Domain-Specific Risk Minimization for Domain Generalization

Yi-Fan Zhang†
Institute of Automation
Beijing, China
yifanzhang.cs@gmail.com

Jindong Wang†
Microsoft Research Asia
Beijing, China
jindong.wang@microsoft.com

Jian Liang
Institute of Automation
Beijing, China
jian.liang@nlpr.ia.ac.cn

Zhang Zhang
Institute of Automation
Beijing, China
zzhang@nlpr.ia.ac.cn

Baosheng Yu
The University of Sydney
Australia
bayu0826@uni.sydney.edu.au

Liang Wang
Institute of Automation
Beijing, China
wangliang@nlpr.ia.ac.cn

Dacheng Tao
The University of Sydney
Australia
dacheng.tao@sydney.edu.au

Xing Xie
Microsoft Research Asia
Beijing, China
xingx@microsoft.com

ABSTRACT

Domain generalization (DG) approaches typically use the hypothesis learned on source domains for inference on the unseen target domain. However, such a hypothesis can be arbitrarily far from the optimal one for the target domain, induced by a gap termed “adaptivity gap”. Without exploiting the domain information from the unseen test samples, adaptivity gap estimation and minimization are intractable, which hinders us to robustify a model to any unknown distribution. In this paper, we first establish a generalization bound that explicitly considers the adaptivity gap. Our bound motivates two strategies to reduce the gap: the first one is ensembling multiple classifiers to enrich the hypothesis space, then we propose effective gap estimation methods for guiding the selection of a better hypothesis for the target. The other method is minimizing the gap directly by adapting model parameters using online target samples. We thus propose Domain-specific Risk Minimization (DRM). During training, DRM models the distributions of different source domains separately; for inference, DRM performs online model steering using the source hypothesis for each arriving target sample. Extensive experiments demonstrate the effectiveness of the proposed DRM for domain generalization. Code is available at: https://github.com/yfzhang114/AdaNPC.

CCS CONCEPTS

• Computing methodologies → Transfer learning; Learning latent representations; Neural networks.

† YF Zhang is also affiliated with School of Artificial Intelligence, University of Chinese Academy of Sciences.
‡ Corresponding author: Jindong Wang.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD ’23, August 6–10, 2023, Long Beach, CA, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8035-0/23/08...
https://doi.org/10.1145/3580305.3599313

KEYWORDS

Domain Generalization, Test-time Adaptation, Adaptivity gap

1 INTRODUCTION

Machine learning models generally suffer from degraded performance when the training and test data are non-IID (independently and identically distributed). To overcome the brittleness of classical empirical risk minimization (ERM), there is an emerging trend of developing out-of-distribution (OOD) generalization approaches [25, 35], where models trained on multiple source domains/datasets can be directly deployed on unseen target domains. Various OOD frameworks are proposed, e.g., disentanglement [39, 60], causal invariance [3, 29, 66], and adversarial training [13, 44, 65].

Existing approaches might rely on two strong assumptions. (i) Hypothesis over-confidence. Most works directly apply a source-trained hypothesis to any unseen target domains [3, 21, 41] by implicitly assuming that the training hypothesis space contains an ideal target hypothesis. However, the IID and OOD performances are not always positively correlated [51], i.e., the optimal hypothesis on source domains might not perform well on any target domains. The distance between the optimal source and target hypothesis is termed adaptivity gap [12], which is even shown can be arbitrarily large [8]. (ii) Pessimistic adaptivity gap reduction. Although the adaptivity gap is ubiquitous, it is almost impossible to identify and minimize due to the unavailability of OOD target samples. As a consequence, there exists no approach that can tackle all kinds of distribution shifts at once (e.g., diversity shift in PACS [24] and correlation shift in the Colored MNIST [3]), but only a specific kind [59]. In a word, it is almost impossible to robustify a model to arbitrarily unknown distribution shift without utilizing the target samples during inference.
To our best knowledge, the two disadvantages are always neglected by the commonly-used domain adaptation and generalization bounds [2, 4, 69], which mostly ignore the terms that are related to the target domain. To this end, we introduce a new generalization bound that independent on the choice of hypothesis space and explicitly considers the adaptivity gap between source and target. The bound motivates two possible test-time adaptation strategies: the first one is to train specific classifiers for different source domains, and then dynamically ensemble them, which is shown able to enrich the set of the hypothesis space [11]. The other is to utilize the arriving target samples, namely once a target sample is given, we update the model by its provided target domain information. To summarize, this paper makes the following contributions:

1. A novel perspective. We provide a new generalization bound that does not depend on the choice of hypothesis space and explicitly considers the adaptivity gap between source domains and the target domain. Our bound is shown tighter than the existing one and provides intuition for reweighting methods, test-time adaptation methods, and classifier ensembling methods for good domain generalization performance.

2. A new approach. We propose DRM method, which consists of two components: (i) During training, DRM constructs specific classifiers for source domains and is trained by reweighting empirical loss. (ii) During the test, DRM performs test-time model selection and retraining for each target sample. Thus, the source classifiers are dynamically changed for each target data and we can enrich the support set of the hypothesis space in this way to minimize the adaptivity gap directly.

3. Extensive experiments. We perform extensive experiments on popular OOD benchmarks showing that DRM (1) achieves very competitive generalization performance on both diversity shift benchmarks and correlation shift benchmarks; (2) beats most existing test-time adaptation methods with a large margin; (3) is orthogonal to other DG methods; (4) reserves strong recognition capability on source domains, and (5) is parameter-efficient and converges even faster than ERM thanks to the structure.

2 RELATED WORK

Domain adaptation and domain generalization Domain/Out-of-distribution generalization [26, 32, 33, 55, 64, 68] aims to learn a model that can extrapolate well in unseen environments. Representative methods like Invariant Risk Minimization (IRM) [3] concentrate on the objective of extracting data representations that lead to invariant prediction across environments under a multi-environment setting. In this paper, we emphasize the importance of considering the adaptivity gap and using online target data for adaptation. Without an invariance strategy, the proposed DRM can attain superior generalization capacity.

Test-time adaptive methods [27] are recently proposed to utilize target samples. Test-time Training methods need to design proxy tasks during tests such as self-consistency [63], rotation prediction [50] and need extra models; Test-time adaptation methods adjust model parameters based on unsupervised objectives such as entropy minimization [54] or update a prototype for each class [19]. Domain-adaptive method [12] needs extra models for adapting to the target domain. Non-Parametric Adaptation [67] needs to store all source domain instances. Our generalization bound indicates that these methods can explicitly reduce the target loss upper bound.

In this paper, we propose other ways to perform test-time adaptation, i.e., multi-classifier dynamic combination and retraining.

**Ensemble learning in domain generalization** learns ensembles of multiple specific models for different source domains to improve the generalization ability, e.g., domain-specific backbones [10], domain-specific classifiers [56], and domain-specific batch normalization [43]. Domain-specific classifiers are also used in this work; however, empirical results show that directly ensemble multiple classifiers with a uniform weight degrades the performance, and the proposed DRM achieves superior results in contrast.

