Domain Generalization by Mutual-Information Regularization with Pre-trained Models

Junbum Cha
Kakao Brain
junbum.cha@kakaobrain.com

Kyungjae Lee
Chung-Ang University
kyungjae.lee@ai.cau.ac.kr

Sungrae Park
Upstage AI Research
sungrae.park@upstage.ai

Sanghyuk Chun
NAVER AI Lab
sanghyuk.c@navercorp.com

Abstract

Domain generalization (DG) aims to learn a generalized model to an unseen target domain using only limited source domains. Previous attempts to DG fail to learn domain-invariant representations only from the source domains due to the significant domain shifts between training and test domains. Instead, we re-formulate the DG objective using mutual information with the oracle model, a model generalized to any possible domain. We derive a tractable variational lower bound via approximating the oracle model by a pre-trained model, called Mutual Information Regularization with Oracle (MIRO). Our extensive experiments show that MIRO significantly improves the out-of-distribution performance. Furthermore, our scaling experiments show that the larger the scale of the pre-trained model, the greater the performance improvement of MIRO. Source code is available at https://github.com/kakaobrain/miro.

1 Introduction

Emerging studies on the generalizability of deep neural networks (DNNs) have revealed that the existing models, that assumes independent and identically distributed (i.i.d.) training and test distribution, are not robust to significant distribution shifts between training and test distribution, *e.g.*, backgrounds [1, 2], geographic distribution [3], demographic statistics [4, 5], textures [6, 7], or day-to-night shifts [8, 9]. Domain generalization (DG) problem aims to learn domain-agnostic representations by accessing multiple source *domains* (*e.g.*, photo, sketch, cartoon) during training. The trained model on multiple source domains is evaluated on an unseen domain (*e.g.*, art painting) to measure the robustness against distribution shifts. The existing DG approaches have tried to learn invariant features across multiple *domains* by minimizing feature divergences between the source domains [10–16], normalizing domain-specific gradients based on meta-learning [17–21], robust optimization [22–25], or augmenting source domain examples [26–32]. However, recent studies [33, 34] have shown that simple baselines without learning invariant features are comparable or even outperform the existing DG methods on the diverse DG benchmarks with a fair hyperparameter selection protocol when a model becomes larger (*e.g.*, from ResNet-18 to ResNet-50 [35]). We presume that it is because training and test distributions differ too significantly to learn domain-invariant features by the training distributions only.

Instead of learning domain-invariant features, we let a model learn similar features to "oracle" representations, *i.e.*, an optimal model generalized to *any* domain. In particular, we re-formulate the domain generalization problem by maximizing the mutual information between the oracle model representations and the target model representations while preserving the training loss on source

domains. However, the oracle model is not achievable in practice. Hence, we use a large pre-trained model (*e.g.*, ImageNet [36] pre-trained ResNet-50 [35]) as an approximation of the oracle model. With this approximation, we derive a tractable variational lower bound of the proposed maximization problem, named Mutual Information Regularization with Oracle (MIRO). At a high level, our MIRO objective consists of two objectives: an original target task (*i.e.*, an ERM objective) and a regularization term between the pre-trained model and the current target model. Note that the standard DomainBed benchmark [33] uses the ImageNet pre-trained ResNet-50 as the initialization of a DG method, thus, we use the pre-trained ResNet as the initialization and the approximation of the oracle model at the same time.

While a naive fine-tuning approach of a large pre-trained model can harm the robustness against distribution shifts [37–40], our proposed algorithm remarkably improves the robustness against unseen domains during fine-tuning in a plug-and-play manner to any scale of the backbone model and datasets. In our experiment, we observe that the naive fine-tuning of a larger pre-trained model can fail to provide better performances, even though the larger pre-trained model is trained with more data and domains. For example, ERM with the ResNet pre-trained on ImageNet (trained with 1.3M images) shows 64.2% of averaged accuracy, while ERM with the ViT pre-trained on CLIP (trained with 400M image-caption pairs) shows 61.1%. On the other hand, we show that our method can significantly improve the average DG performances with backbone models at different scales, *e.g.*, ImageNet pre-trained ResNet (64.2% $\rightarrow$ 65.9%), 400M image-text pre-trained CLIP [37] (61.1% $\rightarrow$ 73.7%) and Instagram 3.6B pre-trained RegNet (SWAG) [41] (68.0% $\rightarrow$ 74.1%). Especially, we observe that the pre-trained knowledge by larger pre-trained models, such as SWAG and CLIP, are more effective to learn domain generalized features than the ImageNet pre-trained model: MIRO with the ViT pre-trained on CLIP outperforms MIRO with the ResNet pre-trained on ImageNet in contrast to the naive fine-tuning. Furthermore, our feature-level regularization method is easily combined with the existing parameter space ensemble methods [34, 39] (74.1% $\rightarrow$ **77.3%** average DG accuracy by combining with SWAD [34] and pre-trained RegNet).

Our contribution is as follows: (1) We re-formulate the DG objective by mutual information with the oracle model. Then, we approximate the oracle model by a large pre-trained model to derive a tractable approximation of the target objective. We propose Mutual Information Regularization with Oracle (MIRO) to solve our objective. (2) We analyze the pre-trained models in terms of the mutual information with the oracle model. Our analysis shows that naive fine-tuning of pre-trained models can harm the mutual information with the oracle model, on the other hand, MIRO shows high mutual information with the oracle. (3) We compare MIRO with state-of-the-art DG methods on DomainBed. MIRO outperforms all methods in all settings, including varying optimizers and pre-trained models. We also provide extensive analysis to understand MIRO. For example, we observe that MIRO shows stronger domain generalization performances with larger pre-trained models, such as SWAG [41] or CLIP [37].

2 Methods

In this section, we first re-formulate the objective for the out-of-domain generalization by introducing an oracle model. Then, we derive a tractable variational bound of the objective by approximating the oracle model to the pre-trained model. The final form consists of the empirical risk and the mutual information regularization by querying the approximated oracle, named Mutual Information Regularization with Oracle (MIRO). We empirically validate our approximation by mutual information between the oracle model and large pre-trained models.

