ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094659

A PRINCIPLED APPROACH TO MODEL VALIDATION IN DOMAIN GENERALIZATION

Boyang Lyu[†⋆] *Thuan Nguyen*[¶⋆] *Matthias Scheutz*[¶] *Prakash Ishwar*[‡] *Shuchin Aeron*[†]

[¶] Department of Computer Science, Tufts University, Medford, MA 02155
[†] Department of Electrical and Computer Engineering, Tufts University, Medford, MA 02155
[‡]Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215

ABSTRACT

Domain generalization aims to learn a model with good generalization ability, that is, the learned model should not only perform well on several seen domains but also on unseen domains with different data distributions. State-of-the-art domain generalization methods typically train a representation function followed by a classifier jointly to minimize both the classification risk and the domain discrepancy. However, when it comes to model selection, most of these methods rely on traditional validation routines that select models solely based on the lowest classification risk on the validation set. In this paper, we theoretically demonstrate a trade-off between minimizing classification risk and mitigating domain discrepancy, *i.e.,* it is impossible to achieve the minimum of these two objectives simultaneously. Motivated by this theoretical result, we propose a novel model selection method suggesting that the validation process should account for both the classification risk and the domain discrepancy. We validate the effectiveness of the proposed method by numerical results on several domain generalization datasets.

Index Terms— Domain generalization, model selection.

1. INTRODUCTION AND RELATED WORK

The success of traditional machine learning methods relies on an important assumption that the training and the test data are independent and identically distributed (*i.i.d*). However, in many real-world scenarios, the distributions of data in the training set and test set are not identical due to the "distribution-shift" phenomenon. Mitigating the problem caused by the distribution shift is the primary goal of the Domain Generalization (DG) problem, where a model is trained using data from several seen domains but later will be applied to unseen (unknown but related) domains with different data distributions.

To address DG problem, a large number of methods consider training a representation function that can learn domain-invariant features[1] by minimizing the domain discrepancy in the representation space [1–6]. Though the domain discrepancy has been accounted for at the training step, few works considered it for model selection at the validation step [7]. Indeed, following traditional machine learning settings, most of the state-of-the-art DG methods form a validation set using a small portion of data from all seen domains and select the model that achieves the lowest classification risk or highest classification accuracy on it. However, unlike the traditional machine learning settings where a model with lower classification risk on the validation set is likely to perform better on the test set, we theoretically show that for DG problem, where the *i.i.d* assumption does not hold, selecting the model with minimum classification risk may enlarge the domain discrepancy, subsequently leading to a non-optimal model on the unseen domain. We thus argue that one needs to consider both the classification risk and the domain discrepancy for selecting good models on unseen domains.

We summarize our contributions as follows:

1. We theoretically show that there is a trade-off between minimizing classification risk and domain discrepancy. This trade-off leads to the conclusion that only targeting a model with the lowest classification risk on the validation set may encourage distribution mismatch between domains (enlarging domain discrepancy), and reduce the model's generalization ability.
2. Based on our theoretical result and considering the limited attention given to DG-specific validation processes, we propose a simple yet effective validation/model selection method that integrates both the classification risk and domain discrepancy as the validation criterion. We further demonstrate the effectiveness of this approach on various DG benchmark datasets.

The trade-off between minimizing the classification risk and domain discrepancy has been mentioned in the literature [8, 9][2]. Shai *et al.* [8] constructed an upper bound on the risk of the target domain, composed of the risk from the source domain and the discrepancy between the target and source domains. The authors suggested that there must be a trade-off between minimizing the domain discrepancy and minimizing the seen domain's risk but did not propose any further details on how

[*]These authors contributed equally to this work.

[1] Domain-invariant features are the features having distributions that are unchanged and stable across domains.