**Labeling function shift and multi-classifiers**. Labeling function shift or correlation shift is not a novel concept and is commonly used in domain adaptation [48, 61, 69] or domain generalization [59]. There are also some studies on DG that are proposed to tackle this problem. CDANN [26] considers the scenario where both $P(X)$ and $P(Y|X)$ change across domains and proposes to learn a conditional invariant neural network to minimize the discrepancy in $P(X|Y)$ between different domains. [31] explores both the correlation and label shifts in DG and aligns the correlation shift via variational Bayesian inference. The proposed DRM is different from these studies because we want the labeling functions $P(Y|X)$ to be more specific to each domain than invariant.

3 A BOUND BY CONSIDERING ADAPTIVITY GAP

**Problem Formulation.** Let $X$, $Y$, $Z$ denote the input, output, and feature space, respectively. We use $X, Y, Z$ to denote the random variables taking values from $X, Y, Z$, respectively. We focus on the domain generalization setting, where a labeled training dataset consisting of several different but related training distributions (domains) is given. Formally, $D = \bigcup_{i=1}^{K} D_i$, where $K$ is the number of domains. Each $D_i$ corresponds to a joint distribution $P_i(X, Y)$ with an optimal classifier $f_i : X \rightarrow \{0, 1\}$.

![Figure 1: A failure case of invariant representations for domain generalization. (a) Four domains in different colors: orange ($\mu_0 = [-3, 0, 3]$, green ($\mu_g = [3, 0, 3]$), red ($\mu_r = [-3, -3]$) and blue ($\mu_b = [3, -3]$). (b) Invariant representations learnt from domain $D_1$ and $D_2$ by feature transformation $g(X) = \mathbf{x}_{1 < 0} \cdot (x_1 + 3) + \mathbf{x}_{1 > 0} \cdot (x_1 - 3)$. The grey color indicates the transformed target domains. (c) The classification boundary learned by DRM.](image-url)
where $l$ indicates the identity matrix. Then, the optimal hypothesis $f^*(X) = 1$ if $x_i \in (−3, 3)$ achieves perfect classification on all domains\(^3\). Let $\mathcal{D}_r, \mathcal{D}_o$ denote source domains and $\mathcal{D}_a, \mathcal{D}_b$ denote target domains. Given hypothesis $\hat{f} : h \circ g$ where the feature transformation function is $g(x) = \mathbb{I}_{x < 0} \cdot (x + 3) + \mathbb{I}_{x > 0} \cdot (x - 3)$ in Figure 1 (b), namely, the invariant representation of $\mathcal{D}_r, \mathcal{D}_o$ is learnt, which is $\mathcal{D}_r b = g \circ \mathcal{D}_o = g \circ \mathcal{D}_b = \mathcal{N}(\{0, −3\}, f)$. However, the labeling functions $f_0$ of $\mathcal{D}_r$ and $f_b$ of $\mathcal{D}_b$ are just the reverse such that $f_b(x) = 1 - f_0(x), \forall X \in \mathcal{D}_b$. In this case, according to Eq. 1, we have that $e_{\mathbb{R}}(\hat{f})$ is equal to:

$$
e_{\mathbb{R}}(h \circ g) + e_{\mathbb{R}}(h \circ g) = P_{X \sim \mathcal{D}_r}(h(X) \neq f_0(X)) + P_{X \sim \mathcal{D}_r}(h(X) \neq f_b(X))$$

$$= 1 - P_{X \sim \mathcal{D}_r}(h(X) \neq f_0(X)) + P_{X \sim \mathcal{D}_r}(h(X) \neq f_b(X)) = 1$$

Therefore, the invariant representation leads to large joint errors on all source and target domains for any prediction head $h$ without considering the adaptivity gap. Motivated by this, we provide a tighter OOD upper bound that considers the adaptivity gap.

**Proposition 2.** Let $\{\mathcal{D}_i, f_i\}_{i=1}^K$ and $\mathcal{D}_T, f_T$ be the empirical distributions and corresponding labeling function for source and target domain, respectively. For any prediction $\hat{f} \in \mathcal{H}$, given mixed weights $\{a_i\}_{i=1}^K \cdot \mathbb{I}_{\|a_i\|_1} \geq 1; a_i \geq 0$, we have:

$$e_T(\hat{f}) \leq \sum_{i=1}^K \left( P_{X \sim \mathcal{D}_i} \left[ a_i P_{T}(X) f_i(X) \right] + \mathbb{I}_{\|f_i - f_T\|} \right)$$

The two terms on the right-hand side have natural interpretations: the first term is the weighted source errors, and the second one measures the distance between the labeling functions from the source domain and target domain. Compared to Eq. 2, Eq. 5 does not depend on $\lambda_i$, i.e., the choice of the hypothesis class $\mathcal{H}$ makes no difference. More importantly, the new upper bound in Eq. 5 reflects the influence of adaptivity gaps between each source domain to the target, i.e., $P_{T}(X) f_i(X)$. The most similar generalization bound to us is $[1]$, in Appendix A.3, we show that the proposed bound is tighter. Although in this work, the density ratio $P_T(x)/P_i(x)$ is ignored and regarded as a constant, it has an interesting connection between reweighting methods.

**Connect the density ratio to reweighting methods.** Intuitively, the density ratio stresses the importance of data sample reweighting, where data samples that are more likely from the target domain should have larger weights. Note that estimating $P_T(x)/P_i(x)$ directly is intractable and the term is significantly

$$\text{A failure case of marginal invariant representation.}$$

We construct a simple counterexample where invariant representations fail to generalize. As shown in Figure 1, given the following four domains: $\mathcal{D}_a \sim \mathcal{N}(−3, 3), \mathcal{D}_b \sim \mathcal{N}(3, 3), \mathcal{D}_r \sim \mathcal{N}(−3, 3), \mathcal{D}_b \sim \mathcal{N}(3, 3), \mathcal{D}_s$, where $X = (x_1, x_2)$ and

$$f_0(X) = \begin{cases} 0 & \text{if } x_1 \leq -3 \\ 1 & \text{otherwise} \end{cases} \quad f_b(X) = \begin{cases} 1 & \text{if } x_1 \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

$$f_b(X) = \begin{cases} 1 & \text{if } x_1 \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

The adaptivity gap is NOT the same as labeling functions difference [69], where the latter measures the difference of two hypotheses: $\min \{E_{\mathcal{D}_r} \left[ f_i - f_T \right], E_{\mathcal{D}_o} \left[ f_i - f_T \right] \}$. However, the error of target hypothesis $f_T$ on the source domain is intractable to estimate and meaningless for DG [20]. The definition of adaptivity gap directly measures if the source classifier performs well on the target.