2.1 Mutual information regularization with oracle

The main idea of the proposed method is to guide the learning process using oracle representations of training datasets. In general, the problem of domain generalization (DG) is to find a model that minimizes an expected loss of *any* domain by using training datasets from only partial domains, which are called source domains. Many existing methods minimize an empirical loss averaged over source domains. More specifically, suppose that training samples $\{\mathcal{S}_d\}_{d=1}^m$ are given in m domains and we consider a hypothesis set $\mathcal{H}$ for optimization. Then, many existing DG frameworks can be formulated as follows:

$$\bar{h} = \arg\min_{h \in \mathcal{H}} \sum_{d=1}^{m} \mathcal{E}_{\mathcal{S}_d}(h), \tag{1}$$

where d indicates an individual source domain and $\mathcal{E}_{\mathcal{S}_d}$ is an empirical loss over the source domain d. Note that majority of existing DG methods can be interpreted as the variant of Equation (1). For example, if we choose a simple cross-entropy loss for $\mathcal{E}_{\mathcal{S}_d}$, then Equation (1) becomes "ERM" baseline used in [33][1]. Otherwise, $\mathcal{E}_{\mathcal{S}_d}$ can be formulated as a regularized ERM, such as IRM [22] or CORAL [13]. However, the formulation (1) still suffers from learning domain-invariant representations using only partial domains when target distribution differs significantly from the training distribution. For example, the state-of-the-art CORAL method shows inconsistent out-of-domain accuracies across domains in DomainNet [42]. While CORAL achieves about 50% top-1 accuracy on four *easy* domains (59.2% for Clipart, 46.6% for Painting, 59.8% Real images, 50.1% for Sketches), it only shows 13.4% for QuickDraw and 19.7% for Infographics where the domains show the significant distribution shift comparing to others.

To alleviate this issue, we re-formulate the DG problem by employing *oracle* representations of source domains. Here, we define an oracle model as a model that can be generalized to *any* possible domain, not only for the source domains. We define a model as a composition of a feature extractor f and a classifier g on the feature space where the whole classifier h can be written as $h = f \circ g$. Then, let f^* be a feature extractor of the oracle model. We first start from a strong assumption: we may assume that f^* is accessible during the training phase. Then, we can obtain additional information from f^* by querying the oracle representations of training samples in the source domains. By using the oracle representations, we can guide the learning process of a target model by maximizing mutual information between oracle representations and target ones. We formulate the proposed oracle-guided DG framework as follows:

$$\begin{aligned} \max_{h} \quad & I(Z_{f^*}; Z_f) \\ \text{s.t.} \quad & \mathcal{E}_{\mathcal{S}}(h) - \mathcal{E}_{\mathcal{S}}(\bar{h}) \leq \epsilon, \end{aligned} \tag{2}$$

where Z_{f^*} is a random feature extracted by f^* and Z_f is a random feature extracted by a target model f. $I(Z_{f^*}; Z_f)$ is mutual information between Z_{f^*} and Z_f, and $\mathcal{E}_{\mathcal{S}}(\cdot) = \sum_{d=1}^{m} \mathcal{E}_{\mathcal{S}_d}(\cdot)$. The inequality constraint ensures the performance of the target model on the source domains. We believe that the guidance of the oracle is beneficial to learning robust representations. Furthermore, maximizing the mutual information will inhibit the target model from learning domain-specific features in the limited source domains.

Unfortunately, the oracle feature extractor f^* is not accessible in practice. Instead, we approximate the oracle feature extractor by using a pre-trained model f^0. Our assumption is that a model pre-trained on large-scale diverse datasets, such as ImageNet [36], partially contains information of diverse domains. In practice, we choose f^0 as the ImageNet pre-trained ResNet-50 [35], the standard initialization choice for evaluating DG algorithms [33]. We also consider models trained by larger diverse datasets, such as CLIP [37] (trained with 400M web crawled image-text pairs) and SWAG [41] (trained with 3.6B noisy image-hashtag pairs crawled from Instagram). Although using CLIP and ResNet is not a fair comparison to the existing DG benchmark, here, we emphasize that naive fine-tuning of large pre-trained models leads to inferior generalizability to extreme distribution shifts at test time [37–40]. In our experiments, we also observe a similar observation: naive fine-tuning of CLIP ResNet shows an inferior DG performance (61.1%) than ERM (64.2%).

Through the approximation of the oracle model, we derive a tractable variational bound of our objective function (2). We assume a pre-trained model f^0 is located near f^* in terms of distance equipped on the hypothesis set of feature extractors and it can provide approximated representation of f^*. Under this assumption, we can obtain a tractable objective function by deriving an approximated lower bound of the mutual information. We first derive the variational lower bound of the mutual information as follows:

$$\begin{aligned} I(Z_{f^*}; Z_f) =& \mathbb{E}_{Z_{f^*}, Z_f}\left[\log \frac{q(Z_{f^*} \mid Z_f)}{p(Z_{f^*})}\right] + KL(p(Z_{f^*} \mid Z_f) \| q(Z_{f^*} \mid Z_f)) \\ \geq& \mathbb{E}_{Z_{f^*}, Z_f}[\log q(Z_{f^*} \mid Z_f)] + H(Z_{f^*}), \end{aligned} \tag{3}$$

[1]Note that the terminology ERM can be unfair because other methods also minimize "empirical risk" but with different loss designs. We use the terminology "ERM" to indicate the cross-entropy baseline as suggested by Gulrajani and Lopez-Paz [33].

Algorithm 1: Mutual Information Regularization with Oracle (MIRO)

Input: feature extractor f, classifier g, mean encoder μ, variance encoder Σ, regularization coefficient λ, batch size N.
Init: initialize f to pre-trained feature extractor f^0.
Output: learned feature extractor f and learned classifier g.
for sampled mini-batch $(\mathbf{x}, \mathbf{y})$ **do**
  $\mathbf{z}_f = f(\mathbf{x})$
  $\mathbf{z}_{f^0} = f^0(\mathbf{x})$
  $\mathcal{L} = \frac{1}{N}\sum_i^N \left[\texttt{CrossEntropy}\left(g(z_f^i), y^i\right) + \lambda\left(\log\left|\Sigma(z_f^i)\right| + \|z_{f^0}^i - \mu(z_f^i)\|^2_{\Sigma(z_f^i)^{-1}}\right)\right]$
  update f, g, μ, Σ to minimize $\mathcal{L}$
end

where q is the variational distribution with a mild regularity condition. More detailed derivation can be found in Barber and Agakov [43]. Then, we approximate the expectation in Equation (3) by using f^0.

$$\begin{aligned} I(Z_{f^*}; Z_f) &\geq \mathbb{E}_{Z_{f^*}, Z_f}\left[\log q(Z_{f^*} \mid Z_f)\right] + H(Z_{f^*}) \\ &\geq \mathbb{E}_{Z_{f^0}, Z_f}\left[\log q(Z_{f^0} \mid Z_f)\right] - C d_{2,\infty}(f^*, f^0) + H(Z_{f^*}), \end{aligned} \tag{4}$$

where C is a constant and $d_{2,\infty}(f^*, f^0) := \sup_x \|f^*(x) - f^0(x)\|_2$. Note that $d_{2,\infty}$ is a proper metric on the hypothesis set of feature extractor. The last inequality of Equation (4) is derived by using the first-order Taylor expansion and assuming the regularity condition of q (See Appendix). We would like to note that the inequality is tight enough due to Taylor's theorem. In other words, equality condition of the last inequality of Equation (4) is $d_{2,\infty}(f^*, f^0) = 0$. Hence, $d_{2,\infty}(f^*, f^0)$ represents the effect of the pre-trained model f^0 on the approximation of the lower bound. Intuitively speaking, the lower bound shows that the smaller $d_{2,\infty}(f^*, f^0)$ is, the tighter the gap between the true lower bound and approximated one is. In summary, the mutual information between Z_{f^*} and Z_f can be maximized by maximizing the term $\mathbb{E}_{Z_{f^0}, Z_f}[\log q(Z_{f^0} \mid Z_f)]$.

