A REMEDY FOR DISTRIBUTIONAL SHIFTS THROUGH EXPECTED DOMAIN TRANSLATION

Jean-Christophe Gagnon-Audet[1, 2], *Soroosh Shahtalebi*[3], *Frank Rudzicz*[3, 4], *Irina Rish*[1, 2]

[1]Mila - Québec AI Institute, Montréal, QC, Canada
[2]Département d'informatique et de recherche opérationnelle, University of Montréal, QC, Canada
[3]Vector Institute for Artificial intelligence, Toronto, ON, Canada
[4] Department of Computer Science, University of Toronto, ON, Canada

ABSTRACT

Machine learning models often fail to generalize to unseen domains due to the distributional shifts. A family of such shifts, "correlation shifts," is caused by spurious correlations in the data. It is studied under the overarching topic of "domain generalization." In this work, we employ multi-modal translation networks to tackle the correlation shifts that appear when data is sampled out-of-distribution. Learning a generative model from training domains enables us to translate each training sample under the special characteristics of other possible domains. We show that by training a predictor solely on the generated samples, the spurious correlations in training domains average out, and the invariant features corresponding to true correlations emerge. Our proposed technique, Expected Domain Translation (EDT), is benchmarked on the Colored MNIST dataset and drastically improves the state-of-the-art classification accuracy by 38% with train-domain validation model selection.

***Index Terms*—** machine learning, domain generalization, out-of-distribution generalization, correlation shift

1. INTRODUCTION

The performance of current machine learning models, although significant in some applications [1, 2, 3], is limited to the implicit assumption that testing and training data are independently and identically distributed (IID). However, this assumption does not hold in many practical applications, which may lead to drastic drops in the accuracy and reliability of their decisions. For example, machine learning models that are trained to classify males vs. females often identify hair color as a strong discriminating feature (e.g., on the CelebA dataset [4]).

More formally, machine learning aim at capturing the joint distribution $P(X = x, Y = y)$, where $x \in \mathcal{X}$ is the input sample and $y \in \mathcal{Y}$ is the class label associated with it. Conventionally, it is assumed that $P_{train}(X, Y) = P_{test}(X, Y)$, while, in practice, this may not be the case. This problem has ignited a surge of interest [5, 6, 7, 8, 9, 10, 11] in developing models that are robust to distribution shifts and has led to a growing body of literature on domain or out-of-distribution (OOD) generalization.

Generally, distribution shifts can be categorized into three groups: (i) **covariate shift** where $P_{train}(X) \neq P_{test}(X)$; (ii) **correlation shift** where $P_{train}(Y \mid X) \neq P_{test}(Y \mid X)$, and (iii) **label shift** where $P_{train}(Y) \neq P_{test}(Y)$. To handle distribution shifts, a fundamental assumption when dealing with these shifts is that a subset of input features are invariant across all training and testing data. These *invariant features* are needed for a model to generalize under distributional shifts.

To achieve *invariant prediction* under distributional shifts, Arjovsky et al. [5] projected data from different domains into a latent space, where the *agreement* between the representation of each class across different domains is maximized. Others [6, 9] aimed for the same objective as before by maximizing the agreement of loss landscapes across different domains. Others [12, 8, 13] approached invariant prediction by minimizing some function of the risk of a model for different domains. Vapnik [13] minimize the average risk across domains, while Krueger et al. [8] minimize an affine combination of risk across domains and Sagawa et al. [12] minimize the maximum risk across domains. Invariant prediction is also looked at through the lens of causality, meaning that causal relationships are in fact, the type of features that remain invariant across different domains [14].

Recently, Robey et al. [15] proposed a new method to tackle covariate shifts in datasets referred to as "model-based domain generalization (MBDG)", which models invariant prediction as a constraint optimization problem and then translates this into an unconstrained problem through non-convex duality theory. MBDG employs multi-modal image-to-image translation networks to learn the domains' manifold and translator recreate samples under the characteristics of other possible domains. Then, a classifier (neural network) is trained to classify the generated samples while minimizing

Irina Rish acknowledges the support from Canada CIFAR AI Chair Program and from the Canada Excellence Research Chairs program.

the distance between the logits of the same sample across different domains.