$\text{Definition 2 (Adaptivity gap). \ The adaptivity gap between } \mathcal{D}_i \text{ and the target domain } \mathcal{D}_T \text{ can be formally defined as } E_{\mathcal{D}_T} \left[ |f_i - f_T| \right], \text{ namely the error incurred by using } f_i \text{ for inference in } \mathcal{D}_T.$

$\text{The adaptivity gap is NOT the same as labeling functions difference [69], where the latter measures the difference of two hypotheses: min } \{E_{\mathcal{D}_r} \left[ f_i - f_T \right], E_{\mathcal{D}_o} \left[ f_i - f_T \right] \}. \text{ However, the error of target hypothesis } f_T \text{ on the source domain is intractable to estimate and meaningless for DG [20]. The definition of adaptivity gap directly measures if the source classifier performs well on the target.}$

$\text{Although Gaussian distributions put some mass on parts of the input space where this } f^* \text{ misclassifies some examples (} x_i > 3 \text{ for } \mathcal{D}_i), \text{ the density of these scopes are very small and can be ignored.}$
problematic with no constraint. However, we can make some safe assumptions and obtain applicable formulations, which is exactly what distributionally robust optimization (DRO) \cite{6} does. Specifically, if we restrict the target domain within a $f$-divergence ball (such as Kullback-Leibler divergence) from the training distribution, which is also known as KL-DRO \cite{17}, then the density ratio will be converted to a reweighting term $e^{f(x)/	au^*}$ used for training, where $f$ indicates the classification error incurred by $(x, y)$ and $\tau^*$ is a hyperparameter. Namely, the reweighting term is actually an approximate estimation of the density ratio. Existing methods \cite{30, 42, 64} use similar reweighting terms and our error bound provides a theoretical explanation for why they work well on DG. Existing methods \cite{30, 42, 64} use similar reweighting strategies and our error bound provides a theoretical explanation for why they work well on DG (See Appendix A.4 for formal derivation).

4 DOMAIN-SPECIFIC RISK MINIMIZATION

Our error bound in Eq. 5 suggests a novel perspective on OOD algorithm design. In this paper, we follow the test-time adaptation setting for domain generalization \cite{19} and try to utilize the online target samples to minimize the adaptivity gap. However, Eq. 5 needs to calculate the expectation and the optimal hypothesis function $f_T$ on the target domain, which are very challenging to obtain. Therefore, we propose a heuristic algorithm, DRM, which avoids the calculation of intractable terms in Eq. 5 and approximately minimizes the bound. The main pipeline of the proposed Domain-Specific Risk Minimization (DRM) is shown in Figure 2.

4.1 Domain-Specific Labeling Function

One natural idea is to use domain-specific classifiers $\{\hat{f}_i\}_{i=1}^K$ rather than a shared classifier $\hat{f}$ for source domains. Each $\hat{f}_i$ is responsible for classification in $D_i$. During training, our goal is to minimize $\frac{1}{K} \sum_{i=1}^K \mathbb{E}_{x \sim D_i} |\hat{f}_i(x) - f(x)|$ by assuming that $K$ training domains are uniformly mixed ($\alpha_i = 1/K$). The generalization results are better with reweighting terms, e.g., using GroupDRO \cite{42}, in the RotatedMNIST dataset, the accuracy of $d = 5$ with reweighting terms is 97.3%, which is better than 96.8% without reweighting. We simply ignore the reweighting term in this work since it is not our focus.

Specifically, given $K$ source domains, DRM utilizes a shared encoder $g$ and a group of prediction head $\{h_i\}_{i=1}^K$ for all domains, respectively. The encoder is trained by all data samples while each head $h_i$ is trained by images from domain $D_i$. It is also possible (but less efficient and accurate) to use specific $g_i$ for each domain.\footnote{Using domain-specific $g_i$ will inevitably increase the computation and memory burden. We observe that $h_i \circ g$ gives an OOD accuracy of 70.1% while the result is only 64.8% for $h_i \circ g_i$ on the Colored MNIST dataset. A possible reason is that a shared encoder $g$ can be seen as an implicit regularization, which prevents the model from overfitting specific domains.}

4.2 Test-Time Model Selection and Adaptation

Test-time adaptive intuitions from the bound. After training, we can get $K$ hypotheses $\hat{f}_i$ that can well approximate source labeling functions. During testing, our error bound provides two strategies to minimize the second term in the upper bound, i.e., $\sum_{i=1}^K a_i \mathbb{E}_{D_T} |\hat{f}_i - f_T|$, one natural strategy is to find $a^* = \arg \min a_i \mathbb{E}_{D_T} |\hat{f}_i - f_T|$, which is termed test-time model selection. The intuition is that if we can find the source domain $D_T$ with a labeling function $f_T$ that minimizes the adaptivity gap $\mathbb{E}_{D_T} |\hat{f}_i - f_T|$, then we have that $a_i = 1, \text{if } i = t^*$, otherwise 0 will minimize this term. Second, if we suppose $\hat{f}_i = f_T$, then minimizing $\sum_{i=1}^K \mathbb{E}_{D_T} |\hat{f}_i - f_T|$ will also minimize the bound. The resulting strategy is termed test-time retraining. Since $f_T$ is unknown, we can update model parameters by the inferred target pseudo labels or use some unsupervised losses such as entropy minimization. Note that these two strategies are orthogonal and can be used simultaneously. In the following, we articulate these two strategies.

4.2.1 Test-time model selection. As mentioned above, we can manipulate $a_i$ to affect the second term in our bound: for every test sample $x \in D_T$, if we can estimate the adaptivity gap $\{H_i = |\hat{f}_i(x) - f_T(x)|\}_{i=1}^K$ and choose $t^* = \arg \min_i |H_i|$, then $a_i = 1, \text{if } i = t^*$, otherwise 0 makes this term the minimum and the prediction will be $\hat{f}_i(x)$. The challenge is estimating $\{H_i\}_{i=1}^K$ and we propose two approximations.

Similarity Measurement (SM). We first reformulate $a_i \mathbb{E}_{D_T} |\hat{f}_i - f_T|$ as follows:

$$a_i \mathbb{E}_{D_T} |\hat{f}_i - f_T| = a_i \mathbb{E}_{D_T} \left[ |\hat{f}_i - \mathbb{E}_{D_T} \hat{f}_i| + \mathbb{E}_{D_T} \hat{f}_i - f_T \right] \leq a_i \left( \mathbb{E}_{D_T} |\hat{f}_i - \mathbb{E}_{D_T} \hat{f}_i| + \mathbb{E}_{D_T} |\mathbb{E}_{D_T} \hat{f}_i - f_T| \right),$$

where $f_T$ is intractable and we then focus on $\mathbb{E}_{D_T} |\hat{f}_i - \mathbb{E}_{D_T} \hat{f}_i|$, which intuitively measures the prediction difference of the given test data $x \in D_T$ and the average prediction result in domain $D_i$. However, taking the average of the prediction labels might produce ill-posed results\footnote{If all source domains have two data samples with different labels, e.g., two different one-hot labels $[0, 1], [1, 0]$. Then the average prediction result of all source domains will be $[0.5, 0.5]$ and have no difference.} and we use $\mathbb{E}_{D_T} |g - \mathbb{E}_{D_T} g| |\$ to approximate this term, where we calculate the representation difference between the test sample and the average representation of the domain $D_i$.}
For each $x \in \mathcal{D}_r$, the estimation $H_i = \text{Dist}(g(x), \mathbb{E}_{\mathcal{D}_i}[g])$, i.e., the distance between $g(x)$ and the average representation of $\mathcal{D}_i$. The Dist function can be any distance metric such as $L_p$-Norm, the negative of cosine similarity, $f$-divergence [38], MMD [25], or Â-distance [4]. We use cosine similarity (CSM) and $L_2$-Norm (L2SM) in our experiments for simplicity.