Finally, to consider the constraint term, we introduce the Lagrangian method to Equation (2), then we can derive an objective function from Equation (4):

$$R(h) = \mathbb{E}_{Z_{f^0}, Z_f}[\log q(Z_{f^0} \mid Z_f)] - \beta \mathcal{E}_{\mathcal{S}}(h), \tag{5}$$

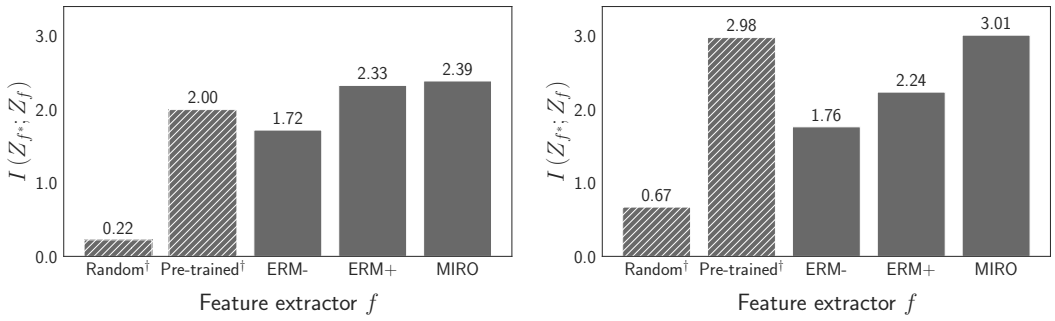
where β indicates the Lagrangian multiplier. Note that the entropy of Z_{f^*} and $d_{2,\infty}(f^*, f^0)$ are omitted, since they are independent to our optimization target $h = f \circ g$. In the implementation, we model the variational distribution as a Gaussian distribution with mean vector $\mu(Z_f)$ and covariance matrix $\Sigma(Z_f)$ and replace the multiplier β with the regularization coefficient λ. Then, our final loss function becomes:

$$\textbf{(MIRO)} \quad \mathcal{L}(h) = \mathcal{E}_{\mathcal{S}}(h) + \lambda \mathbb{E}_{Z_{f^0}, Z_f}\left[\log|\Sigma(Z_f)| + \|Z_{f^0} - \mu(Z_f)\|^2_{\Sigma(Z_f)^{-1}}\right], \tag{6}$$

where $\|x\|_A = \sqrt{x^\intercal A x}$ and constants independent on h are omitted. Then, we optimize the loss function using a stochastic gradient method. The entire learning process is summarized in Algorithm 1. In the following sections, we empirically justify our approximation of f^* and explain implementation details for the mean and variance encoders of the Gaussian distribution q.

2.2 Mutual information with the oracle model

Here, we empirically show how our approximation by pre-trained models is close to the oracle model and how our algorithm is effective to learn representations having high mutual information to the underlying oracle model. More specifically, we compare mutual information between the candidate models and the oracle model on PACS dataset [45]. For the candidate models, we choose "Random" networks (*i.e.*, randomly initialized weights without any training), "Pre-trained" networks from ImageNet 1.3M and Instagram 3.6B (*i.e.*, pre-trained ResNet-50 and RegNetY-16GF, respectively), fine-tuned models from "Random" network and "Pre-trained" networks (we name each model as "ERM−" and "ERM+", respectively), and models trained by our algorithm (MIRO). Since the *true*



(a) ResNet-50 (b) RegNetY-16GF

Figure 1: **Mutual information $I(Z_{f^*}; Z_f)$ with oracle model.** The mutual information is estimated by MINE [44] in PACS. Oracle model is trained using all of the four domains. *Random* and *Pretrained* indicate random and pre-trained model initialization, respectively. *ERM-* and *ERM+* are trained from random and pre-trained model initialization, respectively. † indicates models without fine-tuning.

oracle model (*i.e.*, a model invariant to *any* domain) is not achievable in practice, we train an oracle model by directly optimizing a model on the entire domains in PACS dataset. We train two oracle models using ResNet-50 and RegNetY-16GF backbones, where the average validation accuracies across all domains are 97.2% and 98.4%, respectively. We estimate mutual information between models by mutual information neural estimation (MINE) [44]. We describe the full details of the experiments in Appendix.

Figure 1 illustrates the empirical mutual information between the candidate models and the oracle model. In the figures, we first observe that the larger and more powerful pre-trained backbone ("Pre-trained" in Figure 1b) shows higher mutual information than the smaller backbone ("Pre-trained" in Figure 1a). Both pre-trained models consistently outperform "Random" in mutual information regardless of the backbone models. Our observations imply that a larger and stronger model is closer to the oracle model in terms of mutual information. Similarly, we observe that ERM+ always shows high mutual information than ERM−. However, interestingly, in Figure 1b, we observe that fine-tuning significantly harms mutual information of the pre-trained model ("Pre-trained" vs. "ERM+") when the pre-trained model becomes larger and more powerful. Our observation is aligned to the same line to the previous studies on fine-tuning of large models [37–40]. Lastly, in both scenarios of ImageNet pre-trained ResNet (Figure 1a) and SWAG trained RegNet (Figure 1b), our MIRO shows the highest mutual information with the oracle model.

2.3 Features and encoders design

Here, we describe the design choices for the feature extractor f for MIRO. We employ a multi-level structure to cope with the task shift between pre-training and fine-tuning, and simple encoder designs for the increased feature size.

Multi-scale features. One can use the last high-level features for our regularization. However, high-level features can include pre-training task-related information, often irrelevant to the target task. Instead, we use the intermediate outputs by each model block, *i.e.*, stem output, block 1, 2, 3, and 4 for ResNet [35] and RegNet [46]. For ViT-B [47], we use stem output, block 3, 6, 9, and 12 features.

