOD generalization, as confirmed in our experiments. Further discussion on the tightness of the variational bound is in Appendix B.

3.3 Effective Parameterization

To turn the variational lower bounds in Eqs. (10), (11) and (13) into a tractable, end-to-end objective optimizable by gradient descent, we now parameterize the conditional densities $\hat{p}_{Y|V}$ and $\hat{p}_{Y|XV}$ with energy-based variational families. We instantiate the energy functions on unit hypersphere as:

$$\phi_{\boldsymbol{\theta}}(\mathbf{v}, y) = \frac{g(\mathbf{v})^\top \boldsymbol{\mu}_y}{\kappa}, \quad \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y) = \frac{(g(\mathbf{v}) + g(\mathbf{x}))^\top \boldsymbol{\mu}_y}{2\tau}, \tag{14}$$

where $g(\cdot) \in \mathcal{S}^{d-1}$ is the encoder shared between V and X, $\boldsymbol{\mu}_y \in \mathcal{S}^{d-1}$ is the prototype of class y, and κ, τ are two distinct temperature factors. The two energy functions correspond to two complementary forms of supervision: $\phi_{\boldsymbol{\theta}}(\mathbf{v}, y)$ measures the alignment of the subview alone with the class prototype, while $\phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y)$ measures the alignment of the fused subview-and-full-image representation with the prototype.

Parameterizing $\hat{p}_{Y|V}$. Following the same energy-based form as in Eq. (5), the conditional $\hat{p}_{Y|V}$ takes the standard softmax form:

$$\hat{p}_{Y|V}(y|\mathbf{v};\boldsymbol{\theta}) = \frac{\exp \phi_{\boldsymbol{\theta}}(\mathbf{v}, y)}{\sum_{k=1}^{K} \exp \phi_{\boldsymbol{\theta}}(\mathbf{v}, y_k)}. \tag{15}$$

Parameterizing $\hat{p}_{Y|XV}$. For the joint conditional $\hat{p}_{Y|XV}$, we use an energy-based variational family that takes $\hat{p}_{Y|V}$ as a base distribution scaled by an additional energy term $\phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y)$ that incorporates information from the full image X:

$$\hat{p}_{Y|XV}(y|\mathbf{x}, \mathbf{v};\boldsymbol{\theta}) = \frac{\hat{p}_{Y|V}(y|\mathbf{v};\boldsymbol{\theta}) \exp \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y)}{\sum_{k=1}^{K} \hat{p}_{Y|V}(y_k|\mathbf{v};\boldsymbol{\theta}) \exp \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y_k)}. \tag{16}$$

This design directly reflects the chain-rule decomposition: the base distribution $\hat{p}_{Y|V}$ captures the prediction from the subview, and the additional energy term refines this prediction using the residual information from the full image.

Deriving the final objective. Substituting Eqs. (15) and (16) into the variational bound Eq. (13), and assuming a uniform class prior, we obtain the parameterized form of $I_{\mathrm{VAR}}(X;Y \mid V)$:

$$I_{\mathrm{VAR}}(X;Y \mid V) = \mathbb{E}_{(\mathbf{x},\mathbf{v},y)\sim\mathbb{P}_{XVY}}\left[\log \frac{\exp[\phi_{\boldsymbol{\theta}}(\mathbf{v}, y) + \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y)]}{\sum_{k=1}^{K} \exp[\phi_{\boldsymbol{\theta}}(\mathbf{v}, y_k) + \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y_k)]}\right] + \log K. \tag{17}$$

Combining Eqs. (11) and (17), the full variational lower bound on $I(X;Y)$ becomes:

$$\begin{aligned} I_{\mathrm{VAR}}(X;Y) &\triangleq I_{\mathrm{VAR}}(V;Y) + I_{\mathrm{VAR}}(X;Y \mid V) \\ &= \mathbb{E}\left[\log \frac{\exp \phi_{\boldsymbol{\theta}}(\mathbf{v}, y)}{\sum_{k=1}^{K} \exp \phi_{\boldsymbol{\theta}}(\mathbf{v}, y_k)}\right] + \mathbb{E}\left[\log \frac{\exp[\phi_{\boldsymbol{\theta}}(\mathbf{v}, y) + \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y)]}{\sum_{k=1}^{K} \exp[\phi_{\boldsymbol{\theta}}(\mathbf{v}, y_k) + \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{v}, y_k)]}\right] + 2\log K. \end{aligned} \tag{18}$$

The first term aligns the subview $\mathbf{v}$ with its class prototype, while the second term aligns the fused subview-and-full-image representation with the same prototype, jointly recovering the residual information missed by the first term.

3.4 Learning and Inference

HYPO [Bai et al., 2024] improves OOD generalization by pulling features toward class prototypes while simultaneously pushing those prototypes apart. In practice, the separation is enforced by minimizing the second term of $\mathcal{L}_{\mathrm{HYPO}}$ in Eq. (2). Although effective, this design has two drawbacks: (i) it introduces another loss term whose weight must be tuned, and (ii) it requires a two-stage update in which each prototype is adjusted both by back-propagation and by exponential moving average [Li et al., 2021a]. We address these issues from a different perspective by posing prototype separation as a pure geometric optimization:

$$\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K = \arg \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K} \max_{i \neq j} \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j, \tag{19}$$

where each $\boldsymbol{\mu}_i$ lies on the unit sphere $\mathcal{S}^{d-1}$, i.e., $\boldsymbol{\mu}_i \in \mathcal{S}^{d-1}$, $\forall i = 1, \ldots, K$. Solving the optimization problem in Eq. (19) is equivalent to minimizing the second term of $\mathcal{L}_{\mathrm{HYPO}}$, which is precisely the K-point best-packing (Tammes) problem on a sphere [Musin and Tarasov, 2015]:

$$\arg \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K} \max_{i \neq j} \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = \arg \max_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K} \min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2. \tag{20}$$

Theorem 1. *For any $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathcal{S}^{d-1}$ with $d \geq 2$ and $d + 2 \geq K \geq 2$, the optimizer in Eq. (19) is uniquely given by the vertices of regular $(K-1)$-simplices inscribed in $\mathcal{S}^{d-1}$.*

Theorem 1 shows that this optimization problem has a unique closed-form minimizer given by:

$$\hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j = \frac{K}{K-1} \mathbb{1}\{i = j\} - \frac{1}{K-1}, \quad \forall i, j \in \mathcal{Y}. \tag{21}$$

Since a regular simplex maximizes angular separation [Bezdek and Lángi, 2024], fixing the prototypes at Eq. (21) exactly achieves the objective of HYPO's second term without any iterative updates or extra hyperparameters. This motivates us to approach the maximum discrimination among class prototypes by simply fixing $\boldsymbol{\mu}_i$ at the idealized position $\hat{\boldsymbol{\mu}}_i$ defined in Eq. (21). Consequently, the overall training becomes single-stage and more stable, while retaining the geometric intuition underpinning hyperspherical learning.

Following a batch-based training routine, our overall learning objective can be formulated as follows:

$$\begin{aligned} \mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x}, y) \in \mathcal{B}} \Bigg\{ & \log \frac{\exp[g(\mathbf{v})^\top \hat{\boldsymbol{\mu}}_y / \kappa]}{\sum_{k=1}^K \exp[g(\mathbf{v})^\top \hat{\boldsymbol{\mu}}_k / \kappa]} \\ & + \log \frac{\exp[g(\mathbf{v})^\top \hat{\boldsymbol{\mu}}_y / \kappa + (g(\mathbf{v}) + g(\mathbf{x}))^\top \hat{\boldsymbol{\mu}}_y / (2\tau)]}{\sum_{k=1}^K \exp[g(\mathbf{v})^\top \hat{\boldsymbol{\mu}}_k / \kappa + (g(\mathbf{v}) + g(\mathbf{x}))^\top \hat{\boldsymbol{\mu}}_k / (2\tau)]} \Bigg\}. \end{aligned} \tag{22}$$

In practice, the prototypes $\{\hat{\boldsymbol{\mu}}_i\}_{i=1}^K$ can be effortlessly obtained in an offline manner:

$$\hat{\mathbf{U}} = \sqrt{\frac{K}{K-1}} \mathbf{R} \left(\mathbf{I}_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^\top \right), \tag{23}$$

where $\hat{\mathbf{U}} = [\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K] \in \mathbb{R}^{d \times K}$, $\mathbf{1}_K$ is a K-dimensional all-ones vector, and $\mathbf{R} \in \mathbb{R}^{d \times K}$ is an orthogonal matrix satisfying $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_K$. The matrix $\mathbf{R}$ is constructed deterministically once before training (e.g., via QR decomposition of a random Gaussian matrix with a fixed seed) and remains frozen throughout, introducing no run-to-run variability. We provide further analysis on the role of $\mathbf{R}$, the simplex feasibility condition $d + 2 \geq K$, and the limitations of fixed prototypes in Appendix C.