**Prediction Entropy Measurement (PEM).** During testing, denote the $K$ individual classification logits as $\{y^k_i\}_{k=1}^K$, where $y^k_i = \{y^k_{i1}, ..., y^k_{iC}\}$, and $C$ is the number of classes. Given the following assumption: "the more confident prediction $h_i \circ g$ makes on $\mathcal{D}_r$, the more similar $f_i$ and $f_T$ will be". Then, the prediction entropy of $\tilde{y}^k_i$ can be calculated as $H_k = -\sum_{i=1}^C \frac{y^k_i}{\sum_{j=1}^C y^k_j} \log \frac{y^k_i}{\sum_{j=1}^C y^k_j}$, where the entropy is used as our expected estimation. In our experiments, we find that the prediction entropy is consistent with domain similarities, which is similar to SM.

**Model Ensembling.** A one-hot mixed weight is too deterministic and cannot fully utilize all learned classifiers. Softing mixed weights, on the other hand, can further boost generalization performance and enlarge the hypothesis space, i.e., for ERM, we can generate the final prediction as $\sum_{k=1}^K y^k_i H_k^{-\gamma} / \sum_{k=1}^K H_k^{-\gamma}$, where $H_k$ indicates the contribution of each classifier. We use $-\gamma$, but not $\gamma$ since the smaller the adaptivity gap, the larger the contribution of $f_i$ should be. Specifically, for $\gamma = 0$, we then have a uniform combination, i.e., $a_i = 1/K, \forall i \in \{1, 2, ..., K\}$; for $\gamma \to \infty$, we then have a one-hot weight vector with $a_i = 1$ iff $i = i^*$ otherwise 0. In experiments, we compare the different selection strategies and PEM generally performs the best, thus we use PEM by default.

### 5 EXPERIMENTAL RESULTS

We first conduct case studies on a popular correlation shift dataset (Colored MNIST). Then, we compare DRM with other advanced methods on DG benchmarks (diversity shift). The results verify the argument in the introduction: by utilizing the target data during inference, we can better robustify a model to both distribution shifts. We also compare DRM with different test-time adaptive methods with various backbones. For fair comparisons, We use test-time retraining just when compared to test-time adaptation methods, namely DRM denotes the method wo/ retraining.

**Experimental Setup.** We use five popular OOD generalization benchmark datasets: Colored MNIST [3], Rotated MNIST [14], PACS [24], VLCS [52], and DomainNet [40]. We compare our model with ERM [53], IRM [3], Mixup [57], MLDG [23], CORAL [49], DANN [13], CDANN [26], MTL [7], SagNet [36], ARM [62], VREX [21], RSC [18], Fish [45], and Fishr [41]. All the baselines in DG tasks are implemented using the codebase of Domainbed [15].

**Hyperparameter search.** Following the experimental settings in [15], we conduct a random search of 20 trials over the hyperparameter distribution for each algorithm and test domain. Specifically, we split the data from each domain into 80% and 20% proportions, where the larger split is used for training and evaluation, and the smaller ones are used for select hyperparameters. We repeat the entire experiment twice using different seeds to reduce randomness. Finally, we report the mean over these repetitions as well as their estimated standard error. We observe that the proposed DRM does not converge within 5k iterations on the DomainNet dataset and we thus train it with an extra 5k iterations.

**Implementation details.** During training, we use the average of all classifiers’ losses as the training loss. To further enlarge the hypothesis space, we can simply add an additional prediction head that is trained by all data samples, namely, we have a total of $K + 1$ prediction heads in the test phase, such a simple trick is optional and can bring performance gains on some of our benchmarks.

**Model selection** in domain generalization is intrinsically a learning problem, and we use test-domain validation, one of the three methods in [15]. This strategy is an oracle-selection one since we
choose the model maximizing the accuracy on a validation set that follows the distribution of the test domain.

**Model architectures.** Following [15], we use as encoders ConvNet for RotatedMNIST (detailed in Appendix D.1 in [15]) and ResNet-50 for the remaining datasets.

See Appendix B for dataset details.

5.1 Case Studies on Correlation Shift Datasets

In the following, we conduct thorough experiments and analyze a popular correlation shift benchmark, i.e., the ColoredMNIST dataset [3]. It constructs a binary classification problem based on the MNIST dataset (digits 0-4 are class one and 5-9 are class two). Digits in the dataset are either colored red or green, and there is a strong correlation between color and label but the correlations vary across domains. For example, green digits have a 90% chance of belonging to class 1 in the first domain +90%(d = 0), and a 10% chance of belonging to class 1 in the third domain -90%(d = 2).

**DRM has superior generalization ability on the dataset with correlation shift.** As shown in Table 2, ERM achieves high accuracy in training domains, but lower chance accuracy in the test domain due to its reliance on spurious correlations. IRM [3] forms a trade-off between training and testing accuracy. An ERM model trained on only gray images, i.e., ERM (gray), is perfectly invariant by construction and attains a better tradeoff than IRM. The upper bound performance of invariant representations (OIM) is 90% varying across domains. For example, green digits have a 90% chance of belonging to class 1 in the first domain $+90\%(d = 0)$, and a 10% chance of belonging to class 1 in the third domain $-90\%(d = 2)$.

In contrast, both ERM and DRM retain high generalization performance. Specifically, DRM improves performance compared to ERM on domains $d \in \{0, 5\}$.

**Comparison with test-time adaptative methods.** For fair comparisons, following [19], the base models (ERM and DRM) are trained only on the default hyperparameters and with the fine-grained parametric search. Because [15] omits the BN layer from ResNet when fine-tuning on source domains, we cannot simply use BN-based methods on the ERM baseline. For these methods, their baselines are additionally trained on ResNet-50 with BN. Models with the highest IID accuracy are selected and all test-time adaptation methods are applied to improve generalization performance.

Results show that DRM achieves satisfying in- and out-distribution performance. The baselines include Tent [54], T3A [19], pseudo labeling (PL) [22], SHOT [28], and SHOT-IM [28]. For methods that use gradient backpropagation, we implement both the update of the prediction head (CH) and the full model (Full). Results in Table 5 show that: (i) Simply retraining the classifier or the full model by its own prediction improves the performance on the most challenging domains compared to ERM and beats IRM by a large margin, which indicates that DRM achieves satisfying in- and out-distribution performance.

5.2 Results on General OOD Benchmarks

**OOD results.** The average OOD results on all benchmarks are shown in Table 3. We observe consistent improvements achieved by DRM compared to existing algorithms. The results indicate the superiority of DRM in real-world diversity shift datasets. See the Appendix for multi-target domain generalization and detailed performance on every domain.

**In-distribution results.** Current DG methods ignore the performance of source domains since they focus on target results. However, source domain performance is also of great importance in applications [58], i.e., the in-distribution performance. We then show the in-distribution performances of VLCS and PACS in Table 4. DRM achieves comparable or superior performance in the source domains compared to ERM and beats IRM by a large margin, which indicates that DRM achieves satisfying in- and out-distribution performance.

Results of various backbones. We conduct experiments with various backbones in Table 5, including ResNet-50, ResNet-18, and Vision Transformers (ViT-B16). DRM achieves consistent performance improvements compared to ERM. Specifically, DRM improves 5.3%, 4.7%, and 3.9% for ResNet-50, ResNet-18, and ViT-B16 with evaluation batch size (BSZ) 32, respectively.

