Domain-Specific Risk Minimization

Yi-Fan Zhang[1], Jindong Wang[2], Zhang Zhang[1], Baosheng Yu[3],
Liang Wang[1], Dacheng Tao[3], Xing Xie[2]
[1]Institute of Automation, Chinese Academy of Sciences
[2]Microsoft Research Asia, [3]The University of Sydney.

Abstract

*Learning a domain-invariant representation has become one of the most popular approaches for domain adaptation/generalization. In this paper, we show that the invariant representation may not be sufficient to guarantee good generalization, where **labeling function shift** should be considered. Inspired by this, we first derive a new generalization upper bound on the empirical risk by explicitly considering the labeling function shift. We then propose **Domain-specific Risk Minimization (DRM)** to tackle such shift. DRM can model the distribution shifts of different domains separately and select the most appropriate one for the target domain. Extensive experiments on four popular domain generalization datasets, namely, CMNIST, PACS, VLCS, and DomainNet, demonstrate the effectiveness of DRM for domain generalization with the following advantages: 1) it significantly outperforms competitive baselines; 2) it enables either comparable or superior performance on all training domains comparing to vanilla empirical risk minimization (ERM); 3) it remains very simple and efficient during training, and 4) it is complementary to invariant learning approaches.*

1. Introduction

Domain generalization (DG) [41] aims to learn a generalized model that performs well for unseen domains. Most deep learning-based DG methods seek to learn an *invariant representation* [4, 21, 26, 31], where the feature distributions among all training domains are the same. However, without accessing the data on the target domain, feature alignment can be performed only among source domains, which inevitably raises a question: *is the representation that is invariant to the source domain shift really good enough for unseen domain generalization?*

In an attempt to answer this question, Zhao et al. [50] considers the conditional shift in domain adaptation and shows that only learning invariant representation is insufficient. A surge of methods are then proposed to tackle this problem. However, the target domain is unseen for DG, i.e., its labeling function is totally not accessible, which makes it more challenging to consider labeling function shift. Therefore, most DG methods [2, 9, 30] ignore such shift.

In this paper, we first show through a counterexample that the ignorance of labeling function shift will lead to significantly large errors on all domains even if the domain-invariant representations are well learned. Then, we propose a new generalization error bound to tackle labeling function shifts. The bound is proven tighter than that in [50]. Specifically, an intuitive explanation of the new generalization bound is: *since we cannot guarantee that all labeling functions are the same, we would rather model all labeling functions and choose the most appropriate one for a good generalization for inference*.

Motivated by the proposed error bound, we propose a new DG approach called **Domain-Specific Risk Minimization (DRM)** to reduce the negative impact of domain labeling function shift, which can be easily incorporated into most deep representation learning algorithms. DRM introduces a shared encoder for all source domains with a group of domain-specific classifiers during training. Specifically, each domain-specific classifier is responsible for the labeling function on a specific source domain. During testing, we further propose three test-time model selection strategies for classifier selection. Our contributions are three-fold:

A new perspective. We show the insufficiency of invariant representations and provide a new generalization bound to explicitly consider the conditional shift for DG.

A new approach. We propose DRM to model all labeling functions in a domain-specific manner. The proposed model structure and test-time selection strategy are orthogonal to most of existing methods.

Extensive experiments. Extensive experiments on popular DG benchmarks show that DRM (1) achieves competitive generalization performance; (2) is orthogonal to other DG methods; (3) reserves strong recognition capability on source domains, and (4) is parameter-efficient.