MBDG yields state-of-the-art results on various domain generalization datasets (e.g., PACS and VLCS). Although its theoretical grounding is limited to datasets where the shift is purely covariate, MBDG also gets state-of-the-art results on Colored MNIST, mainly driven by correlation shift. Here, we propose a new theoretical grounding for a model-based approach distinctly equipped to deal with correlation shift.

As thoroughly investigated in the literature [5, 16, 12], correlation shifts happen due to the features that are spuriously correlated with each domain. These features emerge due to selection bias or anti-causal correlations in the data. In this work, which we refer to as Expected Domain Translation (EDT), we take advantage of multi-modal domain translation architectures to regenerate each data point under the characteristics of other domains. Then, we design the learning objective to minimize the risk across the generated samples. Our contributions are:

- EDT significantly improves state-of-the-art accuracy on the CMNIST dataset by 38.6% by achieving 48.8% accuracy on training-domain validation, and improves test-domain validation by 10% with 74.6%.
- EDT matches state-of-the-art classification accuracy on covariate shift problems, meaning that the proposed remedy does not affect the accuracy on these tasks.
- EDT matches state-of-the-art classification accuracy on covariate shift problems.

2. PROBLEM FORMULATION

Here, we provide a formal definition for "correlation shift" and lay the foundations for our definition of "invariant prediction". As mentioned in Section 1, the conventional supervised learning tasks are characterized by capturing the joint probability distribution $P(X, Y)$, such that a predictor $\Phi : \mathcal{X} \to \mathcal{Y}$ can correctly assign each data point X to its label Y. In OOD settings, however, it is assumed that the true joint distribution $P(X, Y)$ cannot be directly sampled, and we can only sample from its realizations in observed domains $P(X^e, Y^e)$, where e belongs to the set of observed domains $\mathcal{E}_{\text{obs}}$. To integrate the notion of the domain in our calculations, we denote the samples obtained from each domain $e \in \mathcal{E}_{train}$ by (x^e, y^e). By sampling IID from each domain and forming datasets $D^e := (x^e_j, y^e_j)_{j=1}^{n_e}$, the invariant prediction problem can be formulated as Problem 1.

Problem 1 *Find an optimal predictor Φ^* such that*

$$\Phi^\star \triangleq \underset{\Phi}{\operatorname{argmin}} \max_{e \in \mathcal{E}_{all}} E_{P(X^e, Y^e)}(\ell(\Phi(X^e), Y^e)), \qquad (1)$$

where $\mathcal{E}_{all} = supp(\mathcal{E})$ is the support of the domain variable $\mathcal{E}$ or the space of all possible domains. Also, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ denotes the loss function.

The invariant predictor Φ^* is said to be *OOD-optimal* as it has learned features that generalize to any domain in $\mathcal{E}_{\text{all}}$. Learning features that are invariant across $\mathcal{E}_{\text{all}}$ is an ultimately challenging task because the set of observed domains is only a finite subset of all the domains that can practically exist, $\mathcal{E}_{\text{obs}} \subseteq \mathcal{E}_{\text{all}}$. In the next section, we formalize our technique to tackle this challenge.

3. PROPOSED METHODOLOGY

Our approach towards the stated problem is grounded on training and employing domain translation networks that approximate the mapping from each domain to another and enable us to translate samples of an observed domain in $\mathcal{E}_{\text{obs}}$ under the characteristics of any other domain in $\mathcal{E}_{\text{all}}$. Together with our training strategy, this technique enables the network to render the spurious correlations of each domain as noisy behaviors that will vanish during the training process.