Multi-target domain generalization. IRM [3] introduces specific conditions for an upper bound on the number of training environments required such that an invariant optimal model can be obtained, which stresses the importance of several training environments. In this paper, we reduce the training environments on the Rotated MNIST from five to three. As shown in Table 8, as the number of training environments decreases, the performance of IRM decreases significantly (e.g., the average accuracy from 97.5% to 91.8%), and the performance on the most challenging domains $d = \{0, 5\}$ declines the most (94.9% → 80.9% and 95.2% → 91.1%). In contrast, both ERM and DRM retain high generalization performances while DRM outperforms ERM on domains $d = \{0, 5\}$.
Table 2: Accuracies (%) of different methods for the Colored MNIST synthetic task. OIM (optimal invariant model) and RG (random guess) are hypothetical mechanisms.

| Method | +90% (d = 0) | +80% (d = 1) | -90% (d = 2) | Avg |
|--------|--------------|--------------|--------------|-----|
|        | train        | test         | train        | test | train | test | train | test | Avg |
| ERM    | 86.1±3.9     | 71.8±0.4     | 83.6±0.5     | 72.9±0.1 | 87.5±3.4 | 28.7±0.5 | 85.7 | 57.8 |
| IRM    | 78.2±9.5     | 72.0±0.1     | 70.6±9.1     | 72.5±0.3 | 85.3±4.7 | 58.5±3.3 | 78.0 | 67.7 |
| DRM    | 81.8±9.8     | 86.7±2.4     | 90.2±0.2     | 80.6±0.2 | 88.0±4.5 | 43.1±7.5 | 86.7 | 70.1 |
| +CORAL | 83.4±8.6     | 85.3±2.3     | 91.6±0.7     | 80.7±0.2 | 89.4±4.9 | 47.2±3.6 | 88.1 | 71.1 |
| RG     | 50           | 50           | 50           | 50    | 50    | 50    | 50    | 50 |
| OIM    | 75           | 75           | 75           | 75    | 75    | 75    | 75    | 75 |
| ERM (gray) | 84.8±2.7   | 73.9±0.3     | 84.3±1.4     | 73.7±0.4 | 83.4±2.3 | 73.8±0.7 | 84.2 | 73.8 |

Table 3: Out-of-distribution generalization performance. No retraining is applied for a fair comparison.

| Method | CMNIST | RMNIST | VLCS | PACS | DomainNet | Avg |
|--------|--------|--------|------|------|-----------|-----|
| ERM [53] | 57.8±0.2     | 97.8±0.1 | 77.6±0.3 | 86.7±0.3 | 41.3±0.1 | 72.2 |
| IRM [3]       | 67.7±1.2     | 97.5±0.2 | 76.9±0.6 | 84.5±1.1 | 28.0±5.1 | 70.9 |
| GDRO [42]   | 61.1±0.9     | 97.9±0.1 | 77.4±0.5 | 87.1±0.1 | 33.4±0.3 | 71.4 |
| Mixup [57]   | 58.4±0.2     | 98.0±0.1 | 78.1±0.3 | 86.8±0.3 | 39.6±0.1 | 72.2 |
| CORAL [49]   | 58.6±0.5     | 98.0±0.0 | 77.7±0.2 | 87.1±0.5 | 41.8±0.1 | 72.6 |
| DANN [13]    | 57.0±1.0     | 97.9±0.1 | 79.7±0.5 | 85.2±0.2 | 38.3±0.1 | 71.6 |
| CDANN [26]   | 59.5±2.0     | 97.9±0.0 | 79.9±0.2 | 85.8±0.8 | 38.5±0.2 | 72.3 |
| MTL [7]      | 57.6±0.3     | 97.9±0.1 | 77.7±0.5 | 86.7±0.2 | 40.8±0.1 | 72.1 |
| SagNet [36]  | 58.2±0.3     | 97.9±0.0 | 77.6±0.1 | 86.4±0.4 | 40.8±0.2 | 72.2 |
| ARM [62]     | 63.2±0.7     | 98.1±0.1 | 77.8±0.3 | 85.8±0.2 | 36.0±0.2 | 72.2 |
| VREx [21]    | 67.0±1.3     | 97.9±0.1 | 78.1±0.2 | 87.2±0.6 | 30.1±3.7 | 72.1 |
| RSC [18]     | 58.5±0.5     | 97.6±0.1 | 77.8±0.6 | 86.2±0.5 | 38.9±0.6 | 71.8 |
| Fish [45]    | 61.8±0.8     | 97.9±0.1 | 77.8±0.6 | 85.8±0.6 | 43.4±0.3 | 73.3 |
| Fishr [41]   | 68.8±1.4     | 97.8±0.1 | 78.2±0.2 | 86.9±0.2 | 41.8±0.2 | 74.7 |
| DRM      | 70.1±2.0     | 98.1±0.2 | 80.5±0.3 | 88.5±1.2 | 42.4±0.1 | 75.9 |
| DRM+CORAL | 71.1±1.3     | 98.3±0.1 | 79.5±2.4 | 88.4±0.9 | 42.7±0.1 | 76.0 |

Figure 3: The entropy of different predictions. (a) Training domain {0,1} and testing domain {2}. (b) The average of training/testing domains {0,1}/(2), {0,2}/(1), and {1,2}/(0). (c) Domain-classifier correlation matrix, the value $c_{ij}$ is the entropy of predictions incurred by predicting samples in the domain $i$ with classifier $j$. Dom.$i$ indicates the classifier for the domain $d = i$. (d) Domain-classifier correlation matrices on Rotated MNIST.
Table 4: In-distribution performance on VLCS and PACS.

| Method  | VLCS       |   |   |   | Avg |
|---------|------------|---|---|---|-----|
| ERM     | 78.2±3.3   | 87.8±9.0 | 86.3±10.2 | 83.3±11.6 | 83.5 |
| IRM     | 76.9±2.9   | 88.2±8.9 | 85.3±9.8  | 77.3±1.0  | 81.9 |
| DRM     | 78.5±2.9   | 87.2±9.2 | 87.3±9.0  | 84.0±10.9 | 84.3 |

Table 5: (Left) Comparison of our method and existing test-time adaptation methods on PACS. (Right) Domain generalization accuracy with different backbone networks on PACS. The reported number is the average generalization performance over P, A, C, S four domains.

| Method   | BSZ=32 | BSZ=8 | Method   | BSZ=32 | BSZ=8 |
|----------|--------|--------|----------|--------|--------|
| ResNet50 | 83.98  | 83.98  | ResNet50 | 83.98  | 83.98  |
| PLClf    | 85.63  | 85.55  | DRM      | 86.57  | 86.57  |
| PLFull   | 86.50  | 85.88  | +Retrain Cls | 87.90  | 87.83  |
| SHOT     | 86.53  | 85.85  | +Retrain Full | 89.30  | 89.33  |
| SHOTIM   | 86.40  | 85.68  | ResNet18 | 79.98  | 79.98  |
| T3A      | 86.23  | 86.00  | DRM      | 80.30  | 80.30  |
| ResNet50-BN | 83.18  | 83.18  | +Retrain Cls | 82.95  | 82.18  |
| TentClf  | 84.15  | 84.15  | +Retrain Full | 84.70  | 84.35  |
| TentNorm | 85.60  | 84.00  | VIT-B16  | 87.10  | 87.10  |
| DRM      | 86.57  | 86.57  | DRM      | 87.85  | 87.85  |
| +Retrain Cls | 87.90  | 87.83  | +Retrain Cls | 90.08  | 90.08  |
| +Retrain Full | 89.30  | 89.33  | +Retrain Full | 90.95  | 90.85  |

5.3 Ablation Studies and Analysis

Different model selection strategies. Here we also conduct another baseline termed Neural Network Measurement (NNM). To fully utilize the modeling capability of the neural network, we propose estimating \( a_1E_{D_\tau}[|f_1 - f_\tau|] \) by NN. Specifically, during training, a domain discriminator is trained to classify which domain is each image from. During test, for \( x \in D_\tau \), the prediction result of the discriminator will be \( (d_i)_K \) and \( (H_1 = -d_i)_K \) is used as the estimation. We compare all the proposed strategies and a simple ensembling learning baseline, which uses a uniform weight for classifier ensembling. Table 6 (left) shows that the simple ensembling method works poorly in all domains. In contrast, the proposed methods achieve consistent improvements and PEM generally performs best.