2. Domain Generalization Bound

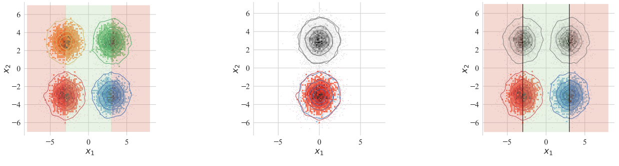
Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ denote the input, output, and feature space, respectively. Let X, Y, Z denote the random variables taking values from $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively. Each domain corresponds to a joint distribution $P_i(X, Y)$ with a labeling function $f_i : \mathcal{X} \to [0,1]$[1]. In the DG setting, we have access to a labeled training dataset that consists of several different but related training distributions (domains): $\mathcal{D} = \cup_{i=1}^{K} \mathcal{D}_i$, where K is the number of domains. In this paper, we focus on a deterministic setting where the output $Y = f_i(X)$ is given by a deterministic labeling function, f_i, which varies from domain to domain. Let $g : \mathcal{X} \to \mathcal{Z}$ denote the encoder/feature transformation and $h : \mathcal{Z} \to \{0,1\}$ denote the classifier/hypothesis. The error incurred by $h \circ g$ under domain $\mathcal{D}_i$ can be defined as $\epsilon_i(h \circ g) = \mathbb{E}_{X \sim \mathcal{D}_i}[|h \circ g(X) - f_i(X)|]$. Given f_i and h as binary classification functions, we have

$$\begin{aligned} \epsilon_i(h \circ g) = \epsilon_i(h \circ g, f_i) &= \mathbb{E}_{X \sim \mathcal{D}_i}\left[|h \circ g(X) - f_i(X)|\right] \\ &= \mathrm{Pr}_{X \sim \mathcal{D}_i}(h \circ g(X) \neq f_i(X)). \end{aligned} \tag{1}$$

During training, $h \circ g$ is trained using all image-label pairs from $\mathcal{D}$. During testing, we perform a retrieval task on the unseen target domain $\mathcal{D}_\mathcal{T}$ without additional model updating and we aim to minimize the error in $\mathcal{D}_\mathcal{T}$: $\min_{h \circ g} \epsilon_\mathcal{T}(h \circ g)$.

2.1. A Failure Case of Invariant Representation

Objective Eq.(1) encodes the goal of learning a model with domain invariant representations [13, 21, 22]. Specifically, a parametric feature transformation $g : \mathcal{X} \to \mathcal{Z}$ is learned such that the induced source distributions on $\mathcal{Z}$ are close to each other. Besides, a hypothesis h over the feature space $\mathcal{Z}$ is found to achieve small empirical errors on source domains. These studies get intuition from the error bound in [5, 46] and for completeness, we show the bound in Appendix C.2. However, we show a counterexample with two source domains and two target domains in Figure 1, where even the optimal invariant representation in Figure 1(b) leads to large errors on both source and target domains (Refer to Appendix A for the details).



(a) (b) (c)

Figure 1. **A failure case of invariant representations for domain generalization.** (a) Four domains in different colors: orange ($\mu_o = [-3.0, 3.0]$), green ($\mu_g = [3.0, 3.0]$), red ($\mu_r = [-3.0, -3.0]$) and blue ($\mu_b = [3.0, -3.0]$). (b) Invariant representations learnt from domain $\mathcal{D}_r$ and $\mathcal{D}_b$ by feature transformation $g(X) = \mathbb{I}_{x_1 < 0} \cdot (x_1 + 3) + \mathbb{I}_{x_1 > 0} \cdot (x_1 - 3)$. The grey color indicates the transformed target domains. (c) The classification boundary learnt by DRM.

2.2. A Bound by Labeling Function Shift

Motivated by the example, we next provide a tighter upper bound for DG that considers labeling function shifts.

Proposition 1. *Let $\{\mathcal{D}_i, f_i\}_{i=1}^{K}$ and $\mathcal{D}_\mathcal{T}, f_\mathcal{T}$ be the empirical distributions and corresponding labeling function. For any hypothesis $h \in \mathcal{H}$ and transformation g, given mixed weights $\{\alpha_i\}_{i=1}^{K}; \sum_{i=1}^{K} \alpha_i = 1, \alpha_i \geq 0$, we have:*

$$\begin{aligned} \epsilon_\mathcal{T}(h \circ g) \leq \sum_{i=1}^{K} (\mathbb{E}_{X \sim \mathcal{D}_i}\left[\alpha_i \frac{P_\mathcal{T}(X)}{P_i(X)} |h \circ g - f_i|\right] + \\ \alpha_i \mathbb{E}_{\mathcal{D}_\mathcal{T}}[|f_i - f_\mathcal{T}|]). \end{aligned} \tag{2}$$

See Appendix C.3 for the proof and interpretations.