Design of the mean and variance encoders. The multi-level structure increases the feature size, resulting in a computational cost increase. We alleviate the issue by employing simple yet effective architectures, identity function for the mean encoder and a bias-only model with diagonal covariance for the variance encoder. We also tested more complicated architectures, but only computational cost was increased without performance improvement.

3 Experiments

3.1 Experiment setups and implementation details

Benchmark datasets. Following the previous DG studies [34, 33], we evaluate the proposed method on the five benchmark datasets, namely `PACS` [45] (4 domains, 7 classes, and $9,991$ examples), `VLCS` [48] (4 domains, 5 classes, and $10,729$ examples), `OfficeHome` [49] (4 domains, 65 classes, and $15,588$ images), `TerraIncognita` [50] (4 domains, 10 classes, and $24,788$ examples), and `DomainNet` [42] (6 domains, 345 classes, and $586,575$ examples).

Evaluation protocols. We use DomainBed [33] as the testbed for DG tasks. Because the original DomainBed requires very heavy computation resources, we apply minor modifications by increasing total training steps and reducing hyperparameter search space, following Cha *et al*. [34]. All performance scores are evaluated by *leave-one-out cross-validation*, where averaging all cases that use a single domain as the target (test) domain and the others as the source (training) domains. We leave 20% of source domain data for validation. Every experiment is repeated three times by different trial seeds.

Implementation details. We use ResNet-50 [35] pre-trained in the ImageNet [36] dataset as default. The model is optimized using Adam [51] optimizer. A mini-batch contains all domains and 32 examples per domain. We tune the λ in [1.0, 0.1, 0.01, 0.001] using training-domain validation set following DomainBed [33]. The other hyperparameters, such as batch size, learning rate, dropout rate, and weight decay, are tuned in similar search space proposed in Cha *et al*. [34]. We provide full implementation details and the hyperparameter search protocol in Appendix.

3.2 Main results

Comparison with domain generalization methods. We provide exhaustive out-of-domain performance comparisons on five DG benchmarks in Table 1. Compared to ERM, the proposed mutual information regularization significantly improves performance on every benchmark dataset, resulting in +1.7pp average improvement (+1.2pp in `PACS`, +1.7pp in `VLCS`, +2.9pp in `OfficeHome`, +2.6pp in `TerraIncognita`, and +0.3pp in `DomainNet`). Compared with the state-of-the-art methods, MIRO achieves the best performances in all benchmarks, except `PACS`. Especially, MIRO remarkably outperforms previous state-of-the-arts: +1.3pp in `OfficeHome` (mDSDI [59]; $69.2\% \rightarrow 70.5\%$) and +1.8pp in `TerraIncognita` (SagNet [29]; $48.6\% \rightarrow 50.4\%$). Considering the experiment setup with 5 datasets and 22 target domains, the experiment results demonstrate the effectiveness of MIRO to the diverse visual data types.

The second part of Table 1 shows the performance with stochastic weight averaging densely (SWAD) [34], a state-of-the-art optimizer for DG by seeking flat minima. Since SWAD is an orthogonal direction to MIRO, we also evaluate the combination of MIRO and SWAD. As shown in the table, the combination of MIRO and SWAD achieves the best performance in all datasets, resulting in +0.8pp average improvement compared to the previous best results.

In the last part of Table 1, we push the limits of the out-of-domain performance by employing a large-scale backbone, RegNetY-16GF pre-trained by SWAG [41]; a weakly-supervised pre-trained model using 3.6 billion noisy Instagram images and hashtags. As shown in our previous study on mutual information with the oracle model, the pre-trained RegNet has higher mutual information to the oracle model than ImageNet pre-trained ResNet (Figure 1). In the experiments, we first observe that the improvement gap by MIRO becomes remarkably large compared to the ResNet pre-trained model (from +1.7pp to +6.1pp). We presume that this significantly large gap originated from the negative effect of the naive fine-tuning as observed by previous works [37–40] and our study (Figure 1b). As shown in Figure 1b, MIRO keeps mutual information with the oracle model high, resulting in remarkable performance gains on large-scale models. We further explore the effect of the scalability of pre-trained models in the later section. Finally, by combining MIRO with RegNet backbone and SWAD, we achieve the best domain generalization results (77.3%) on our evaluation benchmark.

MIRO with various pre-trained models. In this subsection, we investigate the robustness of the proposed method to the choice of pre-trained models. In Table 2, we explore the performance changes of MIRO by varying pre-training datasets, methods, and backbones. From the pre-training

Table 1: **Comparison with domain generalization methods.** Out-of-domain accuracies on five domain generalization benchmarks are shown. We highlight the **best results** in bold. The results marked by †, ‡ are the reported numbers from Gulrajani and Lopez-Paz [33] and Cha *et al*. [34], respectively. The results of Fish, SelfReg, and mDSDI are the reported ones from each paper. Average accuracies and standard errors are reported from three trials.

Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
MMD† [12]	84.7±0.5	77.5±0.9	66.3±0.1	42.2±1.6	23.4±9.5	58.8
Mixstyle‡ [28]	85.2±0.3	77.9±0.5	60.4±0.3	44.0±0.7	34.0±0.1	60.3
GroupDRO† [24]	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
IRM† [22]	83.5±0.8	78.5±0.5	64.3±2.2	47.6±0.8	33.9±2.8	61.6
ARM† [21]	85.1±0.4	77.6±0.3	64.8±0.3	45.5±0.3	35.5±0.2	61.7
VREx† [23]	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
CDANN† [15]	82.6±0.9	77.5±0.1	65.8±1.3	45.8±1.6	38.3±0.3	62.0
DANN† [11]	83.6±0.4	78.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	62.6
RSC† [52]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
MTL† [53]	84.6±0.5	77.2±0.4	66.4±0.5	45.6±1.2	40.6±0.1	62.9
Mixup† [54–56]	84.6±0.6	77.4±0.6	68.1±0.3	47.9±0.8	39.2±0.1	63.4
MLDG† [17]	84.9±1.0	77.2±0.4	66.8±0.6	47.7±0.9	41.2±0.1	63.6
Fish [25]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
ERM‡ [57]	84.2±0.1	77.3±0.1	67.6±0.2	47.8±0.6	44.0±0.1	64.2
SagNet† [29]	**86.3**±0.2	77.8±0.5	68.1±0.1	48.6±1.0	40.3±0.1	64.2
SelfReg [58]	85.6±0.4	77.8±0.9	67.9±0.7	47.0±0.3	42.8±0.0	64.2
CORAL† [13]	86.2±0.3	78.8±0.6	68.7±0.3	47.6±1.0	41.5±0.1	64.5
mDSDI [59]	86.2±0.2	**79.0**±0.3	69.2±0.4	48.1±1.4	42.8±0.1	65.1
MIRO	85.4±0.4	**79.0**±0.0	**70.5**±0.4	**50.4**±1.1	**44.3**±0.2	**65.9**
Combined with SWAD [34]						
ERM + SWAD‡	88.1±0.1	79.1±0.1	70.6±0.2	50.0±0.3	46.5±0.1	66.9
CORAL + SWAD‡	88.3±0.1	78.9±0.1	71.3±0.1	51.0±0.1	46.8±0.0	67.3
MIRO + SWAD	**88.4**±0.1	**79.6**±0.2	**72.4**±0.1	**52.9**±0.2	**47.0**±0.0	**68.1**
Using RegNetY-16GF backbone with SWAG pre-training [41]						
ERM	89.6±0.4	78.6±0.3	71.9±0.6	51.4±1.8	48.5±0.6	68.0
MIRO	**97.4**±0.2	**79.9**±0.6	**80.4**±0.2	**58.9**±1.3	**53.8**±0.1	**74.1**
ERM + SWAD	94.7±0.2	79.7±0.2	80.0±0.1	57.9±0.7	53.6±0.6	73.2
MIRO + SWAD	**96.8**±0.2	**81.7**±0.1	**83.3**±0.1	**64.3**±0.3	**60.7**±0.0	**77.3**