Correlation matrix. From the correlation matrices, we find that (i) the entropy of the predictions between one source domain and its corresponding classifier is minimal. (ii) In the target domain, the classifiers cannot attain a very low entropy as on the corresponding source domains. (iii) The entropy of the predictions has a certain correlation with domain similarity. For example, in Figure 3(d), the classifier for domain \( d = 1 \) (with rotation angle 15°) achieves the minimum entropy in the unseen target domain \( d = 0 \) (no rotation). As the rotation angle increases, the entropy also increases. This phenomenon also occurs in other domains. Refer to the appendix for more analysis.

DRM has comparable model complexity to existing DG methods. As shown in Table 6 (right), methods that require manipulating gradients (Fish [45]) or following the meta-learning pipeline (ARM [62]) have a much slower training speed compared to ERM. The proposed DRM, without the need for aligning representations [13], matching gradient [45], or learning invariant representations [3], has a training speed that is faster than most existing DG methods, especially on small datasets RotatedMNIST. The training speed of DRM is slower than ERM due to the training of additional classifiers. As the number of domains/classes increases or the feature dimension increases, the training time of DRM will increase accordingly; however, DRM is always comparable to ERM and much faster than Fish and ARM (Table 7). For model parameters, since all classifiers in our implementation are just a linear layer, the total parameters of DRM is similar to ERM and much less than existing methods such as CDANN and ARM.

DRM has comparable inference time to ERM. The time cost of prediction for one data sample in the RotatedMNIST, PACS, VLCS, and DomainNet datasets are shown in Table 9. DRM will not introduce significant computational overhead even on the DomainNet dataset, which has the most number of domains.

Softing mixed weights. Figure 4 shows ablation experiments of the hyperparameter \( \gamma \) on three benchmarks. Different benchmarks show different preferences on \( \gamma \). For easy benchmarks, Rotated MNIST and Colored MNIST, softening mixed weights is needless. The reason behind this phenomenon can be found in Figure 3(d), the optimal classifier for the target domain 0 of the Rotated MNIST is exactly the classifier 1 and the prediction entropies will increase as the rotation angle increases. Hence, selecting the most approximate classifier based on the minimum entropy selection strategy is enough to attain superior generalization results. However, prediction entropies on other larger benchmarks, e.g., VLCS, are not so regular as on the Rotated MNIST. On realistic benchmarks, a mixing of classifiers can bring some improvements. Besides, normalization, which is a method to reduce classification confidence\(^1\), is also needless for semi-synthetic datasets (Rotated MNIST and Colored MNIST) and valuable for realistic benchmarks.

DRM brings faster convergence speed. The training dynamics of DRM and several baselines on PACS dataset are shown in Figure 4(d), where \( d = 0 \) is the target domain. IRM is unstable and hard to converge. ARM follows a meta-learning pipeline following the meta-learning pipeline \( (\text{ARM} [62]) \) have a much slower training speed compared to ERM. The proposed DRM, without the need for aligning representations [13], matching gradient [45], or learning invariant representations [3], has a training speed that is faster than most existing DG methods, especially on small datasets RotatedMNIST. The training speed of DRM is slower than ERM due to the training of additional classifiers. As the number of domains/classes increases or the feature dimension increases, the training time of DRM will increase accordingly; however, DRM is always comparable to ERM and much faster than Fish and ARM (Table 7). For model parameters, since all classifiers in our implementation are just a linear layer, the total parameters of DRM is similar to ERM and much less than existing methods such as CDANN and ARM.

Table 9: Comparison between inference times of one data sample in milliseconds.

| Dataset   | DRM | ERM |
|-----------|-----|-----|
| CMNIST    | 1.91| 1.29|
| RMNIST    | 3.31| 1.26|
| PACS      | 10.74| 9.81|
| VLCS      | 10.74| 8.64|
| DomainNet | 11.15| 9.34|

\(^1\)Given two classification results from 2 classifiers \([2.1, 0.4, 0.5], [0.3, 0.6, 0.1]\) and assume the weights are all 1. The result is \([2.4, 1.0, 0.6]\) with normalization and \([1.0, 0.73, 0.27]\) without normalization. The former is more confident than the latter.
Table 6: Comparison of different test-time model selection strategies on the PACS dataset.

| Method          | A      | C      | P      | S      | Avg   |
|-----------------|--------|--------|--------|--------|-------|
| DRM w/ Uniform weight | 81.2 ± 2.2 | 71.2 ± 1.2 | 93.7 ± 0.3 | 78.6 ± 1.5 | 81.2   |
| DRM w/ CSM      | 83.0 ± 2.1 | 74.6 ± 2.5 | 95.6 ± 0.8 | 80.4 ± 1.2 | 83.4   |
| DRM w/ NNM      | 85.5 ± 2.4 | 76.8 ± 2.0 | 96.6 ± 0.4 | 81.8 ± 1.5 | 85.2   |
| DRM w/ L2SM     | 87.7 ± 1.7 | 80.0 ± 0.5 | 96.0 ± 1.6 | 82.1 ± 1.2 | 86.5   |
| DRM w/ PEM      | 88.3 ± 2.9 | 80.1 ± 0.8 | 97.0 ± 0.5 | 80.9 ± 0.7 | 86.6   |

Table 7: Comparisons of different methods on the number of parameters and training time.

| Method   | Colored MNIST | Rotated MNIST | PACS |
|----------|---------------|---------------|------|
|          | Time (sec)    | # Params (M)  | Time (sec)    | # Params (M)  | Time (sec)    | # Params (M)  |
| ERM      | 71.02         | 0.3542        | 168.32        | 0.3546        | 2,717.5       | 22.4326       |
| IRM      | 101.49        | 0.3542        | 236.80        | 0.3546        | 2,786.3       | 22.4326       |
| ARM      | 161.51        | 0.4573        | 360.69        | 0.4562        | 6,616.9       | 22.5398       |
| FISH     | 137.17        | 0.3542        | 251.76        | 0.3546        | 23,849.5      | 22.4326       |
| DRM      | 83.39         | 0.3544        | 203.15        | 0.3595        | 2,895.1       | 22.46         |

Table 8: Generalization performance on multiple unseen target domains.

| Method | Rotated MNIST |
|--------|---------------|
|        | Target domains \{0, 30, 60\} | Target domains \{15, 45, 75\} | Avg   |
| ERM    | 96.0±0.3       | 98.8±0.4       | 98.7±0.1       | 98.8±0.3       | 99.1±0.1       | 96.7±0.3       | 98.0         |
| IRM    | 80.9±3.2       | 94.7±0.9       | 94.3±1.3       | 94.3±0.8       | 95.5±0.5       | 91.1±3.1       | 91.8         |
| DRM    | 97.1±0.2       | 98.8±0.2       | 98.9±0.3       | 98.8±0.1       | 98.8±0.0       | 98.1±0.7       | 98.4         |

Figure 4: Different mixing weights on the (a) Colored MNIST (target domain \(d = 2\)) (b) Rotated MNIST (target domain \(d = 0\)), and (c) PACS datasets (target domain \(d = 3\)). Given a classification vector \(\hat{y} = [y_1, y_2, ..., y_c]\), \(c\) is the number of classes, performing normalization means that let \(y_i = y_i / \sum_{j=1}^{c} y_j\) before mixing. (d) Loss curves of different baselines.