Inference. With the well-trained model g^*, inference for a test sample $\hat{\mathbf{x}}$ is straightforward and only requires computing:

$$\hat{y} = \arg \max_{k \in \mathcal{Y}} \big(g^*(\hat{\mathbf{x}})^\top \hat{\boldsymbol{\mu}}_k \big). \tag{24}$$

An end-to-end pseudo-algorithm is summarized in Appendix D.

Table 1: Comparison with OOD generalization methods on PACS and Office-Home. We report the mean ± standard deviation over three random seeds. The column-wise best is highlighted in bold.

Dataset	PACS				
Domain	**Art**	**Cartoon**	**Photo**	**Sketch**	**Avg.**
IRM	84.8	76.4	96.7	76.1	83.5
GDRO	83.5	79.1	96.7	78.3	84.4
I-Mixup	86.1	78.9	97.6	75.8	84.6
VREx	86.0	79.1	96.9	77.7	84.9
GVRT	87.9	78.4	98.2	75.7	85.0
ARM	86.8	76.8	97.4	79.3	85.1
RSC	85.4	79.7	97.6	78.2	85.2
ERM	84.7	80.8	97.2	79.3	85.5
SelfReg	87.9	79.4	96.8	78.3	85.6
SagNet	87.4	80.7	97.1	80.0	86.3
Pgrad	89.9	80.0	98.0	80.1	87.0
HYPO	87.2	82.3	98.0	**84.5**	88.0
Ours	**89.6**±0.6	**84.7**±0.4	**98.9**±0.0	83.9±0.3	**89.3**±0.2

Dataset	Office-Home				
Domain	**Art**	**Clipart**	**Product**	**Real**	**Avg.**
IRM	58.9	52.2	72.1	74.0	64.3
ARM	58.9	51.0	74.1	75.2	64.8
RSC	60.7	51.4	74.8	75.1	65.5
GDRO	60.4	52.7	75.0	76.0	66.0
VREx	60.7	53.0	75.3	76.6	66.4
ERM	63.1	51.9	77.2	78.1	67.6
SelfReg	63.6	53.1	76.9	78.1	67.9
SagNet	63.4	54.8	75.8	78.3	68.1
I-Mixup	62.4	54.8	76.9	78.3	68.1
Pgrad	65.8	55.4	78.0	80.0	69.8
GVRT	66.3	55.8	78.2	80.4	70.2
HYPO	68.3	57.9	79.0	81.4	71.6
Ours	**70.4**±0.2	**58.7**±0.1	**81.2**±0.1	**81.7**±0.0	**73.0**±0.1

Table 2: Comparison with OOD generalization methods on VLCS and Terra-Incognita. We report the mean ± standard deviation over three random seeds. The column-wise best is highlighted in bold.

Dataset	VLCS				
Domain	**Caltech**	**LabelMe**	**SUN**	**VOC**	**Avg.**
GDRO	97.3	63.4	69.5	76.7	76.7
RSC	97.9	62.5	72.3	75.6	77.1
MTL	97.8	64.3	71.5	75.3	77.2
I-Mixup	98.3	64.8	72.1	74.3	77.4
ERM	97.7	64.3	73.4	74.6	77.5
ARM	98.7	63.6	71.3	76.7	77.6
SelfReg	96.7	65.2	73.1	76.2	77.8
SagNet	97.9	64.5	71.4	77.5	77.8
HYPO	98.1	65.3	73.1	76.3	78.2
VREx	98.4	64.4	74.1	76.2	78.3
IRM	98.6	64.9	73.4	77.3	78.6
GVRT	98.8	64.0	75.2	77.9	79.0
Pgrad	98.3	64.4	74.4	79.9	79.3
Ours	**99.1**±0.0	**70.8**±0.6	**77.1**±0.2	**81.7**±0.3	**82.2**±0.3

Dataset	Terra-Incognita				
Domain	**Loc100**	**Loc38**	**Loc43**	**Loc46**	**Avg.**
GDRO	41.2	38.6	56.7	36.4	43.2
ARM	49.3	38.3	55.8	38.7	45.5
MTL	49.3	39.6	55.6	37.8	45.6
ERM	49.8	42.1	56.9	35.7	46.1
VREx	48.2	41.7	56.8	38.7	46.4
RSC	50.2	39.2	56.3	40.8	46.6
SelfReg	48.8	41.3	57.3	40.6	47.0
IRM	54.6	39.8	56.2	39.6	47.6
I-Mixup	59.6	42.2	55.9	33.9	47.9
GVRT	53.9	41.8	58.2	38.0	48.0
SagNet	53.0	43.0	57.9	40.4	48.6
Pgrad	61.2	45.7	58.2	37.9	50.8
HYPO	58.8	46.6	58.7	42.7	51.7
Ours	**62.5**±0.6	**49.2**±0.7	**60.1**±1.3	**43.9**±0.1	**53.9**±0.3

4 Experiments

4.1 Experimental Setup

Baseline Methods. We compare our method with representative baselines, including ERM [Vapnik, 1999], IRM [Arjovsky et al., 2019], I-Mixup [Xu et al., 2020], GroupDRO [Sagawa et al., 2019], RSC [Huang et al., 2020], VREx [Krueger et al., 2021], MTL [Blanchard et al., 2021], ARM [Zhang et al., 2021], SelfReg [Kim et al., 2021], SagNet [Nam et al., 2021], GVRT [Min et al., 2022], Pgrad [Wang et al., 2023b], PCL [Yao et al., 2022b], MODE [Dai et al., 2023], and HYPO [Bai et al., 2024]. It is worth noting that we have adopted the recommended configurations proposed by prior works, while concurrently standardizing the backbone architecture to ensure equitable comparisons.

Datasets. We conduct experiments on popular benchmarks, including PACS [Li et al., 2017], VLCS [Ghifary et al., 2015], OfficeHome [Venkateswara et al., 2017], Terra-Incognita [Beery et al., 2018], and DomainNet [Peng et al., 2019]. Model selection follows DomainBed's training-domain validation criterion (`IIDAccuracySelectionMethod`) [Gulrajani and Lopez-Paz, 2021], using only validation splits from the source domains. All reported results are averaged over three random seeds. We provide more details of the experimental setups in Appendix E.

Table 3: Comparison with OOD generalization methods on DomainNet. We report the mean ± standard deviation over three random seeds. The column-wise best is highlighted in bold.

Dataset	**DomainNet**						
Domain	**Clipart**	**Infograph**	**Painting**	**Quickdraw**	**Real**	**Sketch**	**Avg.**
IRM	48.5	15.0	38.3	10.9	48.2	42.3	33.9
ARM	49.7	16.3	40.9	9.4	53.4	43.5	35.5
RSC	55.0	18.3	44.4	12.2	55.7	47.8	38.9
I-Mixup	55.7	18.5	44.3	12.5	55.8	48.2	39.2
ERM	58.1	18.8	46.7	12.2	59.6	49.8	40.9
Pgrad	59.4	19.8	49.1	13.7	61.8	51.1	42.5
GVRT	62.4	21.0	50.5	13.8	64.6	52.4	44.1
MODE	68.5	23.1	53.0	15.9	63.7	55.0	46.5
PCL	67.9	**24.3**	55.3	15.7	**66.6**	56.4	47.7
Ours	**69.7**±0.1	**24.3**±0.1	**55.9**±0.1	**17.1**±0.2	**66.6**±0.1	**57.0**±0.1	**48.4**±0.1

Table 4: Ablation study on PACS, Office-Home, VLCS, and Terra-Incognita. We evaluate the contribution of each proposed component. The column-wise best results are highlighted in bold.

	Components				Accuracy (%)					
#	w/ Cutout	w/ Dual Input	w/ Fix Proto.	w/ MI Loss	PACS	Office	VLCS	Terra	Avg.	*Gain*
1	–	–	–	–	88.0	71.7	78.2	51.7	72.4	–
2	✓	–	–	–	88.0	71.8	78.4	51.9	72.5	+0.1
3	✓	✓	–	–	88.3	71.6	79.9	52.4	73.1	+0.7
4	✓	✓	✓	–	88.3	72.2	80.2	53.1	73.5	+1.1
5	✓	✓	✓	✓	**89.3**	**73.0**	**82.2**	**53.9**	**74.6**	**+2.2**

4.2 Main Results

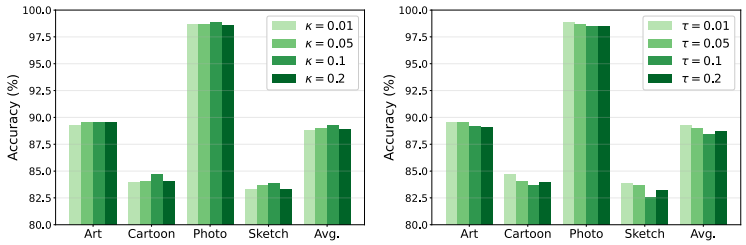
PACS. Table 1 (left) presents the OOD classification results on the PACS dataset. Our approach achieves the highest average OOD accuracy, outperforming the previous state-of-the-art by 1.3%. Notably, in challenging domains such as Art and Cartoon, where images exhibit significant stylistic variation, our approach surpasses the most recent HYPO method by 2.4% on average.