[2] The works in [8, 9] are for domain adaptation, not domain generalization. However, one may derive a similar conclusion by replacing the "source domain" with seen domain and the "target domain" with unseen domain.

this trade-off is determined and characterized. Zhao *et al.* [9] showed that the sum of the risks from source and target domains is lower bounded by the distribution discrepancy between domains. If the discrepancy between domains is large, one can not simultaneously achieve small risks on both domains. Though sharing some similarities, our theoretical result differs from [9] since Zhao *et al.* considered the trade-off between minimizing the risks of different domains rather than the trade-off between optimizing the classification risk and the domain discrepancy. On the other hand, most DG works adopt the model selection methods following the traditional machine learning settings, *i.e.*, a validation set is first formed by combining small portions of data from all seen domains and the model that produces the lowest classification risk or highest classification accuracy on the validation set is then selected. To the best of our knowledge, there are only a few works that explore new model selection methods under DG settings [10–14]. The most related work of this study is [13], where the authors mentioned that they use the training loss (including both classification risk and adversarial domain discrepancy loss) on the validation set for model selection. However, it is not clear from their paper and their released code how the classification risk and the adversarial domain discrepancy loss are used to validate the model and how these two terms are balanced. In contrast, we propose an alternative approach for combining the classification risk and the domain discrepancy loss in a meaningful way in light of our theoretical results.

2. PROBLEM FORMULATION

2.1. Notations

Let $\mathcal{X}$, $\mathcal{Z}$, $\mathcal{Y}$ denote the input space, the representation space, and the label space, $\mathcal{D}^{(s)}$ and $\mathcal{D}^{(u)}$ represent the seen and unseen domain, respectively. $f : \mathcal{X} \rightarrow \mathcal{Z}$ and $g : \mathcal{Z} \rightarrow \mathcal{Y}$ are the representation function and the classifier. We use capital letters for the random variables in different spaces and lowercase letters for samples. Specifically, we denote X as the input random variable, Z as the extracted feature random variable, and Y as the label random variable. The input samples, feature samples, and labels of input samples are denoted as $\boldsymbol{x}, \boldsymbol{z}$, and $y(\boldsymbol{x})$, respectively. Finally, we use $p^{(s)}(\cdot)$ and $p^{(u)}(\cdot)$ to denote the distributions or joint distributions corresponding to the variables inside the bracket on seen domain and unseen domain, respectively.

2.2. Problem formulation

For a representation function f and a classifier g, the classification risk induced by f and g on seen domain is:

$$\begin{aligned} C^{(s)}(f,g) &= \int_{\boldsymbol{x}\in\mathcal{X}} p^{(s)}(\boldsymbol{x})\ell(g(f(\boldsymbol{x})), y^{(s)}(\boldsymbol{x}))d\boldsymbol{x} \\ &= \int_{\boldsymbol{x}\in\mathcal{X}}\int_{\boldsymbol{z}\in\mathcal{Z}} p^{(s)}(\boldsymbol{x},\boldsymbol{z})\ell(g(\boldsymbol{z}), y^{(s)}(\boldsymbol{x}))d\boldsymbol{x}d\boldsymbol{z} \quad (1) \end{aligned}$$

where $\ell(\cdot,\cdot)$ is a distance measure that quantifies the mismatch between the label outputted by classifier g and the true label.

For a representation function f, the distribution discrepancy between seen and unseen domains induced by f is:

$$D(f) = d(p^{(u)}(Y,Z)||p^{(s)}(Y,Z)) \quad (2)$$

where $d(\cdot||\cdot)$ is a divergence measure between two distributions. Indeed, to deal with the "distribution-shift" , one usually looks for a mapping f such that the discrepancy between distributions of seen and unseen domains $D(f)$ is small [15, 16].

A large number of DG works focus on training a model that minimizes both the classification risk $C^{(s)}(f,g)$ and the discrepancy $D(f)$ using data from seen domains [1–6]. Note that while $C^{(s)}(f,g)$ can be directly minimized, one usually need to approximately/heuristically optimize $D(f)$ by optimizing the distribution discrepancy between several seen domains. Since there are already well-established theoretical and empirical works on minimizing the classification risk and domain discrepancy, our work aims to highlight the trade-off between these two objectives (Sec 3) and argues that taking both objectives into account during model selection can improve model's performance on unseen domains (Sec. 4).

3. TRADE-OFF BETWEEN CLASSIFICATION RISK AND DOMAIN DISCREPANCY

We first begin with a definition.