method perspective, we examine two image self-supervised pre-training methods (Barlow Twins [60] and MoCo v3 [61]), one image-language self-supervised pre-training method (CLIP [37]), and one weakly-supervised pre-training method (SWAG [41]), as well as ImageNet supervised pre-training baseline (ImageNet ERM). From the pre-training scale perspective, we employ the ImageNet [36] dataset of 1.3 million examples, the CLIP dataset of 400 million examples, and the Instagram dataset of 3.6 billion examples. We use ResNet-50 [35] backbone architecture as default, but a bigger model is also used for the large-scale pre-training, such as ViT-B [47] for CLIP or RegNetY-16GF [46] for SWAG.

As shown in the table, MIRO improves performances compared with the baseline ERM in all experiments. For the ImageNet pre-training, applying MIRO results in performance improvements of +1.7pp, +3.5pp, and +1.3pp for ERM (supervised learning), Barlow Twins, and MoCo v3, respectively. For the large-scale pre-training, such as CLIP and SWAG, MIRO brings larger performance improvements of +16.3pp, +12.6pp, and +6.1pp for CLIP, CLIP-ViT, and SWAG, respectively. These experiments demonstrate the robustness of the proposed method to the pre-training methods, datasets, and backbone architectures.

Notably, performance improvements of MIRO are remarkable with large-scale pre-trained models, such as CLIP, CLIP-ViT, and SWAG. This is consistent with our observation in Section 2.2. Our method helps large-scale pre-trained models (in terms of the pre-training dataset size) not to be

Table 2: **Comparison with various pre-training datasets, methods, and backbones.** We compare the performance changes according to the scale of the dataset, the method, and the backbone architecture of pre-training. ResNet-50 architecture is used as default. `OH`, `TI`, and `DN` indicate `OfficeHome`, `TerraIncognita`, and `DomainNet`, respectively. Every accuracy is averaged over three trials.

Dataset (size)	**Pre-training**	**Alg.**	PACS	VLCS	OH	TI	DN	Avg.
ImageNet (1.3M)	ERM	ERM	84.2	77.3	67.6	47.8	44.0	64.2
		MIRO	85.4	79.0	70.5	50.4	44.3	65.9 (+1.7)
	Barlow Twins	ERM	78.7	77.3	57.6	36.9	41.7	58.4
		MIRO	80.7	79.4	63.7	43.2	42.6	61.9 (+3.5)
	MoCo v3	ERM	86.7	77.3	61.8	49.1	43.8	63.7
		MIRO	86.3	78.5	66.8	48.4	44.7	65.0 (+1.3)
CLIP (400M)	CLIP (ResNet)	ERM	64.3	69.8	28.2	32.9	29.5	44.9
		MIRO	76.6	78.9	59.5	49.0	42.0	61.2 (+16.3)
	CLIP (ViT)	ERM	83.4	75.9	66.4	35.3	44.4	61.1
		MIRO	95.6	82.2	82.5	54.3	54.0	73.7 (+12.6)
Instagram (3.6B)	SWAG (RegNet)	ERM	89.6	78.6	71.9	51.4	48.5	68.0
		MIRO	97.4	79.9	80.4	58.9	53.8	74.1 (+6.1)

Table 3: **Comparison with learning from pre-trained methods.** Out-of-domain accuracies on five domain generalization benchmarks are shown. Average accuracies and standard errors are reported from three trials.

Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
LP-FT [40]	84.6±0.8	76.7±1.5	65.0±0.2	47.1±0.7	43.0±0.1	63.3
L^2-SP [62]	83.6±0.3	78.8±0.4	65.0±0.3	47.9±2.1	42.5±0.2	63.6
DELTA [63]	83.1±1.1	77.7±0.4	68.5±0.3	45.7±0.9	42.8±0.1	63.6
LwF [64]	83.1±0.8	77.2±0.7	70.0±0.2	49.2±1.2	42.7±0.1	64.5
MIRO	**85.4**±0.4	**79.0**±0.0	**70.5**±0.4	**50.4**±1.1	**44.3**±0.2	**65.9**

biased to the training source domains compared to naive fine-tuning. Especially, naive fine-tuning of CLIP-ViT (61.1%) shows worse out-of-domain performance than fine-tuning ImageNet pre-trained model (64.2%). In contrast, MIRO can leverage the pre-trained knowledge from CLIP-ViT, resulting in superior performance (73.7%) compared with the ImageNet pre-trained model (65.9%). In our later analysis, we show that the knowledge of large-scale pre-trained models is more beneficial to domain generalization than the knowledge of ImageNet pre-trained models.

Comparison with learning from pre-trained methods. Although our approach is the first study to exploit the pre-trained model in the training process for the out-of-domain generalization, there are several studies that utilize the pre-trained model for different purposes. In the transfer learning field, L^2-SP [62] and deep learning transfer using feature map with attention (DELTA) [63] are proposed to improve in-domain performance in the fine-tuning scenario. Learning without forgetting (LwF) [64] is designed to maintain old task performance when learning new tasks in the continual learning setting. LP-FT [40] shows that fine-tuning distorts the pre-trained features and it inhibits out-of-distribution generalization. They propose a simple baseline to alleviate the distortion, freezing the feature extractor in the early training phase. As shown in Table 3, MIRO outperforms the comparison methods with large margins. These results demonstrate the effectiveness of our method design for the out-of-domain generalization.