Office-Home. Table 1 (right) shows that our method achieves an average accuracy of 73.0% on Office-Home, surpassing the best baseline (HYPO, 71.6%) by 1.4%. Notably, our approach improves the Art domain from 68.3% to 70.4% and the Product domain from 79.0% to 81.2%, while also enhancing performance on the Clipart and Real domains. These results indicate that our method effectively promotes robust, domain-invariant representations.

VLCS. Table 2 (left) shows that we attain an average accuracy of 82.2%, which is 2.9% higher than the best-performing baseline (Pgrad at 79.3%). Notably, on the LabelMe domain—a particularly challenging setting due to high variability in scene composition—our approach boosts accuracy from 64.4% (Pgrad) and 65.3% (HYPO) to 70.8%. We also achieve competitive performance on Caltech (99.1%), SUN (77.1%), and VOC (81.7%), showing consistent improvements across all four domains.

Terra-Incognita. Table 2 (right) shows that we achieve an average accuracy of 53.9% on Terra-Incognita, outperforming the best baseline (HYPO, 51.7%) by 2.2%. Notable gains are observed on Loc38 (49.2% vs. HYPO's 46.6%) and Loc100 (62.5% vs. HYPO's 58.8%), suggesting that our method is more robust to severe distributional shifts induced by camera-trap variation. The consistent performance boost across all four locations supports our claim that decomposing MI estimation via the chain rule and leveraging Cutout-generated subviews provides a more robust mechanism for capturing domain-invariant features under challenging conditions.

DomainNet. Table 3 shows our results on six domains of DomainNet, where our method consistently outperforms or matches the best baselines. Notably, on the Clipart domain, we achieve 69.7%, surpassing PCL (67.9%) and MODE (68.5%). We also set new highs on Painting (55.9%) and Quickdraw (17.1%), improving upon PCL and MODE, respectively. For Infograph and Real, our method ties the best-reported scores of 24.3% and 66.6%. Overall, we obtain an average of 48.4%, exceeding the previous best baseline PCL (47.7%).



(a) Temperature κ (b) Temperature τ

Figure 5: Ablation study of the hyperparameters κ and τ on the PACS dataset across all four domains.

4.3 Ablation Studies

We conduct ablation experiments to evaluate the contribution of each proposed component, with results reported on four OOD benchmarks (PACS, Office-Home, VLCS, and Terra-Incognita) in Table 4. Starting from the HYPO baseline (#1, Eq. (2), 72.4% average), simply adding Cutout as augmentation (#2) yields a marginal +0.1% gain, consistent with our observations in Fig. 2. Feeding both the original image $\mathbf{x}$ and its Cutout subview $\mathbf{v}$ as a dual input (#3) further improves performance to 73.1%, indicating that providing additional information from $\mathbf{x}$ alone is insufficient to fully exploit Cutout. Replacing exponentially-moving-averaged prototypes [Li et al., 2021a] with the fixed simplex prototypes (#4) brings a further +0.4% gain, confirming that strong inter-class separation via closed-form prototypes leads to more stable and effective training. Finally, applying the proposed chain-rule MI loss (#5, Eq. (22)) achieves the best average accuracy of 74.6% (+2.2% over the baseline), with the largest single-component gain (+1.1%) on top of the strong dual-input baseline—demonstrating that the principled MI decomposition, rather than the dual input itself, drives the improvement.

4.4 Analyzing the Effect of κ and τ

Fig. 5a reports OOD generalization performance as we vary the temperature factor κ on the four PACS domains. Our method maintains competitive accuracy over a broad range of κ values, with the highest average performance (89.3%) achieved at $\kappa = 0.1$. Fig. 5b shows the corresponding analysis for τ, where smaller values yield stronger OOD performance while larger values (e.g., 0.1) lead to degraded performance. Overall, our method is robust to a wide range of temperature settings, suggesting that the chain-rule MI loss is not overly sensitive to hyperparameter tuning.

Additional analyses. We provide further analyses in the appendix, including the effect of Cutout side length s (Appendix F), generalization on corruption benchmarks (CIFAR-100-C, Appendix G), generality beyond Cutout (Appendix H), evaluation with a ViT-B/16 backbone (Appendix I), and t-SNE visualizations of learned representations (Appendix J).

5 Conclusion

We explored how to make OOD generalization benefit from Cutout. Motivated by the observation that naively plugging Cutout into existing OOD methods offers only marginal gains, we reinterpret hyperspherical prototype learning as MI maximization and decompose the objective via the chain rule into two complementary terms: aligning each Cutout subview with its label, and driving the full image to recover the residual information omitted by the mask. Combined with closed-form fixed prototypes, this yields a single-stage algorithm that fully exploits Cutout's supervision. Extensive experiments on five standard OOD benchmarks validate the effectiveness of our chain-rule MI decomposition.

Acknowledgments and Disclosure of Funding

The author thanks the area chair and the anonymous reviewers for their constructive feedback, which helped improve the quality and clarity of this paper. This work was self-funded by the author.

References

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

Haoyue Bai, Yifei Ming, Julian Katz-Samuels, and Yixuan Li. HYPO: hyperspherical out-of-distribution generalization. In *ICLR*. OpenReview.net, 2024.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV (16)*, volume 11220 of *Lecture Notes in Computer Science*, pages 472–489. Springer, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7(6):1129–1159, 1995.

Károly Bezdek and Zsolt Lángi. From the separable tammes problem to extremal distributions of great circles in the unit sphere. *Discret. Comput. Geom.*, 72(1):269–309, 2024.

Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 22:2:1–2:55, 2021.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *ECCV (23)*, volume 13683 of *Lecture Notes in Computer Science*, pages 440–457. Springer, 2022.

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In *NeurIPS*, 2023.

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018.

Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020.

Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, and Xinmei Tian. Moderately distributional exploration for domain generalization. In *ICML*, pages 6786–6817. PMLR, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):1087–1105, 2023.

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559. IEEE Computer Society, 2015.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*. OpenReview.net, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Yuanwei Hu, Bo Peng, Yadan Luo, Zhen Fang, Ling Chen, and Jie Lu. Respecting modality gap in post-hoc out-of-distribution detection with pre-trained vision-language models. In *Forty-third International Conference on Machine Learning (ICML)*, 2026.

Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV (2)*, volume 12347 of *Lecture Notes in Computer Science*, pages 124–140. Springer, 2020.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *ICLR*. OpenReview.net, 2024.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, pages 9599–9608. IEEE, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Colorado Reed, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *AAAI*, pages 7399–7407. AAAI Press, 2022.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5543–5551. IEEE Computer Society, 2017.

Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Mopro: Webly supervised learning with momentum prototypes. In *ICLR*. OpenReview.net, 2021a.

Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *CVPR*, pages 224–233. Computer Vision Foundation / IEEE, 2021b.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

Wenliang Liu, Guanding Yu, Lele Wang, and Renjie Liao. An information-theoretic framework for out-of-distribution generalization. In *ISIT*, pages 2670–2675. IEEE, 2024a.

Wenliang Liu, Guanding Yu, Lele Wang, and Renjie Liao. An information-theoretic framework for out-of-distribution generalization with applications to stochastic gradient langevin dynamics. *arXiv preprint arXiv:2403.19895*, 2024b.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR (Poster)*. OpenReview.net, 2017.

Mohammadjavad Maheronnaghsh and Taha Akbari Alvanagh. Robustness to spurious correlation: A comprehensive review. In *ECCV Workshops (18)*, volume 15640 of *Lecture Notes in Computer Science*, pages 361–379. Springer, 2024.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

Seonwoo Min, Nokyung Park, Siwon Kim, Seunghyun Park, and Jinkyu Kim. Grounding visual representations with texts for domain generalization. In *ECCV (37)*, volume 13697 of *Lecture Notes in Computer Science*, pages 37–53. Springer, 2022.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022.

Oleg R. Musin and Alexey S. Tarasov. The tammes problem for $N = 14$. *Exp. Math.*, 24(4):460–468, 2015.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, pages 8690–8699. Computer Vision Foundation / IEEE, 2021.

Bo Peng and Wenjie Zhu. Deep structural contrastive subspace clustering. In *Asian Conference on Machine Learning (ACML)*, pages 1145–1160. PMLR, 2021.