Definition 1 (Classification risk-domain discrepancy function)**.** *For any representation function f and classifier g, define:*

$$\begin{aligned} &T(\Delta) = \min_{f:\mathcal{X}\rightarrow\mathcal{Z}} D(f) = \min_{f:\mathcal{X}\rightarrow\mathcal{Z}} d(p^{(u)}(Y,Z)||p^{(s)}(Y,Z)) \\ &\textit{s.t.} \quad C^{(s)}(f,g) = \int_{\boldsymbol{x}\in\mathcal{X}} p^{(s)}(\boldsymbol{x})\ell(g(f(\boldsymbol{x})), y^{(s)}(\boldsymbol{x}))d\boldsymbol{x} \leq \Delta \end{aligned} \quad (3)$$

where Δ is a positive number, $\ell(\cdot,\cdot)$ is a distance measure, and $d(\cdot||\cdot)$ is a divergence measure.

$T(\Delta)$ is the minimal discrepancy between the joint distribution of the unseen domain and seen domain if the classification risk on seen domain $C^{(s)}(f,g)$ does not exceed a positive threshold Δ. Next, we formally show that there is a trade-off between minimizing the distribution discrepancy $D(f)$ and minimizing the classification risk $C^{(s)}(f,g)$.

Theorem 1 (Main result)**.** *If the divergence measure $d(a||b)$ is convex (in both a and b), for a fixed classifier g, $T(\Delta)$ defined in (3) is monotonically non-increasing, and convex.*

Proof. The proof of this theorem is mainly based on the proposed approach in Rate-Distortion theory [17]. Particularly, consider two positive numbers Δ_1 and Δ_2, and assume $\Delta_1 \leq \Delta_2$. For a given classifier g, we use $\mathcal{F}_{\Delta_1}$ and $\mathcal{F}_{\Delta_2}$ to denote the sets of mappings f such that $C^{(s)}(f,g) \leq \Delta_1$ and $C^{(s)}(f,g) \leq \Delta_2$, respectively. First, we show that $T(\Delta)$ is non-increasing. Indeed, from $\Delta_1 \leq \Delta_2$, $\mathcal{F}_{\Delta_1} \subset \mathcal{F}_{\Delta_2}$:

$$\begin{aligned} T(\Delta_1) &= \min_{f\in\mathcal{F}_{\Delta_1}} d(p^{(u)}(Y,Z)||p^{(s)}(Y,Z)) \\ &\geq \min_{f\in\mathcal{F}_{\Delta_2}} d(p^{(u)}(Y,Z)||p^{(s)}(Y,Z)) = T(\Delta_2). \end{aligned}$$

Second, to prove the convexity of $T(\Delta)$, we show that:

$$\lambda T(\Delta_1)+(1-\lambda)T(\Delta_2) \geq T(\lambda\Delta_1+(1-\lambda)\Delta_2), \forall\lambda\in[0,1]. \quad (4)$$

To prove (4), we need some additional notations. Define:

$$f_1 = \underset{f:\mathcal{X}\to\mathcal{Z}}{\arg\min}\, D(f) \quad \text{s.t.} \quad C^{(s)}(f,g) \leq \Delta_1, \quad (5)$$

$$f_2 = \underset{f:\mathcal{X}\to\mathcal{Z}}{\arg\min}\, D(f) \quad \text{s.t.} \quad C^{(s)}(f,g) \leq \Delta_2. \quad (6)$$

Note that for any f, $Y \to X \to Z$ forms a Markov chain, thus:

$$p^{(u)}(Y,Z) = p^{(u)}(Y|X)\,p^{(u)}(X,Z), \quad (7)$$

$$p^{(s)}(Y,Z) = p^{(s)}(Y|X)\,p^{(s)}(X,Z), \quad (8)$$

where $p^{(u)}(Y|X)$ and $p^{(s)}(Y|X)$ are independent of f and only depend on the conditional distributions of label and data on seen and unseen domains.

Let $p_1^{(u)}(Y,Z)$, $p_1^{(s)}(Y,Z)$ be the joint distributions of Y and Z on unseen and seen domain produced by f_1, and similarly $p_2^{(u)}(X,Z)$, $p_2^{(s)}(X,Z)$ be the joint distributions produced by f_2. Define:

$$p_\lambda^{(u)}(X,Z) = \lambda p_1^{(u)}(X,Z) + (1-\lambda)p_2^{(u)}(X,Z), \quad (9)$$

$$p_\lambda^{(s)}(X,Z) = \lambda p_1^{(s)}(X,Z) + (1-\lambda)p_2^{(s)}(X,Z). \quad (10)$$