3.1. Formal assumptions

Although shifts in data distribution come in many forms, we focus on data where both $P(X)$ and $P(Y|X)$ shift between domains and $P(Y)$ is constant. However, we assume that there is an invariant latent variable Z_{inv}. In other words, the realization of an invariant predictor is grounded based on the assumption that there exists a latent variable Z_{inv} that is invariant across $\mathcal{E}_{all}$.

Assumption 1 *We assume there exists an invariant latent variable Z_{inv} which has a joint probability distribution with the label Y that is invariant between domains such that*

$$P(Z_{inv}, Y) = P(Z^e_{inv}, Y) \; \forall \; e \in \mathcal{E}_{all}. \qquad (2)$$

Z_{inv} is the invariant feature that we want the model to learn, and it is reasonable to assume that such feature exists in the data. Without it, there is nothing invariant to learn in the data, and an invariant predictor is impossible.

Assumption 2 *We assume that taking the expected value of the spurious correlation distribution given the label over all possible domains makes spurious correlations uniform over all possible values of $Z_{spu} \in \mathcal{Z}_{spu}$ such that*

$$E_{e' \sim P(\mathcal{E}_{all})}(P(Z^{e'}_{spu}|Y)) = P(Z_{spu}|Y) \overset{d}{=} U(\mathcal{Z}_{spu}). \qquad (3)$$

By taking the expected value of spurious features given the label over all possible domains, we obtain a distribution where spurious correlations are no longer dependent on the domain, thus invariant. Also, spurious correlations become uninformative of the label as they are now uniformly distributed.

Our end goal is to bring the distribution of observed domains, $P(X^e, Y^e)$, to a domain invariant data distribution $P(X_{\text{inv}}, Y)$ during training. Then, the OOD problem narrows down to an in-distribution generalization problem where the goal is to maximize the classification accuracy in a supervised learning setting. For this purpose, we assume a measurable domain translation function T can shift a data point from one domain definition to another.

Assumption 3 *We assume a measurable domain translation function $T : \mathcal{X} \times \mathcal{E}_{all} \to \mathcal{X}$ that injects inter-domain correlation shifts to data samples X^e such that*

$$T(X^e,\, e') = X^{e'} \;\forall\, e, e' \in \mathcal{E}_{all}. \tag{4}$$

In practice, this measurable function is feasible for correlation shift problems with no diversity shift [16] because all spurious features are observed from the data. For example, in CMNIST, all colors and digits are observed in all domains but differently correlated to the label.

Proposition 1 *Under Assumptions 1, 2, and 3, taking the expected translated domain distribution $P(T(X^e, e'),\, Y^e)$ over $e' \in \mathcal{E}_{all}$ averages out the spurious correlations. Thus, a domain-invariant distribution of the data $P(X_{inv}, Y)$ emerges.*

$$P(X_{inv},\, Y) \overset{d}{=} E_{e' \sim P(\mathcal{E}_{all})}(P(T(X^e, e'),\, Y)) \;\forall\, e \in \mathcal{E}_{all} \tag{5}$$

Proof. We assume that X^e is a vector with two features (Z^e_{inv}, Z^e_{spu}): the invariant and spurious latent feature, respectively. Under Assumption 3, the RHS becomes

$$E_{e' \sim P(\mathcal{E}_{all})}(P(T(X^e, e'), Y)) \tag{6}$$

$$= E_{e' \sim P(\mathcal{E}_{all})}(P(Z^{e'}_{inv},\, Z^{e'}_{spu},\, Y)) \tag{7}$$

Under the causal graph $Z^e_{inv} \to Y^e \to Z^e_{spu}$ where the label causes the spurious correlation in the data, we can decompose the joint probability and use Assumption 1

$$E_{e' \sim P(\mathcal{E}_{all})}(P(Z^{e'}_{inv}, Y^e) P(Z^{e'}_{spu} | Y)) \tag{8}$$

$$= P(Z_{inv}, Y) E_{e' \sim P(\mathcal{E}_{all})}(P(Z^{e'}_{spu} | Y)) \tag{9}$$