Although labeling function shift has been considered in domain adaptation error bound [3], in Appendix C.4, we show that the proposed bound is tighter. Besides, the proposed bound can supply a novel perspective for aligning labeling function shifts.

Remark. Eq. (2) provides a new intuition on the design of DG models. Specifically, the density ratio $P_\mathcal{T}(x)/P_i(x)$ has a strong connection with reweighting methods and provides a theoretical explanation for why reweighting data samples works well on DG (See Appendix C.5 for details). The labeling functions $f_i, f_\mathcal{T}$ are constant and cannot be optimized, and **we focus on mixed weights α_i and $h \circ g$**. The first term will be minimized when $h \circ g$ attains low errors in source domains. The second term cannot be optimized directly, however, we can manipulate α to affect this term as follows. Given $f_\mathcal{T}$, if we can find the source domain $\mathcal{D}_{i^*}$ with a labeling function f_{i^*} that minimizes $\mathbb{E}_\mathcal{T}[|f_{i^*} - f_\mathcal{T}|]$, then we have that $\alpha_i = 1$, iff $i = i^*$, otherwise 0 makes this term the minimum. As a whole algorithm, these two procedures correspond to simultaneously finding the domain $\mathcal{D}_{i^*}$ whose labeling function is close to $f_\mathcal{T}$, setting $\alpha_{i^*} = 1$ and learning $h \circ g$ on $\mathcal{D}_{i^*}$ to minimize the source error. Namely, as long as we can accurately estimate $\mathbb{E}_\mathcal{T}[|f_i - f_\mathcal{T}|]$, only one domain is required for training to minimize the error in the target domain. However, calculating $\mathbb{E}_\mathcal{T}[|f_{i^*} - f_\mathcal{T}|]$ is intractable especially when $\mathcal{D}_\mathcal{T}$ is unseen during training. To tackle the challenge and follow the intuition brought by Eq.(2), we propose a new Domain-Specific Risk Minimization (DRM) method for domain generalization.

3. Domain-Specific Risk Minimization

The main pipeline of the proposed Domain-Specific Risk Minimization (DRM) is shown in Figure 2.

3.1. Domain-specific labeling function

One of our main contributions is the modeling of **domain-specific labeling function**. Specifically, given K source

[1] Most theories and examples in this paper considers binary classification for easy understanding and can be easily extended to multi-class classification.

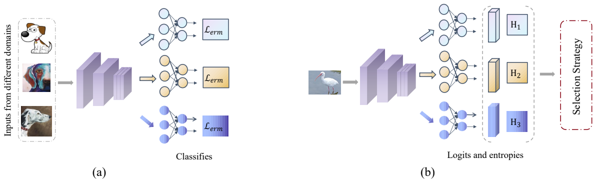


Figure 2. **An illustration of the training and testing pipelines using DRM.** (a) during training, it jointly optimizes an encoder shared by all domains and the specific classifiers for each individual domain. $\mathcal{L}_{erm}$ indicates the cross-entropy loss function. (b) the new image is first classified by all classifiers and a test-time model selection strategy is applied to generate the final result.

domains, DRM utilizes a shared encoder g and a group of classifiers $\{h_i\}_{i=1}^K$ for all domains, respectively. The encoder is trained by all data samples while each classifier h_i is trained by using only images from the domain $\mathcal{D}_i$. If we go back to Eq.(2), with domain-specific classifiers, we then have the bound

$$\sum_{i=1}^{K} \alpha_i \left(\mathbb{E}_{x\sim\mathcal{D}_i} \left[\frac{P_{\mathcal{T}}(x)}{P_i(x)} |h_i \circ g - f_i| \right] + \mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[|f_i - f_{\mathcal{T}}|] \right). \tag{3}$$

Therefore, Eq.(3) shows that it is rather possible to achieve low errors on source domains by using the domain-specific classifiers than just one hypothesis h. It is also possible but not efficient to use specific $h_i \circ g_i$ for each domain. However, we observe that, on the Colored MNIST dataset, it achieves the generalization accuracy 64.8% when using specific $h_i \circ g_i$, while it is 70.1% for using specific h_i. A possible reason is that a shared encoder g can be seen as an implicit regularization, which prevents the model from overfitting specific domains.

