DeconICA: Deconfounding the Dataset Bias for Domain Generalization

Anonymous ACL submission

Abstract

Domain generalization provides a research spot for enhancing the generalization capability of the machine learning model. We focus on a causal perspective for the domain generalization task. In causal theory, a confounder is a factor that affects both the cause and the effect. The confounder is often hidden, which causes problems in correctly performing the intervention. The Deconfounder approach indicates that a factorized multiple causes could be considered a substitute confounder. We choose a non-linear ICA method to factorize the data features to represent the confounder. The confounder is considered to represent the background, and domain biases. Empirical results on text and image classification domain generalization validate the proposed methods.

1 Introduction

Deep neural networks have achieved significant success in various application domains, ranging from image recognition Szegedy et al. (2015); Simonyan and Zisserman (2015) to text embedding Devlin et al. (2019), to games Silver et al. (2016), etc. We consider the problem of the domain generalization (DG) in text and image classification Szegedy et al. (2015); Simonyan and Zisserman (2015); He et al. (2016); Dosovitskiy et al. (2021), due to its great significance.

The DG setting in this paper is when the distribution of the target domain is unknown. The challenge is twofold. First, the built model should have a good generalization capability on an unknown target domain, which is also the ultimate goal of the DG task. Meanwhile, the model should still perform well on the source domains. The State-of-the-art mainly aims to minimize the risk on the source domains via aligning their distributions Wang et al. (2021); Li et al. (2018b). This strategy, however, tends to overfit the model in the source domains, as the last layers of the deep learning models capture the specifics of the source data but fail to generalize well on the target domain. (more in section 2).

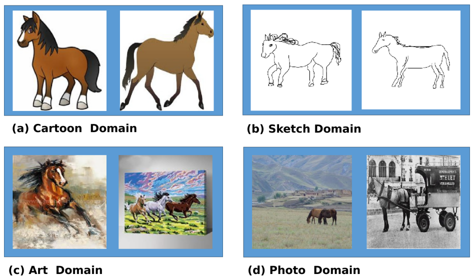

Figure 1: An illustration of the difference between the domain bias and the background difference: (1) Between (b), (c) and (d), the bias includes both domain bias and background bias. (2) Between (b) and (a), the dataset difference equals to domain bias.

The domain generalization strategy proposed in this paper focuses on handling the domain biases resulting from the different specifics of the source datasets, referred to as dataset biases Yang et al. (2020). The background biases are common problems in generic object recognition tasks. While domain biases in DG are more noticeable, many proposed methods do not tackle the issue of the **inherent background bias within the same dataset**. As shown in Figure 1, it is clear that: (1) Between (b), (c) and (d), the bias includes both domain bias and background bias. (2) Between (b) and (a), the bias equals to domain bias. There is **not a simple relationship** between **domain bias and background bias**, they should be considered separately.

Hence, we propose to model both the background and domain biases. From the perspective of causal inference, Pearl and Mackenzie (2018); Peters et al. (2017), the input images are viewed as the cause of the learned model, and its semantic recognition performance is considered as the potential outcomes (effect) Schölkopf et al. (2021); Yang et al. (2020). The dataset biases (the domain biases and background differences) are viewed as hidden confounders, affecting both the causes and outcomes. For instance, a man running on the beach may be incorrectly recognized as swimming, as the seaside background makes a spurious link between the image and the term “swimming”.

On the one hand, from the causal inference perspective, the average effect of the intervention is hard to estimate, given the hardly defined domain differences and the highly complex background of the images in the datasets. On the other hand, the components of a generative model, viewed as substitute hidden confounders (SHCs), are used to block the backdoor effect from the hidden confounders, in *Deconfounder* Wang and Blei (2019). *Deconfounder* relies on a factorized generative model of the data and is receiving increasing attention recently (D'Amour (2019); Gan et al. (2021)). However, the factor model is not always identifiable D'Amour (2019), hindering the validity of the results obtained (more in section 3). The non-identifiability issue is well-explained in non-linear ICA approaches Hyvarinen and Morioka (2016); Xiu et al. (2021).