Under Assumption 2, we see that the spurious feature is no longer a function of the domain and is therefore invariant. With this, the X becomes entirely independent of the domain since the other feature Z_{inv} is already invariant.

$$P(Z_{inv}, Y) P(Z_{spu} | Y) = P(X_{inv}, Y) \tag{10}$$

which directly implies the initial statement. ■

Proposition 2 *Under Proposition 1, the expected randomly translated loss over all domains $\mathcal{E}_{all}$ is equal to the expected loss from an invariant distribution of data.*

$$E_{(x,y) \sim P(X_{inv}, Y)}(\ell(\Phi(x), y)) \tag{11}$$

$$= E_{e' \sim P(\mathcal{E}_{all})}(E_{(x,y) \sim P(T(X^e, e'), Y^e)}(\ell(\Phi(x), y))) \tag{12}$$

3.2. Domain Translation Models

With the purpose of using Proposition 2 in order to learn from an invariant distribution, we must find a function T which fits our assumptions. We turn our attention to data-driven domain translation models, which are employed to learn the mapping and translate the samples of each domain under the characteristics of another domain. We employ a multimodal image to image translation model (MIITN) [17] that is trained on $\mathcal{E}_{\text{obs}}$.

Algorithm 1 Expected Domain Translation (EDT)

Hyperparameters: Step size $\eta > 0$
repeat
  for minibatch $\{(x_j, y_j)\}_{j=1}^m$ in training dataset **do**
    $\tilde{x}_j \leftarrow \textsc{TranslateImage}(x_j)\ \forall j \in [m]$
    $\text{loss}(\theta) \leftarrow (1/m) \sum_{j=1}^m [\ell(\tilde{x}_j, y_j; \varphi(\theta, \cdot))]$
    $\theta \leftarrow \theta - \eta \nabla_\theta \text{loss}(\theta)$
  end for
until convergence
procedure GENERATEIMAGE(x^e)
  $(x, e) \leftarrow H(x^e)$ ▷ Decompose x^e into x and e
  Sample $e' \sim \mathcal{E}_{\text{all}}$
  return $G(x, e')$
end procedure

Particularly, we follow the MUNIT [18] architecture to parameterize the employed MIITN. The MUNIT model consists of two modules; a disentangling model $H : \mathcal{X} \to \mathcal{X} \times \mathcal{E}_{\text{all}}$, and a generative model $G : \mathcal{X} \times \mathcal{E}_{\text{all}} \to \mathcal{X}$. The disentangling module recovers the true sample x from the observed samples for any domain e, i.e. $H(x^e) = (x, e)$, and the generative module translates the recovered true sample x to any domain e', i.e. $G(x, e') = x^{e'}$. This pair enable us to generate the realizations of any sampled data x^e in different domains as $T(x^e, e') = G(H(x^e), e') = x^{e'}$ for all e, e' in $\mathcal{E}_{\text{all}}$.

3.3. Expected Domain Translation

Most recent works in OOD generalization focus on penalizing something in the objective function that promotes invariance in the model [5, 19, 8, 15]. Instead of this approach, we rely on Proposition 2 to learn an invariant predictor. We argue that no spurious correlation can render robust predictors by only looking at randomly translated samples, leading to domain-specific optimal performance. In other words, we implicitly promote invariant learning by hiding observed domains $\mathcal{E}_{\text{obs}}$ from the optimization process and only train on samples that have been translated to domains sampled randomly in $\mathcal{E}_{\text{all}}$. This is achieved by learning the domain manifold of training data through MIITNs [17] and using this manifold to randomize the domain definition of all the training data. From this theoretical background, we formulate the optimization problem in Proposition 3, and its algorithmic procedure is provided in Algorithm 1.