3.3 Analysis of MIRO

Loss function interpretation: Σ distribution analysis. We can interpret the variance term of MIRO, $\Sigma(z_f)$ in Equation (6), as control variables of the distance loss between pre-trained features z_{f^0} and current learning features z_f. During the training phase, if the variance values become smaller then the model will preserve mutual information with the pre-trained model. On the contrary, when the model needs to learn new information, the variance will increase. We illustrate the learned

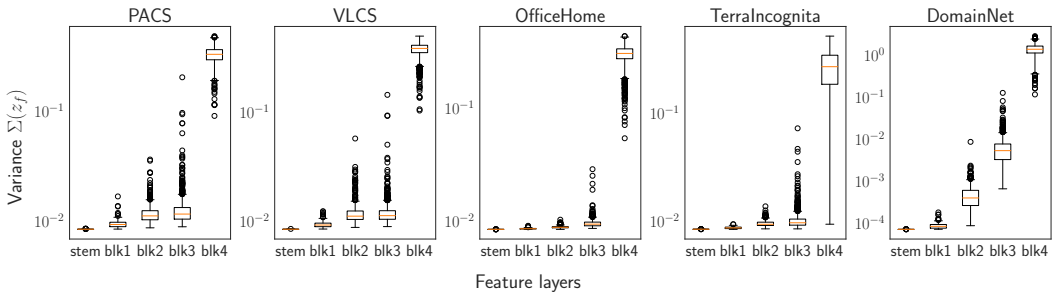


Figure 2: **Distribution of $\Sigma(z_f)$.** We plot the estimated variances, $\Sigma(z_f)$, for each layer. X-axis indicates the feature layer where the features z_f are collected. In all datasets, the variances increase as the layer is closer to the output.

(a) ImageNet (pre-train) (b) `Camelyon17` (fine-tuning)

Figure 3: **Example images of ImageNet and `Camelyon17`.** Large distribution shift occurs between pre-training (ImageNet) and fine-tuning (`Camelyon17`). ImageNet is a multiclass objective recognition task and `Camelyon17` is a binary classification task for reading whether the image contains tumor tissue. Instagram-3.6B examples are omitted since it is not publicly available.

variances in Figure 2. The figure shows that pre-trained information is preserved well in lower layers, while task-specific new information is learned in higher layers. This result is consistent with the interpretation that high layer features represent more task-specific semantic information than low layer features [65]; task shifts during DG fine-tuning make higher layer features learn more semantics than lower layers.

Case study on `Camelyon17`: large distribution shift between pre-training and fine-tuning. As shown in Equation (4), the tightness of the lower bound is directly connected to the divergence between the representations of oracle and pre-trained models. Therefore, we investigate the case that there is a large shift between pre-trained and target datasets using the medical dataset [66, 67], `Camelyon17`. Among the variants of the dataset, we use the patch-based version provided by Koh *et al.* [67]. This dataset consists of whole-slide images of histological lymph node sections from the five hospitals, where each hospital corresponds to each domain. The task is to predict whether the image contains a tumor tissue of breast cancer. As shown in Figure 3, there is a large gap between the pre-

Table 4: **Performance improvements in `Camelyon17` medical dataset.** Even in the large distribution shift setup between pre-training and target datasets, MIRO consistently outperforms ERM. Every accuracy is averaged over three trials.

Pretrain	**Algorithm**	1	2	3	4	5	Avg.
ImageNet ERM	ERM	97.1	94.7	95.7	96.4	90.7	94.9
	MIRO	97.5	94.5	95.6	96.7	93.7	95.6 (+0.7)
SWAG	ERM	97.0	94.1	95.3	96.0	89.5	94.4
	MIRO	97.4	95.5	96.5	96.1	90.9	95.3 (+0.9)

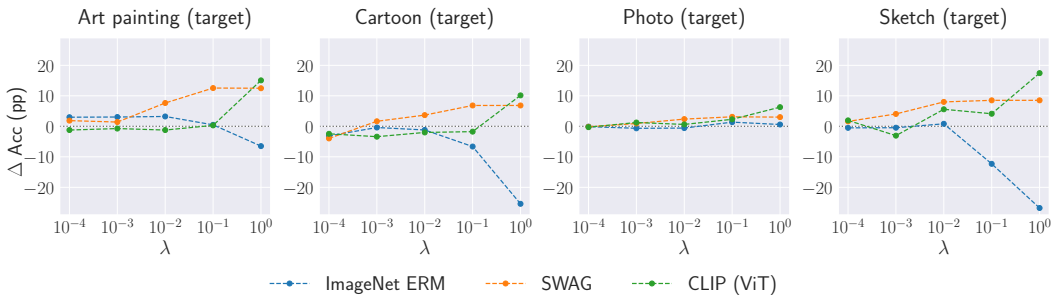


Figure 4: **Comparison of three pre-trained models according to λ.** Y-axis indicates performance difference of MIRO to ERM. λ is the intensity of the mutual information regularization. We compare three models: ResNet-50 pre-trained in ImageNet [35], RegNetY-16GF pre-trained by SWAG [41], and ViT-B pre-trained by CLIP [37].

training distribution (ImageNet or Instagram-3.6B) and the fine-tuning distribution (`Camelyon17`). The results in Table 4 demonstrate MIRO leads the model to learn robust representations even in the large distribution shift setup between pre-training and fine-tuning.

Relationship between the pre-training scale and the intensity of the mutual information regularization. Our method has a control parameter λ, which controls the balance between the cross-entropy loss and the mutual information regularization loss. If λ becomes larger, it implies that the strength of mutual information regularization becomes stronger, while it weakens the strength of ERM objective. Intuitively, if the pre-trained knowledge is informative enough to the target task, larger λ will improve the performances, while if the pre-trained knowledge is uninformative to the target task, then larger λ can harm the performances, because of the penalty on the ERM objective. We compare three pre-trained models (ImageNet pre-trained model, SWAG, and CLIP-ViT) by varying λ. Figure 4 shows how the out-of-domain performance of MIRO with different pre-trained backbones changes by λ. The additional results on different datasets are given in Appendix.

First, we observe that the ImageNet pre-trained backbone has a negative correlation between the performance difference and λ in target domains. When distribution shifts significantly differ, such as cartoon and sketch domains, we can observe an apparent negative correlation. We presume that it is because the ImageNet samples barely contain non-photo images, such as art painting or sketch images. On the other hand, we observe that MIRO with SWAG and CLIP-ViT backbones make significant performance improvements by choosing larger λ. In other words, SWAG and CLIP-ViT pre-trained knowledge are helpful to learn robust features for various target domains compared to the ImageNet pre-trained model. Furthermore, it implies that larger pre-trained models trained with massive diverse domain images show less sensitivity to the choice of λ, not only bringing remarkable performance improvements as shown in Table 2.