Taking inspiration from both *Deconfounder* and the adversarial non-linear ICA factor model Brakel and Bengio (2017), the proposed method, namely, DeconICA scheme, aims to solve the domain biases confounding effect, by extracting substitute hidden confounders and estimating their average effect, with a novel fusion method based on the attention mechanism. The fusion method can be viewed from two perspectives: first, it can be viewed as a more flexible feature fusion mechanism to estimate the average effect; second, it can be seen as an intervention, i.e., the final representation would select the useful features from SHCs to prevent the true confounder from affecting the real causal link.

The contributions of this paper are threefold: 1) The DG task is formulated from a causal inference perspective, considering the background and domain biases as confounders. 2) A novel neural scheme inspired by the *Deconfounder*, and mitigating its unidentifiability issue is proposed. 3) The empirical results on various datasets validate the effectiveness of the proposed scheme.

2 Related Work

2.1 Domain Generalization

Similar to Domain Adaptation Ben-David et al. (2007, 2010), Domain Generalization Huang et al. (2006); Pan et al. (2010); Zhang et al. (2015); Ghifary et al. (2016) aims to transfer learning, and specifically porting models learned from so-called source domain(s) to a target domain. In the case where the source and target distributions are known, one option consists of learning a general model, and adapting to each domain, e.g. via learning a set of bias vectors for each domain (Khosla et al., 2012). Another option is to embed the source and target domains in the same latent space, using e.g. Canonical Correlation Analysis Yang and Gao (2013), or minimizing the distance among the images of the source and target distributions, via minimizing Maximum Mean Discrepancy (MMD) Li et al. (2018b) or KL divergence Wang et al. (2021), or using Adversarial Learning Ganin et al. (2016a). Another option, in the realm of deep learning and computer vision, is to use semantic contrastive loss Motiian et al. (2017); Yoon et al. (2019); Mahajan et al. (2021).

2.2 Causal Inference

The fact that most real-world domains nevertheless involve hidden confounders is tackled by the *Deconfounder* approach Wang and Blei (2019). The *Deconfounder* relies on finding a factor model based on latent variables Z such that the X variables are independent on each other conditionally to the Z:

$$P(X_1, \ldots, X_n) = \Pi_i (P(X_i|Z)P(Z). \quad (1)$$

Under mild assumptions, it is suggested that the Z, referred to as *substitute hidden confounders* (SHCs), can be used to block the true hidden confounders and support a back-door adjustment. Quite a few authors (see in particular D'Amour (2019); Imai and Jiang (2019)) have been arguing however that the non-identifiability of the SHCs (the fact that the solution of Eq. 1 is not uniquely defined) undermines the validity of the *Deconfounder* approach.

2.3 Non-Linear ICA

Non-linear ICA aims to find mutually independent non-linear components, or latent features, defining a generative model of the observational data Hyvarinen et al. (2019). Non-linear ICA is hampered by the fact that simple approaches to non-linear ICA are not identifiable, in stark contrast to the linear ICA case. In the particular case where the data has a structure (e.g. temporal data), Hyvarinen and Morioka (2017, 2016) propose a general contrastive learning scheme for non-linear ICA, using the data structure to define a binary classification problem. For instance, a pair of data fragments $(x[t], x[t'])$ is labelled as 1 (respectively 0) if $t - t'$ is small (resp. big). The model learned to solve this binary classification problem induces auxiliary variables (e.g. the nodes on the last neural layer of the classifier), and the core idea is that the factors are mutually independent given the auxiliary variables. The authors show that the conditional independence of the factors given auxiliary variables is enough to establish the identifiability of the non-linear ICA, without necessarily a strict condition on the marginal independence of the factors (see also Khemakhem et al. (2020)).

An alternative to the use of contrastive losses to extract a non-linear ICA is based on adversarial learning Brakel and Bengio (2017). The authors exploit the permutation-invariant property of the mutually independent components and apply adversarial learning to identify the factorized distribution that best matches the data distribution.

3 Introduction of DeconICA

This section introduces the proposed DeconICA scheme in detail.

Preliminaries The domain generalization (DG) in image or text classification considers a set of N source domains of data, where the i-th domain is associated with a dataset $D_i = \{(x_j^i, y_j^i)\}_{j=1}^{M_i}$, containing M_i labelled samples. The features noted $\mathbf{X} = \{X_1, \ldots X_d\}$ and the label or outcome in the causal literature, noted Y, requiring the same dimension and categories in all domains. DG aims to learn a classifier with good accuracy on all source domains, that still maintains a satisfying accuracy on a target domain, which is not met in the training phase.