Proposition 3 *From Proposition 2, we formulate the optimization problem to minimize the expected value of the expected loss w.r.t. random domain translations.*

$$\theta^\star \triangleq \underset{\theta \in \Theta}{\text{argmin}}\, E_{(X,Y) \sim \cup_{e \in \mathcal{E}_{obs}} D^e}(E_{e' \sim \mathcal{E}_{all}}(\ell(f(T(X, e')), Y; \theta)))$$

4. EXPERIMENTAL RESULTS

This section encapsulates the experimental results of EDT over two major benchmarking datasets in the field of domain generalization. All of our experiments are implemented in the

Table 1: Test Accuracy of CMNIST Dataset

Algorithm	Train-domain Val.	Test-domain Val.
ERM	10.0 ± 0.1	28.7 ± 0.5
IRM	10.2 ± 0.3	58.5 ± 3.3
VREx	10.2 ± 0.0	55.2 ± 4.0
MBDG	10.2 ± 0.1	64.6 ± 0.7
EDT	$\mathbf{48.8 \pm 3.1}$	$\mathbf{74.6 \pm 1.6}$

Domainbed framework [20], which imposes rigorous hyperparameter search and model selection conditions on the domain generalization techniques and allows for a fair comparison among them. For model selection, Domainbed employs two strategies to form the validation sets; either the validation set is formed based on training domains ($\mathcal{E}_{\text{obs}}$) or it is based on data from the last training step of the test domains ($\mathcal{E}_{\text{all}} \setminus \mathcal{E}_{\text{obs}}$).

4.1. Colored MNIST Dataset

CMNIST is a variant of the MNIST handwritten digit classification dataset. Domains $d \in \{80\%, 90\%, 10\%\}$ contain images of digits colored either red or green. The label is a noisy function of the digit and color, such that the color bears a correlation of d with the label, and the depicted digit always bears a correlation of 75% with the label. We train on $\mathcal{E}_{\text{obs}} = \{80\%, 90\%\}$ and test on $\mathcal{E}_{\text{all}} = \{10\%\}$. The color is spuriously correlated with the label, while the depicted digit is invariantly correlated with the label. Minimizing the empirical risk on the training set composed of $\mathcal{E}_{\text{obs}}$ will lead to a predictor relying on the color of the digits as it is the stronger predictor of the label than the shape of the digits.

The ultimate goal is to find an invariant predictor that only considers the shape of the digits to make its prediction, thus only getting 75% classification accuracy across all domains. Table 1 shows that EDT outperforms its counterparts significantly and offers the state-of-the-art close-to-optimal classification accuracy on CMNIST.

Table 2: Test Accuracy over Camelyon17 Dataset

Algorithm	Train-domain Val.	Test-domain Val.
ERM	79.2 ± 4.3	76.2 ± 1.3
MBDG	89.5 ± 1.0	$\mathbf{92.1 \pm 0.1}$
EDT	$\mathbf{89.8 \pm 0.1}$	90.6 ± 0.7

4.2. Camelyon17

This dataset is part of the WILDS domain generalization benchmarking datasets [21] and consists of 400k images of potentially cancerous lymph node sections from different hospitals. The task is to predict whether each section contains a tumor or not. The challenge is that each hospital employs a different coloring agent; thus, models have difficulty generalizing to unseen coloring from new hospitals. This is an instance of a pure covariate shift in the real world. As shown in Table 2, although EDT is not tailored for covariate shift problems, it matches the state-of-the-art accuracy in both validation scenarios. Please note that here we follow the testing conditions as originally published in [21].