Bo Peng, Zhen Fang, Guangquan Zhang, and Jie Lu. Knowledge distillation with auxiliary variable. In *The International Conference on Machine Learning (ICML)*, 2024a.

Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. On the provable importance of gradients for autonomous language-assisted image clustering. In *The International Conference on Computer Vision (ICCV)*, 2025a.

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. An information-theoretical framework for understanding out-of-distribution detection with pretrained vision-language models. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025b.

Bo Peng, Jie Lu, Yonggang Zhang, Guangquan Zhang, and Zhen Fang. Distributional prototype learning for out-of-distribution detection. In *The SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025c.

Bo Peng, Yuanwei Hu, Bo Liu, Ling Chen, Jie Lu, and Zhen Fang. Delving into spectral clustering with vision-language representations. In *ICLR*. proceedings.iclr.cc, 2026a.

Bo Peng, Jie Lu, Zhen Fang, and Guangquan Zhang. A close look at negative label guided out-of-distribution detection in pre-trained vision-language models. In *Forty-third International Conference on Machine Learning (ICML)*, 2026b.

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. Debiased negative mining improves out-of-distribution detection with pre-trained vision-language models. In *The SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026c.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415. IEEE, 2019.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 2019.

Martin Riedmiller and A Lernen. Multi layer perceptron. *Machine learning lab special lecture, University of Freiburg*, 24, 2014.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019.

Claude E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021.

Kuo Shi, Jie Lu, Zhen Fang, and Guangquan Zhang. Unsupervised domain adaptation enhanced by fuzzy prompt learning. *IEEE Trans. Fuzzy Syst.*, 32(7):4038–4048, 2024.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing CLIP to out-of-distributions. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 31716–31731. PMLR, 2023.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

Vladimir Vapnik. An overview of statistical learning theory. *IEEE Trans. Neural Networks*, 10(5): 988–999, 1999.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5385–5394. IEEE Computer Society, 2017.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In *NeurIPS*, pages 2215–2227, 2021.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.*, 35(8):8052–8072, 2023a.

Jingye Wang, Ruoyi Du, Dongliang Chang, Kongming Liang, and Zhanyu Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *ACM Multimedia*, pages 4821–4829. ACM, 2022.

Shanshan Wang, ALuSi, Xun Yang, Ke Xu, Huibin Tan, and Xingyi Zhang. Dual-stream feature augmentation for domain generalization. In *ACM Multimedia*, pages 1111–1119. ACM, 2024a.

Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *ICRA*, pages 6687–6694. IEEE, 2024b.

Zhe Wang, Jake Grigsby, and Yanjun Qi. Pgrad: Learning principal gradients for domain generalization. In *ICLR*. OpenReview.net, 2023b.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, pages 6502–6509. AAAI Press, 2020.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392. Computer Vision Foundation / IEEE, 2021.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *Int. J. Comput. Vis.*, 132(12):5635–5662, 2024.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. Out-of-distribution generalization in natural language processing: Past, present, and future. In *EMNLP*, pages 4533–4559. Association for Computational Linguistics, 2023.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 25407–25437. PMLR, 2022a.

Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *CVPR*, pages 7097–7107, 2022b.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. In *NeurIPS*, pages 23519–23531, 2021.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. Spurious correlations in machine learning: A survey. *CoRR*, abs/2402.12715, 2024.

Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, pages 9035–9044. IEEE, 2019.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proc. IEEE*, 112(10):1583–1609, 2024.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *NeurIPS*, pages 23664–23678, 2021.

Yonggang Zhang, Jie Lu, Bo Peng, Zhen Fang, and Yiu-ming Cheung. Learning to shape in-distribution feature space for out-of-distribution detection. In *NeurIPS*, 2024.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008. AAAI Press, 2020.

Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032. AAAI Press, 2020a.

Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV (16)*, volume 12361 of *Lecture Notes in Computer Science*, pages 561–578. Springer, 2020b.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*. OpenReview.net, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4396–4415, 2023.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *Int. J. Comput. Vis.*, 132(3):822–836, 2024.

Qinli Zhou, Wenjie Zhu, Hao Chen, and Bo Peng. Community detection in multiplex networks by deep structure-preserving non-negative matrix factorization. *Applied Intelligence*, 2025.

Wenjie Zhu and Bo Peng. Manifold-based aggregation clustering for unsupervised vehicle re-identification. *Knowledge-Based Systems*, 2022.

Wenjie Zhu, Bo Peng, Han Wu, and Binhao Wang. Query set centered sparse projection learning for set based image classification. *Applied Intelligence*, 2020.

Wenjie Zhu, Bo Peng, and Chunchun Chen. Self-supervised embedding for subspace clustering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 3687–3691, 2021.

Wenjie Zhu, Bo Peng, Chunchun Chen, and Hao Chen. Deep discriminative dictionary pair learning for image classification. *Applied Intelligence*, 2023.

Wenjie Zhu, Bo Peng, and Wei Qi Yan. Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification. *IEEE Transactions on Multimedia*, 2024.

Wenjie Zhu, Bo Peng, and Wei Qi Yan. Deep inductive and scalable subspace clustering via nonlocal contrastive self-distillation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Limitations

Fixed regular-simplex prototypes impose equal separation between classes and do not explicitly encode their semantic relationships, which may be suboptimal for fine-grained recognition. In addition, the Cutout side length remains a fixed hyperparameter in our current implementation. Future work could explore randomizing the Cutout size during training.

A Related Work

Out-of-Distribution Generalization. Learning discriminative representations [Zhou et al., 2025] is a central objective in image classification [Zhu et al., 2020, 2023]. OOD generalization further requires these representations learned from source domains to remain transferable and robust on unseen target domains, without access to target data [Fang et al., 2023, Shi et al., 2024]. A common strategy is to extract domain-invariant features [Arjovsky et al., 2019, Krueger et al., 2021, Wald et al., 2021, Li et al., 2021b]: IRM [Arjovsky et al., 2019] introduces an invariant risk regularizer, while VREx [Krueger et al., 2021] reduces variance across training risks. On the theoretical side, Ye et al. [2021] bounded OOD error by intra-class variation and inter-class separation, motivating HYPO [Bai et al., 2024], which uses hyperspherical learning to tighten features around class prototypes. Building on this geometric perspective, we reinterpret hyperspherical prototype learning as MI maximization and apply the MI chain rule to decompose the objective, allowing Cutout-generated subviews to provide complementary supervision that markedly boosts OOD performance.

Data Augmentation. Data augmentation has been widely used in visual learning, including subspace clustering [Peng and Zhu, 2021, Zhu et al., 2025], unsupervised person re-identification [Zhu et al., 2024], and knowledge distillation [Peng et al., 2024a]. Augmentation methods for OOD generalization broadly fall into four categories: (i) image generation [Zhou et al., 2020a,b], which trains generators for style-shifted samples but is costly to train; (ii) image transformation [Xu et al., 2021, Wang et al., 2022, Yao et al., 2022a], which perturbs styles (e.g., via Fourier interpolation) but risks distorting semantics; (iii) feature-level augmentation [Yoo et al., 2019, Zhou et al., 2021, Chen et al., 2023] such as MixStyle, which blends instance styles but lacks invariance guarantees; and (iv) masking-based augmentation, which our work builds on. Beyond Cutout [Devries and Taylor, 2017], this category includes Random Erasing and Hide-and-Seek that mask random regions or grid cells, and RSC [Huang et al., 2020] that adversarially erases the most predictive features at the feature level. While sharing the intuition of "learning from incomplete observations", these methods all optimize a single objective on the masked input alone, leaving information in the occluded region unexploited. Our chain-rule decomposition is in principle applicable to any such "information-removal" augmentation: it explicitly recovers the residual information through the conditional term $I(X;Y|V)$, providing a unifying framework that complements rather than competes with these designs.

Mutual Information Approaches in OOD Generalization. MI is a pivotal concept in OOD generalization. Cha et al. [2022] regularize the learning objective with MI to an oracle model; IIB [Li et al., 2022] minimizes MI between inputs and representations to suppress spurious correlations; and recent information-theoretic bounds [Liu et al., 2024a,b] interpolate between Integral Probability Metric and f-divergence. Unlike these methods, which typically minimize MI with nuisance variables, we decompose the target MI $I(X;Y)$ via the chain rule and operationalize complementary supervision through Cutout subviews.