6 CONCLUDING REMARKS

We theoretically and empirically study the importance of the adaptivity gap for domain generalization. Inspired by our theory, we propose a new domain generalization algorithm, DRM to eliminate the negative effects brought by the adaptivity gap. DRM uses different classifier combinations for different test samples and beats existing DG methods and TTA methods by a large margin.

Existing TTA methods for domain generalization need to adapt model parameters continually, therefore, the prediction behavior cannot be thoroughly tested in advance, causing some ethical concerns [19]. DRM alleviates this important issue because model retraining is not necessary. One potential drawback is the additional parameters incurred by the multi-classifiers structure, which can be reduced by advanced techniques and model designs, e.g., varying coefficient technique [16, 37].

ACKNOWLEDGEMENTS

This work was partially funded by the National Key R&D Program of China (2022ZD0117901), and National Natural Science Foundation of China (62236010, and 62141608).
[56] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. 2020. DoFe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging* 39, 12 (2020), 4237–4248.

[57] Shan Yan, Huan Song, Nanxiang Li, Lincan Zou, and Lu Ren. 2020. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.08677* (2020).

[58] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. 2021. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8978–8987.

[59] Nanyang Ye, Kaizan Li, Haoyue Bai, Fanpeng Yu, Lanqing Hong, Fengwei Zhou, Zhengu Li, and Jun Zhu. 2022. OoD-Bench: Quantifying and Understanding Two Dimensions of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7947–7958.

[60] Hanlin Zhang, Yi-Fan Zhang, Weiying Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. 2022. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[61] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. 2013. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, PMLR, 819–827.

[62] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. 2021. Adaptive risk minimization: Learning to adapt to domain shift. *NeurIPS* (2021).

[63] Yi-Fan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. 2021. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249* (2021).

[64] Yi-Fan Zhang, Feng Li, Zhang Zhang, Liang Wang, Dacheng Tao, and Tieu-Tan. 2022. Generalizable Person Re-identification Without Demographics. https://openreview.net/forum?id=VNdFhIS5wq5h

[65] Yi-Fan Zhang, Xue Wang, Jian Liang, Zhang Zhang, Liang Wang, Rong Jin, and Tieu-Tan. 2023. Free Lunch for Domain Adversarial Training: Environment Label Smoothing. *International Conference on Learning Representations (ICLR)* (2023).

[66] Yi-Fan Zhang, Hanlin Zhang, Zachary Chase Lipton, Li Ernan Li, and Eric Xing. 2022. Exploring transformer backbones for heterogeneous treatment effect estimation. In *NeurIPS ML Safety Workshop*.

[67] Yi-Fan Zhang, Xue Wang, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieu-Tan. 2023. AdaNPC: Exploring Non-Parametric Classifier for Test-Time Adaptation. *ICML* (2023).

[68] Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieu-Tan. 2022. Learning domain invariant representations for generalizable person re-identification. *IEEE Transactions on Image Processing* (2022).

[69] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. 2019. On learning invariant representations for domain adaptation. In *ICML*, PMLR.
A PROOFS OF THEORETICAL STATEMENTS

To complete the proofs, we begin by introducing some necessary definitions.

**Definition 3.** (\(\mathcal{H}\)-divergence [5].) Given two domain distributions \(D_S, D_T\) over \(X\), and a hypothesis class \(\mathcal{H}\), the \(\mathcal{H}\)-divergence between \(D_S, D_T\) is \(d_H(D_S, D_T) = 2\sup_{f \in \mathcal{H}} |\mathbb{E}_{X \sim D_S}[f(x)] = 1 - \mathbb{E}_{X \sim D_T}[f(x)] = 1|\).

A.1 Derivation and Explanation of the Learning Bound in Eq. 2

Let \(f^* = \arg\min_{f \in \mathcal{H}} \left( e_T(f) + \sum_{i=1}^{K} \varepsilon_i(f) \right)\), and let \(\lambda_T\) and \(\lambda_i\) be the errors of \(f^*\) with respect to \(D_T\) and \(D_i\) respectively. Notice that \(\lambda_a = \lambda_T + \sum_{i=1}^{K} \lambda_i\). Similar to [5] (Theorem 1), we have

\[
e_T(f) \leq \lambda_T + \sum_{i=1}^{K} \lambda_i + \sum_{i=1}^{K} \varepsilon_i(f) + d_H(\mathcal{D}_a^T, \mathcal{D}_a) + d_H(\mathcal{D}_a, \mathcal{D}_T^T) \leq \lambda_T + \sum_{i=1}^{K} \lambda_i + \sum_{i=1}^{K} \varepsilon_i(f) + d_H(\mathcal{D}_a^T, \mathcal{D}_a) + d_H(\mathcal{D}_a, \mathcal{D}_T^T)
\]

(7)

The forth inequality holds because of the triangle inequality. We provide the explanation for our bound in Eq. 7. The second term is the empirical loss for the convex combination of all source domains. The third term corresponds to “To what extent can the convex combination of the source domain approximate the target domain”. The minimization of the third term requires diverse data or strong data augmentation such that the unseen distribution lies within the convex combination of source domains. For the fourth term, the following equation holds for any two distributions \(D_T', D_T''\), which are the convex combinations of source domains [2]

\[
d_H(D_T', D_T'') \leq \sum_{i=1}^{K} \sum_{k=1}^{K} a_i a_k d_H(D_i, D_k)
\]

(8)

The upper bound will be minimized when \(d_H(D_i, D_k) = 0, \forall i, k \in \{1, \ldots, K\}\). That is, projecting the source domain data into a feature space where the source domain labels are hard to distinguish.