4 Conclusion

Traditional domain generalization (DG) approaches focus to learn a robust representation using multiple source domains. However, in the recent trends of scaling up pre-training, the use of a large-scale pre-trained model becomes more important than the use of DG algorithms for the real-world DG. In line with this trend, we propose Mutual Information Regularization with Oracle (MIRO) to robustly exploit the pre-trained model by approximating an oracle model. To do this, we first re-formulate the domain generalization objective by introducing a concept of an oracle model. Then, we derive a tractable variational bound of the objective by approximating the oracle model with the pre-trained model. Our experimental results demonstrate both the effectiveness and the potential of the proposed method. MIRO achieves state-of-the-art performance in the DomainBed benchmarks. Furthermore, when combining MIRO with large-scale pre-trained backbones, such as CLIP [37] or SWAG [41], the performance improvements remarkably increases. We hope that this study promotes a new research direction of exploiting pre-trained backbones to learn robust representations for domain generalization.

References

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

[2] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2020.

[3] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[4] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Conference on Fairness, Accountability, and Transparency*, 2020.

[5] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. In *International Conference on Learning Representations (ICLR)*, 2022.

[6] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[7] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning (ICML)*, 2020.

[8] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

[9] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[10] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[12] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[13] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450, 2016.

[14] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[15] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Neural Information Processing Systems*, 33, 2020.

[17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[18] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Neural Information Processing Systems*, 31, 2018.

[19] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Neural Information Processing System*, 2019.

[20] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[21] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020.

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

[23] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[24] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

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

[26] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018.

[27] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[28] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

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

[30] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[31] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, pages 561–578. Springer, 2020.

[32] Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. *AAAI Conference on Artificial Intelligence*, 2021.

[33] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

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

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[38] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[39] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

[40] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[41] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. *arXiv preprint arXiv:2201.08371*, 2022.

[42] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

[43] David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16(320):201, 2004.

[44] 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.

[45] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, 2017.

[46] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

[47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

[49] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

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

[51] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[52] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *European Conference on Computer Vision*, 2, 2020.

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

[54] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020.

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

[56] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[57] V Vapnik. Statistical learning theory. *NY: Wiley*, 1998.

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

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

[60] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[61] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[62] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pages 2825–2834. PMLR, 2018.

[63] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *International Conference on Learning Representations*, 2019.

[64] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[65] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[66] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018.

[67] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

A Derivation of Lower Bound

Assumption 1. *The variational distribution $q(\cdot|z)$ satisfies the regularity condition such that, for any $\mathbb{P}_{X|z} \in \{\mathbb{P}'_{X|z} \mid \mathbb{E}_{X|z}[|X|^2] < \infty\}$,*

$$\mathbb{E}_{X|z}\left[(\nabla_x \log q(x|z)|_{x=X})^\intercal \nabla_x \log q(x|z)|_{x=X}\right] < \infty, \tag{7}$$

where $\mathbb{E}_{X|z}$ is a conditional expectation of X given z.

Remark 1. *Note that the Gaussian distribution used in our implementation satisfies the regularity condition. To check the regularity condition of Gaussian distribution, we first compute the gradient as follows,*

$$\begin{aligned}
&\nabla_x \log q(x|z)|_{x=X} && (8)\\
&= \nabla_x \left(C + \frac{1}{2}\log|\Sigma(z)| + \frac{1}{2}(x-\mu(z))^\intercal \Sigma(z)^{-1}(x-\mu(z)) \right)|_{x=X} && (9)\\
&= \Sigma(z)^{-1}(X-\mu(z)). && (10)
\end{aligned}$$

Hence, we get,

$$\begin{aligned}
&\mathbb{E}_{X|z}\left[(\nabla_x \log q(x|z)|_{x=X})^\intercal \nabla_x \log q(x|z)|_{x=X}\right] && (11)\\
&\quad = \mathbb{E}_{X|z}\left[(X-\mu(z))^\intercal \Sigma(z)^{-2}(X-\mu(z))\right] < \infty. && (12)
\end{aligned}$$

since $\mu(z)$ and $\Sigma(z)$ are finite and $\mathbb{E}_{X|z}[|X|^2]$ is bounded. Hence, the Gaussian distribution satisfies the regularity condition.

Under the assumption of q, we derive the lower bound.

Derivation of the Lower Bound. Based on the regularity condition, we derive the lower bound of the term, $\mathbb{E}_{Z_{f^*},Z_f}[\log q(Z_{f^*} \mid Z_f)]$. Before starting the derivation, let us define $d_{2,\infty}(f,g) := \sup_x \|f(x)-g(x)\|_2$. Then, the derivation starts from Taylor's theorem for a differentiable multivariate function. From Taylor's theorem, there exists a point c such that $c = tx + (1-t)x_0$ for some $t \in [0,1]$ and the following equality holds,

$$\log q(x \mid y) = \log q(x_0 \mid y) + \nabla_x \log q(x \mid y)|^\intercal_{x=c}(x-x_0). \tag{13}$$

Then, we can derive the following upper bound as follows,

$$\begin{aligned}
\log q(x \mid y) &= \log q(x_0 \mid y) + \nabla_x \log q(x \mid y)|^\intercal_{x=c}(x-x_0) && (14)\\
&\le \log q(x_0 \mid y) + |\nabla_x \log q(x \mid y)|^\intercal_{x=c}(x-x_0)| && (15)\\
&\le \log q(x_0 \mid y) + \|\nabla_x \log q(x \mid y)|_{x=c}\|_2 \, \|x-x_0\|_2 && (16)
\end{aligned}$$

By using this bound, we can derive the following lower bound,

$$\begin{aligned}
\mathbb{E}_{Z_{f^*},Z_f}[\log q(Z_{f^*} \mid Z_f)] &= \mathbb{E}_{X,X'}\left[\log q(f^*(X) \mid f(X'))\right] && (17)\\
&\ge \mathbb{E}_{X,X'}\left[\log q(f^0(X) \mid f(X'))\right]\\
&\quad - \mathbb{E}_{X,X'}\left[\|\nabla \log q(c(X) \mid f(X'))\|_2 \left\|f^0(X) - f^*(X)\right\|_2\right] && (18)\\
&\ge \mathbb{E}_{X,X'}\left[\log q(f^0(X) \mid f(X'))\right]\\
&\quad - \mathbb{E}_{X,X'}\left[\|\nabla \log q(c(X) \mid f(X'))\|_2\right] d_{2,\infty}(f^*, f^0) && (19)\\
&\ge \mathbb{E}_{Z_{f^0},Z_f}\left[\log q(Z_{f^0} \mid Z_f)\right] - C d_{2,\infty}(f^*, f^0), && (20)
\end{aligned}$$

where $c(x)$ is the function between f^0 and f^*, which selects the point satisfying Taylor's theorem, and C is a constant derived from the regularity condition. □