Out-of-Distribution Detection. OOD detection [Yang et al., 2024] is closely related to OOD generalization, offering insights into robust representation learning under distribution shifts. Representative methods derive detection scores from classifier energy [Liu et al., 2020] or class-conditional feature statistics [Lee et al., 2018]. Subsequent work develops tractable density estimation with ConjNorm [Peng et al., 2024b], shapes features toward prescribed mixture distributions [Zhang et al., 2024], and models intra-class variation using distributions of prototypes [Peng et al., 2025c]. Pretrained vision-language models (VLMs) enable detection through image–text alignment: MCM [Ming et al., 2022] compares inputs with in-distribution (ID) class descriptions, while NegLabel [Jiang et al., 2024] introduces negative labels from text corpora. Recent developments address negative-label sampling bias [Peng et al., 2026c], robustness to uncertain negative-label distributions [Peng et al., 2026b], and the modality gap through online visual prototype learning [Hu et al., 2026]. Closely

related to our formulation, Peng et al. [2025b] apply the chain rule of pointwise MI to image subviews for post-hoc OOD scoring. Their framework supports masking and cropping, with random crops used by default. Compared to their post-hoc scoring with pretrained VLMs, we use the MI chain rule to train a shared encoder on Cutout subviews and full images with fixed class prototypes for known-class classification across unseen domains.

B Discussion on the Tightness of the Variational Bound

In this section, we provide an in-depth analysis of the variational lower bounds used in our method, addressing both their relationship to alternative MI estimators and the implications of their tightness for OOD generalization.

B.1 Recap of the Variational Bounds

Our method relies on three variational MI estimators:

$$I_{\mathrm{VAR}}(X;Y) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log \hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})\right] + \log K, \tag{25}$$

$$I_{\mathrm{VAR}}(V;Y) \triangleq \mathbb{E}_{(\mathbf{v},y)\sim\mathbb{P}_{VY}}\left[\log \hat{p}_{Y|V}(y|\mathbf{v};\boldsymbol{\theta})\right] + \log K, \tag{26}$$

$$I_{\mathrm{VAR}}(X;Y\mid V) \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{v},y)\sim\mathbb{P}_{XVY}}\left[\log \frac{\hat{p}_{Y|XV}(y|\mathbf{x},\mathbf{v};\boldsymbol{\theta})}{\hat{p}_{Y|V}(y|\mathbf{v};\boldsymbol{\theta})}\right]. \tag{27}$$

All three are instances of the variational lower bound originally introduced by Barber and Agakov [2004], which approximates the true conditional density with a learnable parametric distribution $\hat{p}$. Specifically, for any random variables A and B, the Barber-Agakov bound states:

$$I(A;B) = \mathbb{E}_{(a,b)\sim\mathbb{P}_{AB}}\left[\log \frac{p_{B|A}(b|a)}{p_B(b)}\right] \geq \mathbb{E}_{(a,b)\sim\mathbb{P}_{AB}}\left[\log \frac{\hat{p}_{B|A}(b|a;\boldsymbol{\theta})}{p_B(b)}\right] \triangleq I_{\mathrm{BA}}(A;B), \tag{28}$$

where the gap $I(A;B) - I_{\mathrm{BA}}(A;B) = \mathbb{E}_{a\sim\mathbb{P}_A}\left[D_{\mathrm{KL}}(p_{B|A}(\cdot|a)\,\|\,\hat{p}_{B|A}(\cdot|a;\boldsymbol{\theta}))\right] \geq 0$, with equality iff the variational family is rich enough to capture the true conditional. In our setup, the energy-based family in Eq. (5) reduces to a softmax classifier when p_Y is uniform—a flexible parameterization in practice given a sufficiently expressive encoder $g(\cdot)$.

B.2 Comparison with Alternative MI Estimators

The MI estimation literature offers a spectrum of estimators with different bias-variance trade-offs [Poole et al., 2019]. Among the most representative alternatives, MINE [Belghazi et al., 2018] adopts the Donsker-Varadhan dual representation, $I(X;Y) \geq \sup_T \mathbb{E}_{\mathbb{P}_{XY}}[T(\mathbf{x},y)] - \log \mathbb{E}_{\mathbb{P}_X\mathbb{P}_Y}[e^{T(\mathbf{x},y)}]$, which is asymptotically tight but suffers from high variance and unstable optimization due to the exponential term in the denominator, especially in high-dimensional settings. InfoNCE [van den Oord et al., 2018] is a contrastive bound, $I(X;Y) \geq \log K + \mathbb{E}[\log(e^{f(\mathbf{x},y)}/\frac{1}{K}\sum_{k=1}^{K} e^{f(\mathbf{x},y_k)})]$, which is upper-bounded by $\log K$ but enjoys low variance and stable training. By contrast, the Barber-Agakov bound we adopt [Barber and Agakov, 2004] is governed by the expressiveness of the variational family $\hat{p}$.

We deliberately adopt the Barber-Agakov bound because in K-way classification, the conditional $p_{Y|X}$ is a categorical distribution, and the energy-based softmax in Eq. (5) forms a natural and expressive variational family. Crucially, our bound coincides with the standard cross-entropy loss up to an additive constant, which yields three concrete advantages. First, in terms of stability, the bound is bounded above by $\log K$ (a constant), unlike MINE-type bounds whose estimates can diverge during training. Second, in terms of compatibility, the cross-entropy form integrates seamlessly with hyperspherical prototype learning—the dot product $g(\mathbf{x})^\top \boldsymbol{\mu}_y/\kappa$ naturally serves as the energy function $\phi_{\boldsymbol{\theta}}$, requiring no additional MI estimator network. Third, in terms of sample efficiency, unlike contrastive bounds whose tightness improves with the number of negative samples, our bound's quality depends on the expressiveness of the encoder rather than on batch size.

B.3 What Tightness Means for OOD Generalization

It is well-known that variational lower bounds are not necessarily tight in absolute terms—a known limitation of MI estimation in high-dimensional settings [Poole et al., 2019, McAllester and Stratos,

2020]. We argue that this concern does not undermine our framework. Our method is built on the chain-rule decomposition $I(X;Y) = I(V;Y) + I(X;Y \mid V)$, and the empirical objective optimizes lower bounds for both terms. The key insight is that what matters for OOD generalization is not whether $I_{\mathrm{VAR}}(X;Y)$ recovers the true MI value, but whether the decomposed optimization directs the encoder to learn richer, more transferable features than the single-term baseline $I_{\mathrm{VAR}}(V;Y)$. Formally, suppose the gaps of the two bounds are $\Delta_1 = I(V;Y) - I_{\mathrm{VAR}}(V;Y)$ and $\Delta_2 = I(X;Y \mid V) - I_{\mathrm{VAR}}(X;Y \mid V)$, with $\Delta_1, \Delta_2 \geq 0$. Then:

$$I(X;Y) - \left[I_{\mathrm{VAR}}(V;Y) + I_{\mathrm{VAR}}(X;Y \mid V)\right] = \Delta_1 + \Delta_2, \tag{29}$$

and the relative improvement over the naive baseline becomes:

$$\underbrace{I_{\mathrm{VAR}}(V;Y) + I_{\mathrm{VAR}}(X;Y \mid V)}_{\text{ours}} - \underbrace{I_{\mathrm{VAR}}(V;Y)}_{\text{naive}} = I_{\mathrm{VAR}}(X;Y \mid V) \geq 0. \tag{30}$$

This relative gain is independent of Δ_1 (the gap of the $I(V;Y)$ estimator), meaning that even a loose absolute bound on $I(V;Y)$ does not contaminate the relative advantage from the conditional term. The conditional term provides a strictly non-negative contribution as long as it is optimized faithfully.

C Further Analysis on Fixed Prototypes

This section provides extended discussions of the fixed-prototype design introduced in Section 3.4, including (i) the construction and role of the orthogonal matrix $\mathbf{R}$, (ii) the feasibility of the regular simplex solution under our experimental settings, (iii) the limitations of fixed prototypes for fine-grained semantic structures, and (iv) the respective roles of the separation loss and fixed prototypes.

C.1 Construction of the Orthogonal Matrix R

The orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d \times K}$ in Eq. (23) satisfies $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_K$ and serves to embed the abstract regular $(K-1)$-simplex into the d-dimensional feature space. In practice, we obtain $\mathbf{R}$ by orthogonalizing a randomly initialized matrix (e.g., via QR decomposition of a Gaussian matrix), construct it once before training based on a chosen random seed, and keep it frozen throughout optimization.

A natural question is whether different choices of the random seed—and hence different realizations of $\mathbf{R}$—lead to different empirical results. We argue that this is not the case. The training loss in Eq. (22) depends on $\mathbf{R}$ only through inner products of the form $g(\mathbf{x})^\top \hat{\boldsymbol{\mu}}_y$ and $\hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j$. By construction (Eq. (21)), the prototype-prototype inner products $\hat{\boldsymbol{\mu}}_i^\top \hat{\boldsymbol{\mu}}_j$ are determined entirely by K and are independent of $\mathbf{R}$. For the feature-prototype inner products, suppose two valid orthogonal matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ are related by an orthogonal transformation $\mathbf{R}_2 = \mathbf{Q}\mathbf{R}_1$. Then any solution $g_1(\cdot)$ trained with prototypes from $\mathbf{R}_1$ corresponds to an equivalent solution $g_2(\mathbf{x}) = \mathbf{Q} g_1(\mathbf{x})$ that produces identical loss values when trained with prototypes from $\mathbf{R}_2$. In other words, different choices of $\mathbf{R}$ define equivalent training problems related by an isometry of the feature space; the choice of seed introduces no run-to-run variability beyond standard SGD-induced fluctuations.