3.1 Problem Statement

The model for classification problems commonly aims to estimate class Y as a function of $\mathbf{X}$, e.g. the Bayes classifier $\mathbf{E}[Y|\mathbf{X} = x]$. The challenge, as discussed previously, is that each domain usually involves unobserved confounders $\mathbf{U}$ (e.g. the background of images) affecting both the extracted features $\mathbf{X}$ and the outcome (outputs of the model) Y thus causing spurious correlations. Such confounders induce a serious bias in the estimation of the outcome ($\mathbf{E}[Y|X = x] \neq Y$).

From the causal perspective, the back adjustment Pearl (2009) takes into account the con-

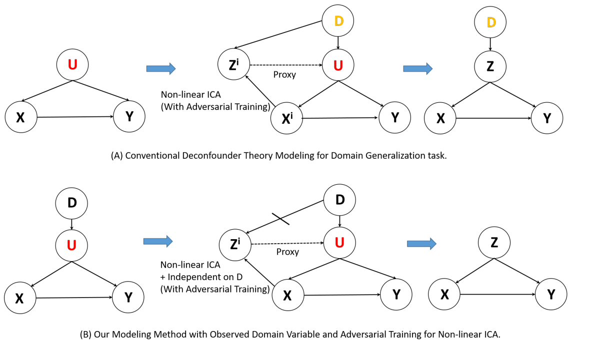

Figure 2: The DeconICA scheme. Left: Domain Generalization involves features $\mathbf{X}$, outcome (label) Y and the hidden confounders $\mathbf{U}$ depend on the domain index D. Middle: Substitute Hidden Confounders $\mathbf{Z}$ are extracted as in the *Deconfounder* scheme, and $\mathbf{Z}$ are made independent of the domain. Right: DeconICA searches for a model expressing the relationship between $\mathbf{X}$ and Y while being independent of the SHCs $\mathbf{Z}$.

founders and their impact on the extracted data features by computing

$$\mathbf{E}_u[Y|\mathbf{X} = x, \mathbf{U} = u] = Y. \quad (2)$$

The DG challenge here, is even more critical, as the spurious correlations among $\mathbf{X}$ and Y due to the confounders generally depend on the considered domain, preventing the learned model from being accurately applied in new domains with different confounders. DG thus needs to cancel out the effects of confounders.

3.2 Principle of DeconICA

The proposed DeconICA is illustrated in Figure 2. As an example, $\mathbf{X}, Y$ and $\mathbf{U}$ might respectively correspond to the data, the semantic label, and unobserved confounders such as the measurement bias. Following the *Deconfounder* principles Wang and Blei (2019), substitute hidden confounders $\mathbf{Z}$ are extracted by searching for a factorized model of $\mathbf{X}$ (Eq. 1). Specifically, the mutually independent $\mathbf{Z}$ are obtained by applying non-linear ICA Brakel and Bengio (2017) to factorize $\mathbf{X}$. The SHCs $\mathbf{Z}$ are used to further process the model: an attention mechanism is used to tune the impact of the SHCs onto the prediction, akin to a front-door intervention mechanism Pearl (1995); Yang et al. (2021). The structure of the attention mechanism, trained using a standard predictive loss, has the potential to automatically adjust the impact of the $\mathbf{Z}$ on the model depending on $\mathbf{X}$.

An originality of the proposed approach is to introduce the D variable, standing for the domain itself. By definition, D has an impact on the other confounders $\mathbf{U}$, and it could be rightly considered as part of $\mathbf{U}$. The point is that D is observed, as opposed to $\mathbf{U}$: we can thus enforce the independence of $\mathbf{Z}$ s w.r.t. D (Fig. 2, middle). By cutting off the link from the domain variable D to the SHC $\mathbf{Z}$, the latter is made invariant and robust w.r.t. the different domain biases. Therefore, the SHCs $\mathbf{Z}$ are both mutually independent and invariant w.r.t. the domain variable D. The model learns to estimate the expectation of outcome Y conditionally to both $\mathbf{X}$ and $\mathbf{Z}$.