3.2. Test-time model selection

We do not aim at a lower source error but also want to know *"how to determine mixed weights α such that low target domain error can be achieved?"*. As mentioned above, the second term $\alpha_i\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[|f_i - f_{\mathcal{T}}|]$ cannot be optimized directly, however, we can manipulate α_i to affect this term: for every test sample $x \in \mathcal{D}_{\mathcal{T}}$, if we can estimate $\{H_i = |f_i(x) - f_{\mathcal{T}}(x)|\}_{i=1}^K$ and choose $i^* = \arg\min\{H_i\}_{i=1}^K$. Then $\alpha_i = 1, \text{iff } i = i^*, \text{otherwise } 0$ makes this term the minimum and the final prediction result will be $f_{i^*} \circ g(x)$. The challenge here is estimating $\{H_i\}_{i=1}^K$. To this end, we propose three different techniques to estimate $\{H_i\}^K$ given an assumption "*the learnt $h_i \circ g$ can well approximate f_i*".

Similarity Measurement (SM). We first reformulate $\alpha_i\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[|f_i - f_{\mathcal{T}}|]$ as follows:

$$\begin{aligned}\alpha_i\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[|f_i - f_{\mathcal{T}}|] &= \alpha_i\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}\left[|f_i - \mathbb{E}_{\mathcal{D}_i}[f_i] + \mathbb{E}_{\mathcal{D}_i}[f_i] - f_{\mathcal{T}}|\right] \\ &\leq \alpha_i \left(\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}\left[|f_i - \mathbb{E}_{\mathcal{D}_i}[f_i]|\right] + \mathbb{E}_{\mathcal{D}_{\mathcal{T}}}\left[|\mathbb{E}_{\mathcal{D}_i}[f_i] - f_{\mathcal{T}}|\right]\right),\end{aligned} \tag{4}$$

where $f_{\mathcal{T}}$ is intractable and we then focus on $\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}\left[|f_i - \mathbb{E}_{\mathcal{D}_i}[f_i]|\right]$, which intuitively measures the prediction difference of the given test data $x \in \mathcal{D}_{\mathcal{T}}$ and the average prediction result in domain $\mathcal{D}_i$. However, take average of the prediction labels is meaningless[2] and we use $\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}\left[g - \mathbb{E}_{\mathcal{D}_i}[g]|\right]$ to approximate this term, where we calculate the representation difference between the test sample and average representations of domain $\mathcal{D}_i$. Estimation H_i here is 1 minus the representation similarity between $g(x); x \in \mathcal{T}$ and domain $\mathbb{E}_{\mathcal{D}_i}[g]$. The similarity can be calculated by any distance metric such as l_p-Norm, cosine similarity, $f-$divergence, MMD/$\mathcal{A}$ distance, and we use cosine similarity (CSM) and l_2-Norm (L2SM) in our experiments for example.