By definition, the left hand side of (4) can be rewritten by:

$$\begin{aligned}
& \lambda T(\Delta_1) + (1-\lambda)T(\Delta_2) \\
= \;& \lambda d(p_1^{(u)}(Y,Z)\,||\,p_1^{(s)}(Y,Z)) \\
+ \;& (1-\lambda)d(p_2^{(u)}(Y,Z)\,||\,p_2^{(s)}(Y,Z)) \\
= \;& \lambda d(p^{(u)}(Y|X)p_1^{(u)}(X,Z)||p^{(s)}(Y|X)p_1^{(s)}(X,Z)) \quad (11) \\
+ \;& (1-\lambda)d(p^{(u)}(Y|X)p_2^{(u)}(X,Z)||p^{(s)}(Y|X)p_2^{(s)}(X,Z)) \quad (12) \\
\geq \;& d(p^{(u)}(Y|X)p_\lambda^{(u)}(X,Z)||p^{(s)}(Y|X)p_\lambda^{(s)}(X,Z)) \quad (13)
\end{aligned}$$

where (11) and (12) are due to (7) and (8); (13) is due to (9), (10), and the convexity of $d(\cdot||\cdot)$.

Let f_λ be the corresponding function that induces the joint distribution $p_\lambda^{(u)}(X,Z)$ and $p_\lambda^{(s)}(X,Z)$. Define:

$$\Delta_\lambda = \int_{\boldsymbol{x}\in\mathcal{X}}\int_{\boldsymbol{z}\in\mathcal{Z}} p_\lambda^{(s)}(\boldsymbol{x},\boldsymbol{z})\ell(g(\boldsymbol{z}),y^{(s)}(\boldsymbol{x}))\,d\boldsymbol{x}d\boldsymbol{z}. \quad (14)$$

By definition of $T(\Delta)$ in Definition 1, we have:

$$d(p^{(u)}(Y|X)\,p_\lambda^{(u)}(X,Z)||p^{(s)}(Y|X)\,p_\lambda^{(s)}(X,Z)) \geq T(\Delta_\lambda). \quad (15)$$

Combine (13) and (15):

$$\lambda T(\Delta_1) + (1-\lambda)T(\Delta_2) \geq T(\Delta_\lambda). \quad (16)$$

That said, the left-hand side of (4) is greater or equal to $T(\Delta_\lambda)$. Next, we show that:

$$T(\Delta_\lambda) \geq T(\lambda\Delta_1 + (1-\lambda)\Delta_2). \quad (17)$$

Since $T(\Delta)$ is non-increasing, (17) is equivalent to:

$$\Delta_\lambda \leq \lambda\Delta_1 + (1-\lambda)\Delta_2. \quad (18)$$

Indeed, we have:

$$\begin{aligned}
\Delta_\lambda = \;& \int_{\boldsymbol{x}}\int_{\boldsymbol{z}} p_\lambda^{(s)}(\boldsymbol{x},\boldsymbol{z})\ell(g(\boldsymbol{z}),y^{(s)}(\boldsymbol{x}))d\boldsymbol{x}d\boldsymbol{z} \quad (19) \\
= \;& \lambda\int_{\boldsymbol{x}}\int_{\boldsymbol{z}} p_1^{(u)}(\boldsymbol{x},\boldsymbol{z})\ell(g(\boldsymbol{z}),y^{(s)}(\boldsymbol{x}))d\boldsymbol{x}d\boldsymbol{z} \quad (20) \\
+ \;& (1-\lambda)\int_{\boldsymbol{x}}\int_{\boldsymbol{z}} p_2^{(u)}(\boldsymbol{x},\boldsymbol{z})\ell(g(\boldsymbol{z}),y^{(s)}(\boldsymbol{x}))d\boldsymbol{x}d\boldsymbol{z} \quad (21) \\
\leq \;& \lambda\Delta_1 + (1-\lambda)\Delta_2 \quad (22)
\end{aligned}$$

with (19) due to (14), (20) and (21) due to (9), (22) due to (5) and (6), respectively. From (18) and (22), (17) follows. Finally, from (16) and (17), (4) follows. The proof is complete. □

It is worth noting that the convexity of $d(\cdot||\cdot)$ is not a restricted condition, indeed, most of the divergence functions, for example, the Kullback-Leibler (KL) divergence is convex.

Theorem 1 shows that only enforcing a small distribution discrepancy between domains will increase the classification risk and vice-versa.