3.3 The DeconICA Algorithm

The system of DeconICA is presented in Figure 3. The backbone model is to represent the features $\mathbf{X} = (X_1, \ldots X_d)$. These features are processed via an autoencoder with 1d convolutional operations, yielding the latent representation $\mathbf{V}$ (of the same dimension d as $\mathbf{X}$ for convenience).

This latent representation is trained using a standard reconstruction loss: denoting $\mathbf{v}$ as the encoding of $\mathbf{x}$ and $\hat{\mathbf{x}}$ as the decoding of $\mathbf{v}$ (realized as 1d convolutional blocks), it comes for the i-th domain

$$\mathcal{L}_{MSE}(i) = \sum_{j=1}^{M_i} \|\mathbf{x}_j - \hat{\mathbf{x}}_j\|^2. \tag{3}$$

The search for non-linear independent components is achieved using adversarial learning Brakel and Bengio (2017). Noting σ a random permutation on $[[1, d]]$, the discriminator D_{ICA} aims to discriminate among the latent representation $\mathbf{v}$ of the data and their permuted image $\mathbf{v}_\sigma$ (with $\mathbf{v} = (v_1, \ldots, v_d)$ and $\mathbf{v}_\sigma = (v_{\sigma(1)}, \ldots v_{\sigma(d)})$). Overall, the non-linear ICA loss is defined as the sum of the AE loss (Eq. 3) and the adversarial loss: on the i-th domain,

$$\begin{aligned} &\mathcal{L}(i) = \mathcal{L}_{MSE}(i) - \mathcal{L}_{Adv}(i), \\ &\mathcal{L}_{Adv}(i) = \sum_{j=1}^{M_i} (log(D_{ICA}(\mathbf{v}_j)) + log(1 - D_{ICA}(\mathbf{v}_{j,\sigma}))). \end{aligned} \tag{4}$$

This loss does not guarantee the identifiability of the model Khemakhem et al. (2020). To mitigate this, the loss term is augmented with a third term, an adversarial loss imposing that the latent factors be independent of the domain variable. Formally, letting $\mathbf{w} = (\mathbf{v}, i)$ denote a *paired term* if $\mathbf{v}$ is the latent representation of a sample in the i-th domain, and $\mathbf{w}' = (\mathbf{v}', j)$ denote an *unpaired term* if $\mathbf{v}'$ is the latent representation of a sample in the i-th domain with $i \neq j$, then the $\mathcal{L}_{DeconICA}$ is expressed as,

$$\begin{aligned} \mathcal{L}_{DeconICA} &= \sum_{i=1}^{N} (\mathcal{L}_{MSE}(i) - \mathcal{L}_{Adv}(i)) - \mathcal{L}_{Dom}, \\ \mathcal{L}_{Dom} &= \sum_{\mathbf{W}\ paired} log(D_{Dom}(\mathbf{w})) \\ &+ \sum_{\mathbf{W}'\ unpaired} log(1 - D_{Dom}(\mathbf{w}')). \end{aligned} \tag{5}$$

The pseudo-code of the proposed DeconICA algorithm is displayed in Algorithm 1.

Algorithm 1 DeconICA

Input data X **Output** The trained model; Encoder, Decoder, discriminators D_{ICA}, D_{Dom}.
Not converged get a batch of examples $\mathbf{x}_i$ in the source domains
$L_{AE} \leftarrow 0$
$L_{ICA} \leftarrow 0$
$L_{Dom} \leftarrow 0$
i in batch
$\mathbf{v}_i \leftarrow Encoder(\mathbf{x}_i)$
$\hat{\mathbf{x}}_i \leftarrow Decoder(\mathbf{v}_i)$
$\mathbf{w}_i = (\mathbf{v}_i, k)$ for k the domain index of $\mathbf{x}_i$
$\mathbf{w}'_i = (\mathbf{v}_i, j)$ for $j \neq k$, j in $[[1, N]]$
Draw σ permutation on $[[1, d]]$
$L_{ICA} \leftarrow L_{ICA} + log(D_{ICA}(\mathbf{v}_i)) + log(1 - D_{ICA}(\mathbf{v}_{i,\sigma}))$
$L_{AE} \leftarrow L_{AE} + \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2$
$L_{Dom} \leftarrow L_{Dom} + log(D_{Dom}(\mathbf{w}_i)) + log(1 - D_{Dom}(\mathbf{w}'_i))$
Update D_{ICA} to maximize L_{ICA}
Update D_{Dom} to maximize L_{Dom}
Update Encoder and Decoder to minimize $L_{AE} - L_{ICA} - L_{Dom}$