C.2 Feasibility of the Regular Simplex Solution

Theorem 1 states that the regular $(K-1)$-simplex inscribed in $\mathcal{S}^{d-1}$ exists uniquely when $d+2 \geq K$, i.e., when the feature dimension is at least $K-1$. We verify that this condition is satisfied in all our experiments. We use a feature dimension of $d = 512$ throughout (following Bai et al. [2024]), and the largest number of classes among our benchmarks is $K = 345$ (DomainNet). Since $512 + 2 = 514 \geq 345$, the regular simplex prototypes are guaranteed to exist for every benchmark we evaluate on. For tasks with $K > d + 1$ (e.g., extreme-class classification with thousands of categories), the regular simplex no longer exists; in such cases, one would need to either increase the feature dimension or fall back to learnable prototypes. We do not explore this regime in this paper.

C.3 Limitations of Fixed Prototypes for Fine-Grained Tasks

The regular-simplex construction assigns equal angular distances between every pair of class prototypes, treating all classes as semantically equidistant. This design is well-suited to standard OOD

Table 5: Ablation of the separation loss and fixed prototypes based on HYPO. We report classification accuracy (%) on four OOD benchmarks. The best result in each column is highlighted in bold.

Method	PACS	Office-Home	VLCS	Terra-Incognita	Avg.
HYPO	88.0	71.7	78.2	51.7	72.4
HYPO without Separation Loss	87.3	71.0	77.6	50.5	71.6
HYPO using Fixed Prototypes	**88.2**	**71.9**	**78.9**	**52.4**	**72.9**

benchmarks where classes are semantically distinct (e.g., "dog" vs "elephant" in PACS), and where maximizing inter-class separation directly translates to better generalization. However, this assumption may be suboptimal for fine-grained classification tasks where some class pairs are inherently more similar than others (e.g., "Husky" vs "Malamute" versus "Husky" vs "Tabby cat"). In such settings, learnable or hierarchically-structured prototypes that preserve semantic relationships might outperform our equidistant construction. Extending our framework to incorporate semantic priors—for instance, by initializing prototypes from a class taxonomy or text embeddings, while keeping them fixed during training—is an interesting direction for future work.

C.4 Fixed Prototypes and the Separation Loss

To clarify the respective roles of the separation loss and fixed prototypes, we conduct an additional ablation study directly based on HYPO [Bai et al., 2024] across the four benchmarks used in Table 4.

We compare three variants: (1) HYPO: dynamic prototypes with the original prototype-separation loss. (2) HYPO without Separation Loss: dynamic prototypes without the separation loss. (3) HYPO using Fixed Prototypes: dynamic prototypes are replaced by our fixed regular-simplex prototypes. The separation loss is omitted because it becomes a parameter-independent constant once the prototypes are fixed.

As shown in Table 5, removing the separation loss while retaining dynamic prototypes decreases the average accuracy from 72.4% to 71.6%, with performance drops of 0.6–1.2 percentage points across all four benchmarks. These results show the importance of explicit separation regularization when the prototypes are dynamically learned. The fixed-prototype variant outperforms the original HYPO with its separation loss by 0.5 percentage points on average.

Together, these results show that the improvement from fixed prototypes cannot be attributed to removing the separation loss alone. The closed-form simplex construction directly places the class prototypes at maximally separated positions, replacing the need for explicit separation regularization.

D Pseudo Algorithm

The training procedure of our method is summarized in Algorithm 1.

E More Experimental Details

Software and hardware. Our implementation is based on the PyTorch framework (version 1.10), and all experiments were conducted on systems equipped with NVIDIA GeForce A40 GPUs.

Details of datasets. We provide the following descriptions of the datasets used in this work:

PACS [Li et al., 2017] is a widely used benchmark for OOD generalization. It contains 9,991 images of size 224×224, spread across four distinct domains, including photo, art painting, cartoon, and sketch, and covers a total of seven categories.

VLCS [Ghifary et al., 2015] comprises four domains: Caltech101, LabelMe, SUN09, and VOC2007. This dataset contains 10,729 images of size 224×224, covering five classes in total.

Office-Home [Venkateswara et al., 2017] has four domains, including art, clipart, product, and real, and comprises 15,588 images of size 224×224, distributed across 65 categories.

Algorithm 1: The training procedure of our method.

Input: Training dataset $\mathcal{D}$, deep neural network g, temperatures κ and τ, batch size b
Pre-compute: fixed class prototypes $\{\hat{\boldsymbol{\mu}}_k\}_{k=1}^{K}$ via Eq. (23)
for *epoch* $= 1, 2, \ldots$ **do**
  for *iter* $= 1, 2, \ldots$ **do**
    Sample a mini-batch $\mathcal{B} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{b}$
    Generate Cutout subviews $\mathbf{v}_i = T_{\text{cutout}}(\mathbf{x}_i; \Omega)$ for each $\mathbf{x}_i \in \mathcal{B}$
    `// compute` $\mathcal{L}_1$`: align subview v with class prototype`
    $\mathcal{L}_1 = -\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{B}} \log \frac{\exp[g(\mathbf{v}_i)^\top \hat{\boldsymbol{\mu}}_{y_i} / \kappa]}{\sum_{k=1}^{K} \exp[g(\mathbf{v}_i)^\top \hat{\boldsymbol{\mu}}_k / \kappa]}$
    `// compute` $\mathcal{L}_2$`: align fused` $g(\mathbf{v}) + g(\mathbf{x})$ `with class prototype`
    $\mathcal{L}_2 = -\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{B}} \log \frac{\exp[g(\mathbf{v}_i)^\top \hat{\boldsymbol{\mu}}_{y_i} / \kappa + (g(\mathbf{v}_i) + g(\mathbf{x}_i))^\top \hat{\boldsymbol{\mu}}_{y_i} / (2\tau)]}{\sum_{k=1}^{K} \exp[g(\mathbf{v}_i)^\top \hat{\boldsymbol{\mu}}_k / \kappa + (g(\mathbf{v}_i) + g(\mathbf{x}_i))^\top \hat{\boldsymbol{\mu}}_k / (2\tau)]}$
    `// compute overall loss and update`
    $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$
    Update network weights $\boldsymbol{\theta}$ by minimizing $\mathcal{L}$

Terra Incognita [Beery et al., 2018] consists of wildlife camera images from four distinct locations: location100, location38, location43, and location46. The dataset includes 24,788 images at a resolution of 224×224, grouped into 10 classes.

DomainNet [Peng et al., 2019] contains a total of 586,575 images across six distinct domains, including clipart, real, sketch, infograph, painting, and quickdraw, each containing the same 345 object categories (e.g., bracelet, plane, bird). This large-scale dataset is frequently used for multi-domain and cross-domain learning.

CIFAR-100 [Krizhevsky, 2009] consists of 60,000 color images of size 32×32, categorized into 100 classes. The training set has 50,000 images and the test set has 10,000 images, with 500 and 100 samples per class, respectively.

CIFAR-100-C [Hendrycks and Dietterich, 2019] is derived from CIFAR-100 by applying various corruptions, such as Gaussian noise, zoom blur, impulse noise, defocus blur, snow, brightness, contrast, elastic transform, fog, frost, Gaussian blur, glass blur, JPEG compression, motion blur, pixelate, saturate, shot noise, spatter, and speckle noise. This dataset provides a rigorous test bed for evaluating robustness to a wide range of image distortions.

Setups. Following prior works [Gulrajani and Lopez-Paz, 2021, Bai et al., 2024], we adopt ResNet50 [He et al., 2016] pre-trained on ImageNet1K [Deng et al., 2009] as the default backbone for all datasets, including PACS, VLCS, Office-Home, Terra Incognita, and DomainNet. We further append a two-layer MLP projection head [Riedmiller and Lernen, 2014] at the end of the backbone to extract features for classification, with the feature dimension fixed to $d = 512$ for all datasets. The training process runs for 10 epochs on DomainNet and 50 epochs on the remaining datasets, using stochastic gradient descent (SGD) with a momentum of 0.9, cosine scheduling [Loshchilov and Hutter, 2017], and a weight decay of 1×10^{-4}. Each domain is held out in turn as the unseen target domain for OOD evaluation, while the remaining domains serve as sources. For model selection, we follow DomainBed's training-domain validation criterion (`IIDAccuracySelectionMethod`) [Gulrajani and Lopez-Paz, 2021]. Each source domain is split into training and validation subsets, and hyperparameters are selected using the mean validation accuracy across source domains. No target-domain data are used for training or model selection. The search distribution in our experiments for the learning rate hyperparameter is: $lr \in \{0.005, 0.001, 0.0005, 0.0001, 0.00005\}$, and the temperature parameters τ and κ are selected from $\{0.01, 0.05, 0.1\}$. In addition, the Cutout transform is configured with a side length $s = 112$, which is half of the image length.