4. A NEW VALIDATION METHOD

Based on Theorem 1, we argue that to select a good model for unseen domains, one must account for both the classification risk and the domain discrepancy not only in the training process but also in the validation process. Note that state-of-the-art model evaluation methods for DG are mainly based on the classification risk or, equivalently, the classification accuracy [7] [18] on the validation set to select the models. Given this fact, we propose to select a model that minimizes the following objective function on the validation set:

$$L_{\text{Validation loss}} = \beta(1-\alpha)L_{\text{Classification risk}} + \alpha L_{\text{Domain-discrepancy loss}} \quad (23)$$

where α is the convex combination hyper-parameter and β is the scale hyper-parameter that supports the combination of objectives with different scales.

It is pretty clear that the cross-entropy loss is a good representation of classification risk. However, it is hard to choose the measure for quantifying the domain-discrepancy loss. Indeed, there exist various definitions of domain discrepancy. Several works characterize the domain-discrepancy via the difference in the marginal distributions [5,6], other works measure it by the mismatch in conditional distributions [1]. We believe that finding a good measure for domain discrepancy is still an open problem. Therefore, in this short paper, we decide to use the widely accepted Maximum Mean Discrepancy (MMD) loss [6] in the feature space to quantify the domain discrepancy. We also acknowledge that though MMD measure is extensively used, it may not be the optimal choice.

In practice, we found that MMD loss is at the same scale as the cross-entropy loss when the training process is stable, we thus choose β as 1. For α, we consider the classification performance as the more important goal and thus heuristically

Table 1. Classification accuracy of 12 tested algorithms on PACS, VLCS, and C-MNIST datasets using the Training-domain validation method (Traditional) proposed in [18] *vs.* using our new validation method.

Algorithm	**Fish** [19]	**IRM** [1]	**GDRO** [12]	**Mixup** [20]	**CORAL** [5]	**MMD** [6]	**DANN** [21]	**CDANN** [22]	**MTL** [23]	**VREx** [24]	**RSC** [25]	**SagNet** [26]	**Wins**
PACS (Traditional)	**84.6**	84.9	84.2	83.3	**85.1**	83.6	84.6	**86.4**	83.0	**84.5**	**85.2**	83.7	
PACS (Ours)	82.0	**85.3**	**84.3**	**85.3**	84.9	**85.0**	**84.9**	82.0	**84.2**	84.2	81.3	**85.1**	7/12
VLCS (Traditional)	**79.4**	76.0	78.1	77.4	76.8	**78.5**	77.8	79.2	77.3	76.4	**78.6**	**80.5**	
VLCS (Ours)	77.5	**79.2**	**79.6**	**77.6**	**78.8**	78.0	**78.5**	**80.3**	**78.2**	**78.6**	76.1	79.3	8/12
CMNIST (Traditional)	**10.0**	10.0	10.2	**10.4**	9.7	**10.4**	10.0	9.9	10.5	10.2	10.2	10.4	
CMNIST (Ours)	9.7	**10.9**	**12.6**	10.3	**11.2**	9.9	**11.1**	**10.2**	**11.5**	**15.6**	**13.8**	**10.5**	9/12

choose α as 0.2. From our experiments, we found that the performance of our validation method is robust to small values of α within the range of $[0.1, 0.3]$. One more insight from Theorem 1 is that it is advisable to avoid extreme points in Δ (classification error) to maintain a balance between the model's generalization and prediction capabilities. This means the classification error should not be too small or too large. Thus, for each hyper-parameter configuration, we sort the validation cross-entropy loss in ascending order and only pick the models that produce 5% to 50% percentile of the validation cross-entropy loss as a subset of candidates for model selection. Our implementation is released at this link[3].

5. NUMERICAL RESULTS

We compare the proposed model selection method with the Training-domain validation method described in [18] on three datasets: PACS, VLCS, and Colored-MNIST (C-MNIST) using DomainBed package and 12 different DG algorithms provided there [18]. Recall that the Training-domain validation method chooses the model that produces the highest validation accuracy, while our method selects the model that minimizes the objective function in (23). For PACS and VLCS datasets, we report the average test accuracy over 4 different tasks with each time leaving one domain out as the unseen domain. For the C-MNIST dataset, we only focus on the most difficult domain, where the correlation between the label and the color of the unseen domain is completely different from the seen domains and no algorithm can achieve more than 10.5% points accuracy [18].