3.4 DG Classifier

The classifier is learned on the top of the $\mathbf{X}$ and $\mathbf{V}$ representations learned by DeconICA, with a novel fusion method based on the attention mechanism. Formally,

$$\begin{aligned} &\beta = \mathbf{X} \odot \mathbf{V}, && \textit{dot product attention} \\ &sim = \exp(\frac{\beta^i}{\sum_{j=1}^{d} \beta^j}), && \textit{attention score} \\ &\mathbf{F}_c = \mathbf{V} + sim * \mathbf{X}, && \textit{confounder features} \\ &\mathbf{F}_{final} = \mathbf{X} + \alpha * \mathbf{F}_c && \textit{final features} \end{aligned} \tag{6}$$

with α the $d \times d$ matrix, dot product attention of the region features $\mathbf{X}$ and the SHCs $\mathbf{V}$; α^i is the average of the i-th column in α; sim defines the attention score and F_c additively aggregates the information from the SHCs and the initial description biased according to sim.

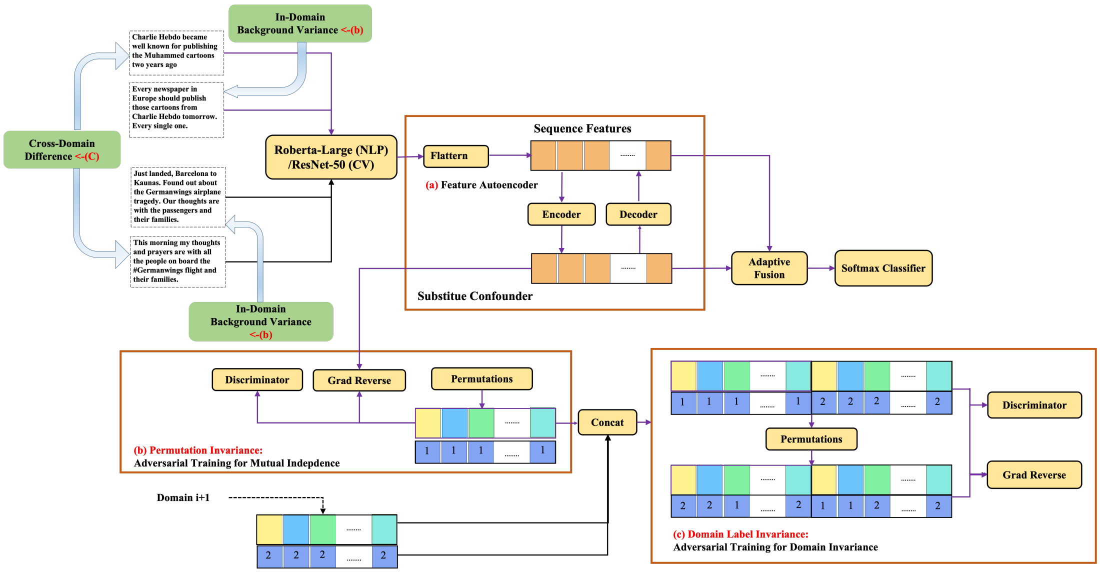

Figure 3: Neural architecture of DeconICA. The top box is an auto-encoder defining a latent representation $\mathbf{Z}$ of the features, using a standard MSE loss. The middle box enforces the mutual independence of the $\mathbf{Z}$, via an adversarial loss distinguishing the $\mathbf{Z}$ and their random permutations. The bottom box enforces the independence of the $\mathbf{Z}$ w.r.t. the domain indices, via an adversarial loss distinguishing the paired $(\mathbf{Z}, domain)$ and their unpaired equivalent (see text). Overall, the scheme learns latent variables which are independent of each other, and independent from the domain variable D.

The final features F_{final} used by the classifier elementally add the F_c and the initial $\mathbf{X}$, along the neural architecture.

The overall loss includes the standard classifier loss and λ times the DeconICA loss (Eq. 7):

$$Loss = \mathcal{L}_{Classify} + \lambda \mathcal{L}_{DeconICA}. \tag{7}$$

Note that the gradients of the classifier module operate on the whole neural architecture (the backbone network) while the gradients of the DeconICA modules are stopped and do not operate on the backbone, to achieve more stability during training.