Table 6: Impact of varying the Cutout side length s on VLCS. We report OOD accuracy (%) per domain and overall average.

s	**Caltech101**	**LabelMe**	**SUN09**	**VOC2007**	**Avg.**
112	99.10	70.75	77.11	81.69	82.16
64	99.10	69.62	76.39	79.25	81.09
32	98.96	69.28	75.03	80.34	80.90
16	98.96	69.58	75.33	80.19	81.02

Table 7: Impact of varying the Cutout side length s on PACS. We report OOD accuracy (%) per domain and overall average.

s	**Art**	**Cartoon**	**Photo**	**Sketch**	**Avg.**
160	89.0	83.9	98.5	84.0	88.8
128	89.4	84.9	98.8	83.7	89.2
112	89.6	84.7	98.9	83.9	89.3
64	89.1	83.6	98.8	84.2	88.9
32	88.7	83.0	98.5	83.3	88.4
16	88.4	83.2	98.4	83.0	88.3

F Effect of Cutout Side Length

We investigate how the side length s (the size of the square region masked out by Cutout) influences OOD generalization. As shown in Table 6, we take the VLCS dataset as an example. Using a larger side length ($s = 112$) achieves the highest average accuracy of 82.16%, while smaller values (e.g., $s = 16$ or $s = 32$) lead to slightly lower performance. This suggests that masking a sufficiently large portion of the image more effectively challenges the model to focus on domain-invariant cues, thereby enhancing OOD generalization.

We furtherconduct the same sensitivity study on PACS with side lengths of 16, 32, 64, 112, 128, and 160, keeping all other settings unchanged. As shown in Table 7, among the evaluated sizes, $s = 112$ achieves the highest average accuracy of 89.3%, exceeding $s = 64$, $s = 32$, and $s = 16$ by 0.4, 0.9, and 1.0 percentage points, respectively. Performance remains similar at $s = 128$, with an average of 89.2%, only 0.1 percentage points below $s = 112$. Further increasing the side length to $s = 160$ reduces the average to 88.8%. The per-domain results are not uniformly monotonic: Cartoon achieves its highest accuracy at $s = 128$, whereas Sketch performs best at $s = 64$. These results are consistent with a trade-off between the complementary supervision provided by masking and the preservation of label-relevant information in the subview V.

G Generalization on Corruption Benchmarks

In practice, distribution shifts between training and test domains can also arise from input corruptions. Recent studies [Maheronnaghsh and Alvanagh, 2024, Ye et al., 2024] identify the corrupted CIFAR datasets as prototypical spurious-correlation benchmarks, since class labels covary with specific corruption patterns (noise, blur, etc.). We therefore evaluate our method under this scenario.

We use CIFAR-100 [Krizhevsky, 2009] as the in-distribution data and CIFAR-100-C [Hendrycks and Dietterich, 2019] as the OOD test data, which applies 19 diverse common corruptions (e.g., noise, blur, compression) to CIFAR-100. We train a ResNet-18 model exclusively on the clean CIFAR-100 images, with no exposure to corrupted samples during training. Specifically, we train from scratch for 500 epochs with an initial learning rate of 0.5, a cosine learning-rate schedule, and a batch size of 512. An MLP projection head with one hidden layer produces 512-dimensional features, and we set both temperatures κ and τ to 0.1.

As shown in Fig. 6, our method brings significant improvements on challenging cases such as Gaussian noise, raising OOD accuracy from 51.42% to 57.30% ($+5.88\%$), with an overall improvement from 64.65% to 67.64% ($+2.99\%$) over HYPO. Compared to HYPO (65.40%), our approach gains over $+2.2\%$, reflecting better resilience to diverse corruptions including Shot noise, Fog, and Elastic

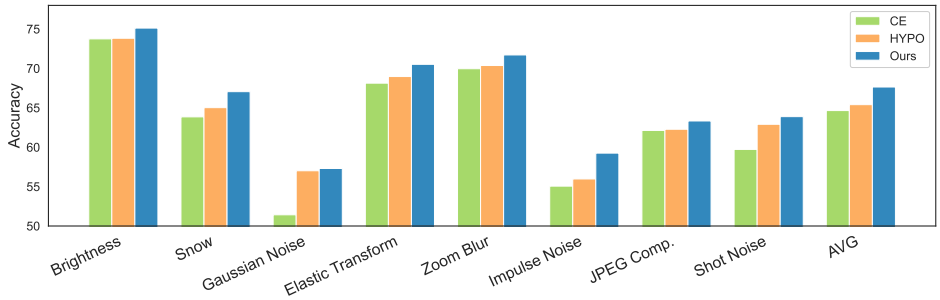


Figure 6: OOD generalization performance on selected corruption types from CIFAR-100-C. Models are trained on CIFAR-100 (ID) and tested on each corruption type in CIFAR-100-C, measuring OOD test accuracy under varied corruptions.

Table 8: Main results for verifying OOD generalization performance on the 19 different covariate shift datasets. We train on CIFAR-100 as ID, using CIFAR-100-C as the OOD test dataset. Acc. denotes the accuracy on the OOD test set.

Method	Corruptions	Acc.	Corruptions	Acc.	Corruptions	Acc.	Corruptions	Acc.	Corruptions	Acc.
CE	Gaussian noise	51.42	Zoom blur	69.96	Impulse noise	55.06	Defocus blur	74.20	Snow	63.85
HYPO	Gaussian noise	57.01	Zoom blur	70.37	Impulse noise	55.97	Defocus blur	73.71	Snow	65.02
Ours	Gaussian noise	57.30	Zoom blur	71.71	Impulse noise	59.25	Defocus blur	75.46	Snow	67.05
CE	Brightness	73.75	Contrast	73.12	Elastic transform	68.13	Fog	73.20	Frost	67.16
HYPO	Brightness	73.83	Contrast	73.42	Elastic transform	68.97	Fog	72.85	Frost	66.50
Ours	Brightness	75.12	Contrast	74.82	Elastic transform	70.52	Fog	74.45	Frost	68.09
CE	Gaussian blur	74.21	Glass blur	20.56	JPEG compression	62.13	Motion blur	69.31	Pixelate	71.61
HYPO	Gaussian blur	73.66	Glass blur	23.85	JPEG compression	62.28	Motion blur	68.82	Pixelate	70.74
Ours	Gaussian blur	75.50	Glass blur	37.40	JPEG compression	63.33	Motion blur	70.39	Pixelate	72.17
CE	Saturate	69.22	Shot noise	59.72	Spatter	67.67	Speckle noise	59.16	**Average**	64.65
HYPO	Saturate	69.48	Shot noise	62.91	Spatter	67.88	Speckle noise	62.91	**Average**	65.40
Ours	Saturate	71.20	Shot noise	63.88	Spatter	70.85	Speckle noise	62.97	**Average**	**67.64**

Table 9: Classification accuracy (%) on clean CIFAR-100 (ID) and CIFAR-100-C (OOD). All models are trained only on clean CIFAR-100. The best result in each column is highlighted in bold.

Method	CIFAR-100 ID Acc.	CIFAR-100-C OOD Acc.
ERM/CE	74.99	64.65
ERM + Cutout	75.60	64.87
HYPO	77.85	65.40
Ours	**78.28**	**67.64**

transform. The consistent improvements across all 19 corruptions confirm that our method yields more corruption-robust representations than standard cross-entropy training and the HYPO baseline. In addition, these gains on the smallest-resolution dataset in our study (32×32) suggest that the proposed method does not introduce bias when sample size is limited; the method remains stable and advantageous even in low-resolution regimes. Complete results are reported in Table 8.

We further compare the ID result on CIFAR-100. As shown in Table 9, our method achieves the best ID accuracy of 78.28% and the best OOD accuracy of 67.64%. Adding Cutout directly to ERM improves ID accuracy by 0.61 percentage points but improves CIFAR-100-C accuracy by only 0.22 points. Our method improves over its direct HYPO baseline by 0.43 points on CIFAR-100 and by 2.24 points on CIFAR-100-C.