The validation set is formed using 20% data from each seen domain, denoted as the training-domain validation set in [18]. We follow exactly the same settings and training routine used in DomainBed and conduct 20 trials of random search over a joint distribution of hyper-parameters for each task per algorithm. For the MMD loss implementation, we directly use the code provided in DomainBed package. We train each model for 5000 steps. The validation cross-entropy loss, MMD loss, and validation accuracy are recorded every 100 steps for VLCS dataset and every 300 steps for PACS and C-MNIST datasets.

With $\alpha = 0.2, \beta = 1$, the performance of each algorithm under different validation methods on PACS ,VLCS and Colored-MNIST datasets is shown in Table 1. We refer to the Training-domain validation method as "Traditional" and the proposed method as "Ours". For the PACS dataset, the proposed validation method can select slightly better models for seven out of twelve DG algorithms. For the remaining five DG algorithms, our method achieves comparable performance with the "Traditional" method on CORAL [5] and VREx [24]. However, for Fish [19], CDANN [22] and RSC [25], we observe a performance deterioration. The effectiveness of the proposed method can be more easily observed on VLCS dataset, where eight out of twelve DG algorithms get an improved model selected, with the improvement varies from 0.2% to 3.2%. For the C-MNIST dataset, the proposed validation method consistently selects models with better performance compared with the "Traditional" validation method. Accuracy improves for nine out of twelve tested algorithms with the most significant improvement for VREx [24] method by 5.4%.

6. CONCLUSION

By showing the trade-off between minimizing the classification risk and domain discrepancy, we demonstrate that the traditional model selection methods may not be suitable for DG problem and propose a new model selection method that considers both objectives. While our approach outperforms traditional methods on several DG algorithms and datasets, it lacks an automatic hyper-parameter tuning strategy. Note that the domain discrepancies may vary across different datasets, one may not expect the same optimal values of α and β for all datasets. Determining the "optimal" ones could be a hard problem both practically and theoretically. We thus leave it as an open problem for future work. Despite this limitation, we believe our approach provides insight and initial results for exploring new model selection methods specific for DG problem.

[3] https://github.com/thuan2412/A-principled-approach-for-model-validation-for-domain-generalization

7. REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz, "Invariant risk minimization," *stat*, vol. 1050, pp. 27, 2020.

[2] Boyang Lyu, Thuan Nguyen, Prakash Ishwar, Matthias Scheutz, and Shuchin Aeron, "Barycentric-alignment and reconstruction loss minimization for domain generalization," *arXiv preprint arXiv:2109.01902*, 2021.

[3] Thuan Nguyen, Boyang Lyu, Prakash Ishwar, Matthias Scheutz, and Shuchin Aeron, "Conditional entropy minimization principle for learning domain invariant representation features," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 3000–3006.

[4] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao, "Invariant information bottleneck for domain generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 7399–7407.

[5] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[6] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[7] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.

[8] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, 2006.

[9] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon, "On learning invariant representations for domain adaptation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7523–7532.

[10] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit, "On calibration and out-of-domain generalization," *Advances in neural information processing systems*, vol. 34, pp. 2215–2227, 2021.

[11] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang, "Towards a theoretical framework of out-of-distribution generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23519–23531, 2021.

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

[13] 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.

[14] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong, "Ensemble of averages: Improving model selection and boosting performance in domain generalization," *Advances in Neural Information Processing Systems*, 2021.

[15] Thuan Nguyen, Boyang Lyu, Prakash Ishwar, Matthias Scheutz, and Shuchin Aeron, "Joint covariate-alignment and concept-alignment: a framework for domain generalization," in *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2022, pp. 1–6.

[16] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál, "Impossibility theorems for domain adaptation," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 129–136.

[17] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.

[18] Ishaan Gulrajani and David Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2020.

[19] 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.

[20] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6502–6509.

[21] 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*, vol. 17, no. 1, pp. 2096–2030, 2016.

[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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 624–639.

[23] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott, "Domain generalization by marginal transfer learning," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 46–100, 2021.

[24] 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 *International Conference on Machine Learning*. PMLR, 2021, pp. 5815–5826.

[25] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang, "Self-challenging improves cross-domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 124–140.

[26] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo, "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8690–8699.