Methods	Caltech101	LabelMe	Sun09	VOC2007	Avg
Mixup (Yan et al., 2020)	98.3	**64.8**	72.1	74.3	77.4
MLDG (Li et al., 2018a)	97.4	65.2	71.0	75.3	77.2
MMD (Li et al., 2018b)	97.7	64.0	72.8	75.3	77.5
CDANN (Li et al., 2018c)	97.1	65.1	70.7	77.1	77.5
MTL (Blanchard et al., 2021a)	97.8	64.3	71.5	75.3	77.2
SagNet (Nam et al., 2021)	97.9	64.5	71.4	77.5	77.8
ARM (Zhang et al., 2021)	98.7	63.6	71.3	76.7	77.6
VREx (Krueger et al., 2021)	98.4	64.4	74.1	76.2	78.3
RSC (Huang et al., 2020)	97.9	62.5	72.3	75.6	77.1
SelfReg (Kim et al., 2021)	48.8	41.3	57.3	40.6	47.0
PCL (Yao et al., 2022)	96.6	58.1	72.4	75.2	75.6
AdaNPC (Zhang et al., 2023)	98.9	64.5	73.5	75.6	78.1
DeconICA	**99.10**	64.12	**74.51**	**79.76**	**79.37**

Table 1: Comparative assessment of DeconICA on the VLCS dataset.

Methods	Art	Clipart	Product	Real World	Avg
MMD-AAE (Saito et al., 2018)	56.5	47.3	72.1	74.8	62.7
CCSA (Motiian et al., 2017)	59.9	49.9	74.1	75.7	64.9
JiGen (Carlucci et al., 2019)	53.0	47.5	71.5	72.8	61.2
CrossGrad (Shankar et al., 2018)	58.4	49.4	73.9	75.8	64.4
FAR (Jin et al., 2020)	61.4	**52.9**	74.5	75.4	66.0
VREx (Krueger et al., 2021)	60.7	53.0	75.3	76.7	66.4
RSC (Huang et al., 2020)	60.7	51.4	74.8	75.1	65.5
DANN (Ganin et al., 2016a)	59.9	53.0	73.6	76.9	65.9
CDANN (Li et al., 2018c)	61.5	50.4	74.4	76.6	65.7
MTL (Blanchard et al., 2021b)	61.5	52.4	74.9	76.8	66.4
SagNet (Nam et al., 2021)	63.4	54.8	75.8	78.3	68.1
ARM (Zhang et al., 2021)	58.9	51.0	74.1	75.2	64.8
SelfReg (Kim et al., 2021)	63.6	53.1	76.9	78.1	67.9
PCL (Yao et al., 2022)	62.7	54.0	76.9	78.5	68.0
AdaNPC (Zhang et al., 2023)	62.9	52.3	75.1	75.6	66.5
DeconICA	**69.8**	52.2	**77.7**	**82.2**	**70.5**

Table 2: Comparative assessment of DeconICA on the Office-Home dataset.

4 Experimental Setting

4.1 Datasets

Two publicly available benchmark datasets in computer vision are considered: The **VLCS** dataset Torralba and Efros (2011), with 5 classes, involves four datasets respectively shared by the PASCAL VOC 2007, LabelMe, Caltech and Sun. The **Office-Home** dataset Saenko et al. (2010), with 65 classes, includes 15,500 images of everyday objects in the office and home scenarios, divided into four domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr) and Real-World images (Rw). For VLCS, we follow the offi-

Methods	DVD	Electronics	Kitchen	Book	Avg
Guo et al. (Guo et al., 2018a)	87.70	89.50	90.50	87.90	88.90
Wright et al. (Wright and Augenstein, 2020)	88.90	90.30	90.80	90.00	90.00
Roberta-Large (Liu et al., 2019)	90.00	93.95	93.40	92.65	92.50
MMD (Li et al., 2018b)	89.85	94.15	93.70	92.55	92.56
MoE (Guo et al., 2018b)	90.25	94.04	93.99	92.50	92.69
Intra (Wen et al., 2016a; Ye et al., 2020)	90.06	94.00	94.06	92.75	92.72
Adv (Ganin et al., 2016b)	90.25	94.45	94.60	92.85	93.04
SCL (Tan et al., 2022a)	89.95	94.25	93.10	93.45	92.69
SCL+M=128 (Tan et al., 2022a)	91.45	95.10	95.10	93.70	93.85
DeconICA	91.75	95.00	94.50	94.75	94.00
DeconICA+M=128	**92.00**	**96.00**	**95.75**	**95.00**	**94.69**

Table 3: Comparative assessment of DeconICA on the Multi-Domain Sentiment dataset.