H Generality Beyond Cutout

The decomposition in Section 3.2 depends on the conditional independence of the subview V and the label Y given the full image X. A sufficient construction is $V = T(X; \Omega)$, where the transformation randomness Ω is independent of Y conditional on X. This gives $p(v \mid x, y) = p(v \mid x)$: once X is known, V provides no additional label information. Hence $I(V; Y \mid X) = 0$, and Eq. (8) reduces

Table 10: OOD generalization performance with Random Erasing on PACS and VLCS. Avg. denotes the mean across target domains. Best result is highlighted in bold.

PACS					
Method	Art	Cartoon	Photo	Sketch	Avg.
HYPO	87.2	82.3	98.0	**84.5**	88.0
HYPO + Random Erasing	87.4	82.5	98.1	84.3	88.1
Ours (Cutout)	**89.6**	**84.7**	**98.9**	83.9	**89.3**
Ours + Random Erasing	89.3	84.4	98.8	83.5	89.0
VLCS					
Method	Caltech	LabelMe	SUN	VOC	Avg.
HYPO	98.1	65.3	73.1	76.3	78.2
HYPO + Random Erasing	98.2	65.8	73.6	76.5	78.5
Ours (Cutout)	**99.1**	**70.8**	**77.1**	**81.7**	**82.2**
Ours + Random Erasing	99.0	69.8	76.7	80.9	81.6

Table 11: OOD classification accuracy (%) on PACS with a ViT-B/16 backbone. Avg. denotes the mean across target domains. The best result in each column is highlighted in bold.

Method	Art	Cartoon	Photo	Sketch	Avg.
ERM	87.6	82.9	98.4	82.4	87.8
Ours	**90.2**	**85.3**	**99.1**	**85.0**	**89.9**

exactly to

$$I(X;Y) = I(V;Y) + I(X;Y \mid V). \tag{31}$$

The identity therefore applies to label-independent information-removal transformations beyond Cutout.

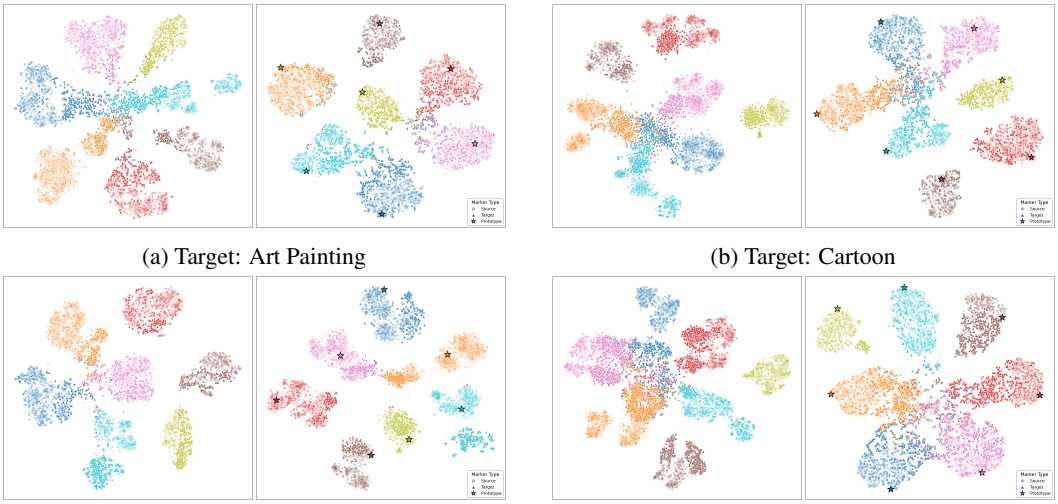
Evaluation with Random Erasing. We replace Cutout with Random Erasing [Zhong et al., 2020] and evaluate both its use as an ordinary augmentation for HYPO [Bai et al., 2024] and its integration into our framework on PACS [Li et al., 2017] and VLCS [Ghifary et al., 2015]. We use `torchvision.transforms.RandomErasing` with `p=1.0`, `scale=(0.02, 0.33)`, `ratio=(0.3, 3.3)`, and `value=0`, corresponding to an erased-area fraction of 0.02–0.33, an aspect-ratio range of 0.3–3.3, and zero filling. The transform is applied at the same preprocessing stage as Cutout to generate V for each training image, while retaining X unchanged. HYPO + Random Erasing trains on V as an ordinary augmentation; Ours + Random Erasing uses both X and V with the chain-rule objective in Eq. (22) and fixed simplex prototypes. All other training settings and data splits follow the original experiments. Erasing is disabled at inference and introduces no additional inference cost.

As shown in Table 10, adding Random Erasing to HYPO improves average accuracy by 0.1 percentage points on PACS and 0.3 on VLCS. Integrating the same transformation into our framework achieves 89.0% and 81.6%, respectively, exceeding HYPO + Random Erasing by 0.9 and 3.1 percentage points. Although Random Erasing is slightly weaker than Cutout, these results show that our chain-rule benefit is not tied to a fixed square mask.

I Evaluation with a Vision Transformer Backbone

We evaluate ERM and our method on PACS using ViT-B/16 [Dosovitskiy et al., 2021] to assess performance with a transformer backbone. Both backbones are initialized with ImageNet-1K pretrained `ViT_B_16_Weights.IMAGENET1K_V1` weights. We follow the evaluation protocol in Appendix E, holding out each domain in turn for OOD evaluation and using only source-domain validation splits for model selection.

As shown in Table 11, our method improves average accuracy from 87.8% to 89.9% ($+2.1$ percentage points), with improvements on all four target domains. These results show that our method also improves OOD generalization with a ViT-B/16 backbone on PACS. A future direction is to integrate



(a) Target: Art Painting

(b) Target: Cartoon

(c) Target: Photo

(d) Target: Sketch

Figure 7: t-SNE visualization of learned feature representations on the PACS dataset. We compare the ERM baseline (left) with Ours (right), with each of the four domains held out in turn for OOD evaluation. Different colors correspond to the 7 semantic classes. Marker shapes indicate the data source: circles denote source domains, triangles denote the unseen target domain, and large stars represent the fixed class prototypes used in our method.

our MI-based Cutout objective with ViT-based vision-language models, which have been successfully applied to various visual recognition tasks [Shu et al., 2023, Peng et al., 2025a, 2026a].

J Visualization of Learned Representations

To intuitively verify the effectiveness of our method in learning domain-invariant and discriminative features, we provide t-SNE visualizations of the feature embeddings on the PACS dataset. Fig. 7 compares the feature distributions learned by ERM (left column) and Ours (right column), with Art Painting, Cartoon, Photo, and Sketch each held out in turn for OOD evaluation. In these visualizations, different colors represent different semantic classes, while marker shapes denote the data source: circles for source domains and triangles for the unseen target domain. In our method's plots, stars indicate the positions of the fixed class prototypes.

The representations learned by ERM (left column) exhibit two issues: (i) loose clustering—intra-class distributions are spread out, indicating weak discriminative power; and (ii) domain misalignment—there is a noticeable separation between source (circles) and target (triangles) features within the same class, particularly under large distribution shifts (e.g., Sketch and Art Painting), confirming that ERM struggles to bridge the domain gap. By contrast, our method (right column) achieves (i) high intra-class compactness—features from all domains are tightly clustered around their respective class prototypes, validating the effectiveness of the fixed-prototype strategy and the $I(V;Y)$ term; and (ii) robust domain alignment—source and target features are well-mixed within each cluster, even on the most challenging Sketch domain, demonstrating that our MI maximization objective drives the encoder to capture domain-invariant structural cues rather than domain-specific superficial statistics.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our contributions: (i) identifying the failure mode of naively integrating Cutout into the hyperspherical prototype framework, (ii) the proposed method that addresses this failure, and (iii) empirical gains on five standard OOD generalization benchmarks. The scope is consistent with the experimental and theoretical results presented in the main paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation, as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our approach in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are stated with explicit assumptions in the main paper, and full proofs are provided in the appendix with proper cross-referencing.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All training details, hyperparameters, optimizer choices, and architecture specifications are provided in the experimental section and the appendix.

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All datasets used are publicly available. The code and reproduction instructions have not yet been released, as code cleanup and documentation are still in progress.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and test details, including data splits, optimizer (SGD), and hyperparameter selection, are reported in the experimental section and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean accuracies along with error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources (e.g., GPU type) are reported in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. We use only publicly available benchmark datasets and do not involve human subjects, sensitive data, or potentially harmful applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational research on out-of-distribution generalization for image classification. It does not have a direct path to specific deployments and we do not foresee immediate societal impacts beyond improving the robustness of image classifiers.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained generative models, scraped datasets, or other high-risk assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (PACS, Office-Home, VLCS, Terra-Incognita, DomainNet) and baseline methods are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology of this research. Any use was limited to minor writing or formatting assistance, which does not require declaration per the NeurIPS LLM policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.