A.2 Derivation the Learning Bound in Eq. 5

**Proposition 3.** Let \(\{D_i, f_i\}_{i=1}^{K}\) and \(D_T, f_T\) be the empirical distributions and the corresponding labeling function. For any hypothesis \(\hat{f} \in \mathcal{H}\), given mixed weights \((a_i)_{i=1}^{K}, \sum_{i=1}^{K} a_i = 1, a_i \geq 0\), we have:

\[
e_T(\hat{f}) \leq \sum_{i=1}^{K} \mathbb{E}_{X \sim D_i} \left[ a_i P_T(X) f(X) - f_i(X) + a_i d_H(D_i, D_T) \right]
\]

(9)

Then we will prove that Eq. 9 is upper bounded by Eq. 10. At first, the second line in Eq. 11 is bounded by

\[
e_T(\hat{f}, f_S) \leq \mathbb{E}_{X \sim D_S} \left[ a_i P_T(X) f(X) - f_S(X) + \varepsilon_S(f_X, f_T) \right] = \mathbb{E}_{X \sim D_S} \left[ a_i P_T(X) f(X) - f_S(X) + d_H(D_S, D_T) \right]
\]

(10)

**Theorem 1.** (Lemma 4.1 and Theorem 4.1 in [1].) Given two distributions in the image space \(\langle D_S, f_S \rangle, \langle D_T, f_T \rangle\) and \(\hat{f} \in \mathcal{H}\), we have

\[
|\varepsilon_S(f_S, f_T) - \varepsilon_T(\hat{f})| \leq d_H(D_S, D_T).
\]

(11)

The error in the target domain can then be bounded by

\[
e_T(\hat{f}) \leq \varepsilon_S(f_S, f_T) + d_H(D_S, D_T) + \min(\varepsilon_S(f_S, f_T), \varepsilon_T(\hat{f}))
\]

(12)

where the result is based mainly on the inequality in Eq. 9.

A.3 Comparison of the Proposed Bound to Existing Bound.

Before we derive our main result, we first introduce some necessary theorems. For simplicity, given hypothesis \(\hat{f}, f^* \in \mathcal{H}\) and label function \(f_S\) for \(D_S\), denote \(\varepsilon_S(f, f^*) = \mathbb{E}_{Z \sim D_S}[|f - f^*|]\) and \(\varepsilon_S(f) = \mathbb{E}_{Z \sim D_S}[|f - f_S|]\), we have

\[
e_T(\hat{f}) \leq \mathbb{E}_{X \sim D_T}[|\hat{f}(X) - f_T(X)|] = \sum_{i=1}^{K} a_i \mathbb{E}_{X \sim D_T}[|\hat{f}(X) - f_T(X)|] + \sum_{i=1}^{K} \mathbb{E}_{X \sim D_T}[|f_i(X) - f_T(X)|]
\]

(10)

which completes our proof.
Also, since the density ratio $\frac{P_T(X)}{P_S(X)}$ is intractable and during implementation, this term is set to a constant and ignored. That is, the last line of Eq. 11 is approximately equal to

$$
\mathbb{E}_{X \sim D_{S}} \left[ \frac{P_T(X)}{P_S(X)} \right] \left| \hat{f} - f_S \right| + \epsilon_T(f_S, f_T)
$$

$$
\leq \epsilon_S(\hat{f}) + \epsilon_T(f_S, f_T)
$$

Combining Eq. 12 and Eq. 13 we can get the error bound in Eq. 11 is upper bounded by $\epsilon_S(\hat{f}) + d_H(D_S, D_T) + \min\{\epsilon_S(f_S, f_T), \epsilon_T(f_S, f_T)\}$, which completes our proof.

### A.4 Reformulation of the Density Ratio.

In this subsection, we first introduce some important definitions of the distributionally robust optimization (DRO) framework [6] and then reformulate the density ratio under some necessary assumptions. In DRO, the expected worst-case risk on a predefined family of distributions $Q$ (termed uncertainty set) is used to replace the expected risk on the unseen target distribution $T$ in ERM. Therefore, the objective is as follows.

$$
\min_{\theta \in \Theta} \max_{Q \in Q} \mathbb{E}_{(x, y) \in Q} \left[ \ell(x, y; \theta) \right].
$$

Specifically, the uncertainty set $Q$ encodes the possible test distributions on which we want our model to perform well. If $Q$ contains $T$, the DRO object can upper bound the expected risk under $T$.

The construction of uncertainty set $Q$ is of vital importance. Here we reformulate the density ratio based on the KL-divergence ball constraint and other choices (e.g., using the moment constraint [9], $f$-divergence [34], Wasserstein/MMD ball [46, 47]) will lead to different reweighting methods. Given the KL upper bound (radius) $\eta$, denote the empirical distribution $P$, we have the uncertainty set $Q = \{Q : \text{KL}(Q||P) \leq \eta\}$. The Min-Max Problem in Eq. 14 can then be reformulated as

$$
\min_{\theta \in \Theta} \max_{Q \in Q : \text{KL}(Q||P) \leq \eta} \mathbb{E}_{(x, y) \in Q} \left[ \ell(x, y; \theta) \right].
$$

Then we have the following theorem, which derives the optimal density ratio and converts the original problem to a reweighting version.

**Theorem 2.** (Modified from Section 2 in [17]) Assume the model family $\theta \in \Theta$ and $Q$ to be convex and compact. The loss $\ell$ is continuous and convex for all $x \in X$, $y \in Y$. Suppose that the empirical distribution $P$ has density $p(x, y)$. Then the inner maximum of Eq. 15 has a closed-form solution: $q^*(x, y) = \frac{P(x, y)}{p(x, y)} e^{r \ell(x, y; \theta)}$, where $r^*$ satisfies $\mathbb{E}_P \left[ e^{r \ell(x, y; \theta)} \right] = \frac{\epsilon_S(\hat{f}) + d_H(D_S, D_T) + \min\{\epsilon_S(f_S, f_T), \epsilon_T(f_S, f_T)\}}{\eta}$ and $q^*(x, y)$ is the optimal density of $Q$. The min-max problem in Eq. 15 is equivalent to

$$
\min_{\theta \in \Theta} \max_{r > 0} r \log \mathbb{E}_P \left[ e^{r \ell(x, y; \theta)/r} \right] + \eta r.
$$

### B DATASET AND IMPLEMENTATION DETAILS

#### B.1 Dataset Details

**Colored MNIST** [3] consists of digits in MNIST with different colors (blue or red). The label is a noisy function of the digit and color. First, a preliminary label $\tilde{y}$ is assigned to images based on their digits, $\tilde{y} = 0$ for digits 0-4 and $\tilde{y} = 1$ for digits 5-9. The final label is obtained by flipping $\tilde{y}$ with probability 0.25. The color signal $z$ of each sample is obtained by flipping $\tilde{y}$ with probability 0.25. The color signal $z$ of each sample is obtained by flipping $\tilde{y}$ with probability 0.25. The color signal $z$ of each sample is obtained by flipping $\tilde{y}$ with probability 0.25.

**Rotated MNIST** [14] consists of 10,000 digits in MNIST with different rotated angles where the domain is determined by the degrees of each sample is obtained by flipping $\tilde{y}$ with probability 0.25.

**PACS** [24] includes 9,991 images with 7 classes $\{\text{dog, elephant, giraffe, guitar, horse, house, person}\}$ from 4 domains $d \in \{\text{art, cartoons, photos, sketches}\}$.

**VLCS** [52] is composed of 10,729 images, 5 classes $\{\text{bird, car, chair, dog, person}\}$ from domains $d \in \{\text{Caltech101, LabelMe, SUN09, VOC2007}\}$.

**DomainNet** [40] has six domains $d \in \{\text{clipart, infographic, painting, quickdraw, real, sketch}\}$. This dataset contains 586,575 examples of size (3, 224, 224) and 345 classes.