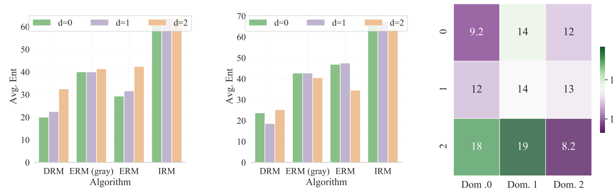
Prediction Entropy Measurement (PEM). Given the following assumption: *"the more confident prediction $h_i \circ g$ makes on $\mathcal{D}_{\mathcal{T}}$, the more similar f_i and $f_{\mathcal{T}}$ will be"*. We then have, during testing, the K individual classification logits as $\{\bar{\mathbf{y}}^k\}_{k=1}^K$, where $\bar{\mathbf{y}}^k = [y_1^k, ..., y_c^k]$, and c is the number of classes. Then, the prediction entropy of $\bar{\mathbf{y}}^k$ can be calculated as $H_k = -\sum_{i=1}^{c} \frac{y_i^k}{\sum_{j=1}^c y_j^k} \log \frac{y_i^k}{\sum_{j=1}^c y_j^k}$, where the entropy is used as our expected estimation. In our experiments, we find that the prediction entropy consistent with domain similarities, which is similar to SM.

Neural Network Measurement (NNM). To fully utilize the modeling capability of neural network, we finally propose to estimate $\alpha_i\mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[|f_i - f_{\mathcal{T}}|]$ by NN. Specifically, during training, a domain discriminator is trained to classify which domain is each data sample from. During test, for $x \in \mathcal{D}_{\mathcal{T}}$, the classification vector of the discriminator will be $\{d_i\}_{i=1}^K$, and $\{H_i = -d_i\}_{i=1}^K$ is used as the estimation.

Model Ensembling. A one-hot mixed weight is too deterministic and cannot fully utilize all learned classifiers. **Softing mixed weights** can further boost generalization performance, *i.e.* for ERM, we can generate the final prediction as $\sum_{k=1}^K \bar{\mathbf{y}}_k \frac{H_k^{-\gamma}}{\sum_{i=1}^K H_i^{-\gamma}}$, where $H_k^{-\gamma}$ indicates the contribution of each classifier. We use $-\gamma$ not γ because the smaller the predicted labeling function difference between f_i and $f_{\mathcal{T}}$, the larger the contribution of f_i should be. Specifically, for $\gamma = 0$, we then have a uniform combination, *i.e.* $\alpha_i = 1/K, \forall i \in [1, 2, ..., K]$; for $\gamma \to \infty$, we then have a one-hot weight vector with $\alpha_i = 1$ iff $i = i^*$ otherwise 0.

Remark. By modeling domain-specific labeling functions, DRM can further reduce source errors (*i.e.* the first term in our upper bound); For the second term, the test-time model selection strategies strategy allows us to select appropriate mixed weights and avoid directly calculating labeling function difference. In Appendix B, we show that DRM performs well on the counterexample, where invariant learning fails. Refer to Appendix Algorithm 1,2,3 for the detail of the

[2] if all source domain has 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.



(a) (b) (c)

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 v_{ij} is the entropy of predictions incurred by predicting samples in domain i with classifier j. Dom.i indicates the classifier for the domain $d = i$.

training and test pipelines of the proposed three strategies. In the experimental section, we compare the proposed three strategies and PEM generally performs the best, thus we later use PEM as the default choice.

3.3. Case Studies

In this subsection, we perform case study analysis on the Colored MNIST dataset [4], where spurious correlations are manually created and can thus be a good indicator, to verify the following remarks:

- *DRM has better generalizability than invariant learning-based methods.*
- *DRM retains high accuracies on source domains and is orthogonal to invariant learning-based methods.*
- *PEM implicitly reduces prediction entropy and the entropy-based strategy performs well on finding a proper labeling function for inference.*

As shown in Table 1, ERM achieves high accuracies on training domains but below-chance accuracy on the test domain due to relying on the spurious correlations. IRM forms a tradeoff between training and testing accuracy [4]. 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 a hypothetical model that not only knows all spurious correlations but also has no modeling capability limit. For averaged generalization performance, DRM, without any invariance regularization, outperforms IRM by a large margin (more than 2.4%). Besides, the training accuracy attained by DRM is even higher than ERM and significantly higher than IRM and OIM. Note that DRM is complementary with invariant learning-based methods, where incorporating CORAL [37] can further boost both training and testing performances. Though the Colored MNIST dataset is a good indicator to show the model capacity for avoiding spurious correlation, these spurious correlations therein are unrealistic and utopian. Therefore, when testing on large DG benchmarks (*e.g.* PACS, VLCS, DomainNet), ERM outperforms IRM. Different from them, DRM not only performs well on the semi-synthetic dataset but also attains SOTA performance on large benchmarks.