B Additional Implementation Details

B.1 Hyperparameter tuning

We split the hyperparameters (HPs) into two groups: algorithm-specific HPs and algorithm-agnostic HPs. The algorithm-agnostic HPs consist of batch size, learning rate, dropout, and weight decay, and

MIRO has only one algorithm-specific HP, λ. To reduce the computational cost, we tune the algorithm-specific HPs and algorithm-agnostic HPs independently. We first search algorithm-specific HPs with default algorithm-agnostic HPs, then search algorithm-agnostic HPs with the tuned algorithm-specific HPs. That is, the λ is searched in [1.0, 0.1, 0.01, 0.001] with the batch size of 32, the learning rate of 5e-5, no dropout, and no weight decay. Then, we search algorithm-agnostic HPs with the searched λ following Cha *et al*. [34]. They propose reduced HP search space for efficiency compared to DomainBed [33]. The protocol searches the learning rate in [1e-5, 3e-5, 5e-5], dropout in [0.0, 0.1, 0.5], and weight decay in [1e-4, 1e-6]. The batch size per domain is fixed to 32. Since MIRO is a regularization method, we add a case of no weight decay.

Even though we use the efficient HP search protocol, it still requires heavy computational resources. Therefore, we tune λ only for the non-main experiments, including combination with SWAD, combination with various pre-trained backbone, and the case study on `Camelyon17`. Also, we use the batch size of 16 for SWAG [41] due to the GPU memory limitation. Note that there is room for further performance improvement by intensive HP tuning and additional usage of GPU memory, considering the simplified HP search protocol and limited computational resources.

B.2 Implementation details

The variance encoder is initialized to estimate the variance of 0.1. It is chosen by observing the convergence point of the variance. Softplus function is employed to ensure non-negativity of the variance. Also, we empirically apply the 10 times larger learning rate for the mean and variance encoders than the feature extractor and the classifier.

B.3 Mutual information estimation

In Section 2.2, we estimate the mutual information using Mutual Information Neural Estimator (MINE) [44]. The mutual information is estimated by MINE as follows:

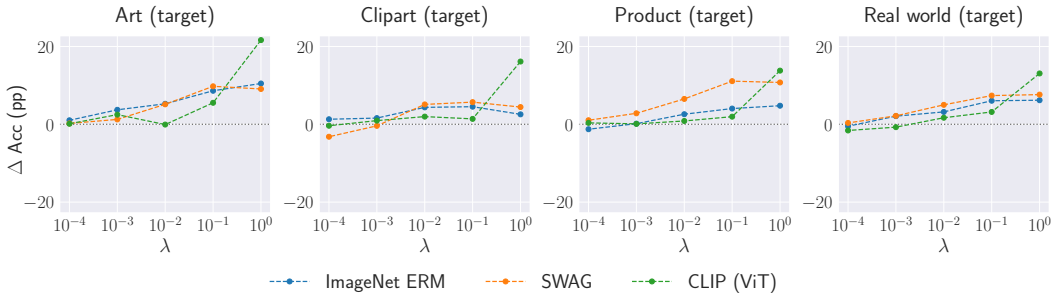
$$\widehat{I(Z_{f^*}; Z_f)} = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{Z_{f^* Z_f}}} [T_\theta] - \log \left(\mathbb{E}_{\mathbb{P}_{Z_{f^*}} \otimes \mathbb{P}_{Z_f}} \left[e^{T_\theta} \right] \right). \tag{21}$$

For the features Z_{f^*} and Z_f, the features after global average pooling are uniformly collected by domains. The statistics network, T_θ, consists of two hidden linear layers with 512 dimensions and ELU activation functions, following [44]. In the case of the fine-tuning, such as ERM$-$, ERM$+$, and MIRO, the models are trained as many as the number of target domains. Therefore, we estimate the mutual information for each model and report their average value.

C Additional Results

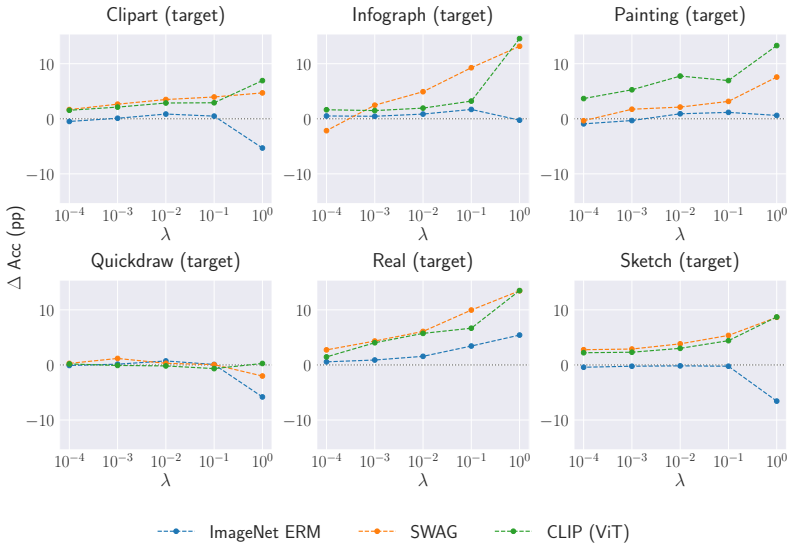
C.1 Relationship between the pre-training scale and the intensity of the mutual information regularization

In this section, we provide the extended results of Figure 4 in the main text. Figure 5 shows the additional comparison of three pre-trained backbones according to λ about `OfficeHome`, `TerraIncognita`, and `DomainNet`. The comparisons show similar trends with the results in `PACS`. ImageNet pre-trained backbone, such as ResNet-50 pre-trained in ImageNet [35], has a negative correlation between the performance difference and λ in some target domains. Large-scale pre-trained backbones, such as SWAG [41] and CLIP [37], tend to consistently make significant performance improvements at high λ and become less sensitive to the choice of λ.



(a) `OfficeHome`

(b) `TerraIncognita`



(c) `DomainNet`

Figure 5: **Comparison of three pre-trained models according to λ.** Y-axis indicates performance difference of MIRO to ERM. λ is the intensity of the mutual information regularization. We compare three models: ResNet-50 pre-trained in ImageNet [35], RegNetY-16GF pre-trained by SWAG [41], and ViT-B pre-trained by CLIP [37].