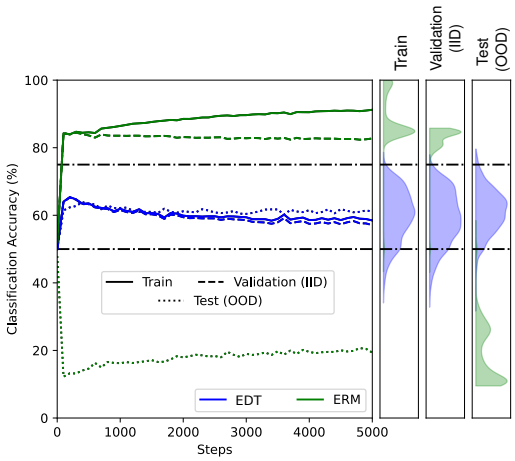

Fig. 1: Classification accuracy of EDT on the CMNIST dataset compared to ERM. The "Train" partition is the model's data for training where the color is 90% and 80% correlated with the label. The "Validation (IID)" partition is data unseen by the model but sampled IID from the training domains. The "Test (OOD)" partition is the test domain where the color is only 10% correlated with the label. Each curve is the average performance across 20 random hyperparameter configurations and three trial seeds. The plots on the right are violin plots of all accuracy points for all three partitions.

4.3. Is the objective in Problem 1 satisfied?

Here, we explore the behavior of different algorithms in solving Problem 1. The goal is to minimize the maximum risk across all possible domains $\mathcal{E}_{\text{all}}$ – a solution should yield an invariant predictor that achieves similar accuracy over the data coming from any set of domains. To investigate if standard OOD algorithms can find an invariant predictor, we look at the accuracy during training across different splits (Train, IID or OOD) of the CMNIST dataset in Fig. 1. In CMNIST, any predictor above 75% accuracy is leveraging spurious correlations. We see that this is the case for ERM on the Train and IID splits, but not for EDT. Also, the risk across domains is not constant for ERM, while EDT manages to achieve a constant risk across the seen and unseen domains. This behavior indicates that EDT has not developed over-confident decision rules based on spurious correlations, even though these correlations are typically the easiest ones to pick up. These observations point to an invariant predictor learned by EDT.

5. CONCLUSION

In this work, we proposed EDT, which offers a remedy to tackle correlation shifts in distributional shifts. EDT takes advantage of domain translation models to augment different realizations of one dataset under different environmental conditions. EDT is evaluated on CMNIST dataset and offers state-of-the-art classification accuracy.

6. REFERENCES

[1] Adrien Bardes, Jean Ponce, and Yann LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[2] Soroosh Shahtalebi, Seyed Farokh Atashzar, Olivia Samotus, Rajni V Patel, Mandar S Jog, and Arash Mohammadi, "Phtnet: Characterization and deep mining of involuntary pathological hand tremor using recurrent neural network models," *Scientific reports*, vol. 10, no. 1, pp. 1–19, 2020.

[3] Soroosh Shahtalebi, S Farokh Atashzar, Rajni V Patel, Mandar S Jog, and Arash Mohammadi, "A deep explainable artificial intelligent framework for neurological disorders discrimination," *Scientific reports*, vol. 11, no. 1, pp. 1–18, 2021.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

[6] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Scholkopf, "Learning explanations that are hard to vary," *arXiv preprint arXiv:2009.00329*, 2020.

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

[8] 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)," *arXiv preprint arXiv:2003.00688*, 2020.

[9] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish, "Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization," *arXiv preprint arXiv:2106.02266*, 2021.

[10] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney, "Empirical or invariant risk minimization? a sample complexity perspective," *arXiv preprint arXiv:2010.16412*, 2020.

[11] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar, "Invariant risk minimization games," in *International Conference on Machine Learning*. PMLR, 2020, pp. 145–155.

[12] 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.

[13] Vladimir N Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[14] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

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

[16] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li, "Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms," *arXiv preprint arXiv:2106.03721*, 2021.

[17] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner, "Towards a definition of disentangled representations," *arXiv preprint arXiv:1812.02230*, 2018.

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

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

[20] Ishaan Gulrajani and David Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[21] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al., "Wilds: A benchmark of in-the-wild distribution shifts," *arXiv preprint arXiv:2012.07421*, 2020.