	+90% ($d = 0$)		**+80%** ($d = 1$)		**-90%** ($d = 2$)		Avg	
Method	train	test	train	test	train	test	train	test
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	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
DRM+CORAL	83.4±8.6	85.3±2.3	9**1.6±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 1. **Accuracies (%) of different methods on training/testing domains for the Colored MNIST synthetic task.** OIM (optimal invariant model) and RG (random guess) are hypothetical.

The prediction entropy is often related to the fact that more confident predictions tend to be correct [40]. In Figure 3a, we find that the entropy in target domain ($d = 2$) tends to be greater than the entropy in source domains, where the source domain with stronger spurious correlations ($d = 1$) also has larger entropy than easier one ($d = 0$). Fortunately, with the entropy minimization strategy, we can find the most confident classifier for a given data sample, and DRM can reduce the entropy of predictions (Figure 3b). To further analyze the entropy minimization strategy, we visualize the domain-classifier correlation matrix in Figure 3c, where the entropy between the domain and its corresponding classifier is minimal, verifying the efficiency of the entropy minimization strategy. Please refer to Appendix F.4 for more analysis on the domain-classifier correlation matrix.

We also conduct experiments on popular DG datasets, *e.g.* PACS, VLCS, and DomainNet, analyze the model complexity, training convergence, and compare the proposed three test-time selection strategies with ensembling learning baselines (Refer to Appendix F for the details). For space limit, we discuss related works of DG ethods, ensembling learning methods, and labeling function shift in Appendix E.

4. Conclusion

In this paper, we study the important problem of labeling function shifts for domain generalization theoretically and empirically. We first construct an example to show that learning an invariant representation without considering the labeling function shift is not sufficient for a good generalization. We then prove a novel upper bound for the target error, which motivates us to propose DRM to eliminate the negative effects brought by labeling function shifts. DRM achieves not only a superior generalization performance but also maintain low source errors simultaneously. We hope that our results can shed new light on the model design for domain generalization problems. One possible direction is to estimate $\alpha_i P_{\mathcal{T}}(x)/P_i(x)$ and then reweight data samples, which will be the subject of our future study.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *ICML*, 2020.

[2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

[3] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

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

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, 2006.

[6] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

[7] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[8] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 2021.

[9] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[10] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 2010.

[11] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017.

[12] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[14] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015.

[15] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[16] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[17] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV 2020*.

[18] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, 2021.

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[21] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018.

[22] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018.

[23] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

[24] Xiaofeng Liu, Bo Hu, Linghao Jin, Xu Han, Fangxu Xing, Jinsong Ouyang, Jun Lu, Georges EL Fakhri, and Jonghye Woo. Domain generalization under conditional and label shifts via variational bayesian inference. *arXiv preprint arXiv:2107.10931*, 2021.

[25] Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Modeling the second player in distributionally robust optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

[26] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[27] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, 2021.

[28] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[29] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint arXiv:2109.02934*, 2021.

[30] Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *arXiv preprint arXiv:2102.11436*, 2021.

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

[32] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *arXiv preprint arXiv:2011.12672*, 2020.

[33] 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.

[34] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[35] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[36] Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime G. Carbonell, and Kun Zhang. Domain adaptation with invariant representation learning: What transformations to learn? In *NeurIPS*, 2021.

[37] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

[38] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[39] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

[41] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[42] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12):4237–4248, 2020.

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

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

[45] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. 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*, pages 7947–7958, 2022.

[46] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. *arXiv preprint arXiv:2111.13839*, 2021.

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

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

[49] YiFan Zhang, Feng Li, Zhang Zhang, Liang Wang, Dacheng Tao, and Tieniu Tan. Generalizable person re-identification without demographics, 2022.

[50] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *ICML*. PMLR, 2019.