Methods	charlieh	ferguson	germanw	ottawashoo	sydneysiege	Avg
Wright et al. (Wright and Augenstein, 2020)	67.90	45.40	74.50	62.60	64.70	63.02
Roberta-Large (Liu et al., 2019)	64.78	43.03	69.87	60.42	62.02	60.02
MMD (Li et al., 2018b)	63.80	43.44	69.04	63.94	63.27	60.70
MoE (Guo et al., 2018b)	65.84	43.61	72.23	61.63	64.25	61.51
Intra (Wen et al., 2016a; Ye et al., 2020)	64.14	42.89	70.77	61.84	64.21	60.41
Adv (Ganin et al., 2016b)	64.83	42.23	65.74	61.47	62.81	59.45
SCL (Tan et al., 2022a)	65.57	43.22	73.03	63.50	63.52	61.77
SCL+M=128 (Tan et al., 2022a)	68.08	44.55	75.41	66.52	65.19	63.95
DeconICA	81.14	76.23	67.35	69.27	72.80	73.36
DeconICA+M=128	**83.74**	**76.81**	**74.41**	**77.07**	**74.28**	**77.26**

Table 4: Comparative assessment of DeconICA on the Multi-Domain Sentiment dataset.

cial training-val-testing split, for the Office-Home dataset, similar to the previous methods, we randomly split the training-validation into 90% and 10% samples of the original datasets.

Two publicly available benchmark datasets in natural language processing are used: The **Multi-Domain Sentiment Dataset** for cross-domain sentiment classification. The dataset consists of 8,000 Amazon product reviews, evenly distributed across four domains: DVD, Electronics, Kitchen and Book. Within each domain, there are 1,000 positive and 1,000 negative reviews. To ensure a fair comparison with previous studies, we followed the same data split Ganin et al. (2016a); Du et al. (2020); Guo et al. (2020), resulting in 1,600 training examples and 400 test examples for each domain. The **PPHEME Rumour Detection Dataset**, which includes 5,802 annotated tweets from 5 different events ((C)harlie(H)ebdo, (F)erguson, (G)erman(W)ings, (O)ttawa(S)hooting, and (S)ydneySiege) labelled as rumour or non-rumour (1,972 rumours, 3,830 non-rumours). On each benchmark, the DeconICA classifier is trained on all domains but one and tested on the remaining one. For both datasets, we follow the official training-val-testing split to perform the experimental evaluation.

4.2 Implentation Details

The DeconICA architecture implemented on the PyTorch platform (Fig. 3) is built on top of the **X** representation consisting of the last and second last convolutional features of the ResNet-50 backbone network, for the sake of a fair comparison with the baselines.

The classifier takes as input the SHCs **V** (Alg. 1) fused with **X** via an attention mechanism, and with **X** directly, along a residual connection scheme. As said, the gradient from the DeconICA module is stopped and has no impact on the backbone network.

Instead of utilizing the fully connected layers, the operations in DeconICA are realized via 1d convolutions to not only learn the relationship between data points but also reach a higher efficiency. The backbone network in image classification and DeconICA are trained with learning rate 1e-5 using Adam optimizer, with batch size 48,[1] for at most 200 epochs. Early stopping based on the validation set performance (available from the benchmark) is used. The size of the 1dconvolutional kernel is set to 7 in all benchmarks (all domains) after preliminary experiments.

All the experiments are conducted on a computing server equipped with a GPU of Nvidia Geforce 2080-Ti. The code is implemented in Python, referencing the evaluation protocol from related research. We will make the codes public upon the acceptance of our paper.

5 Experimental Validation

5.1 Comparison with other State-of-the-arts

5.1.1 Image Classification

VLCS Benchmark. The considered baselines include Domainbed Gulrajani and Lopez-Paz (2021), which proposes a platform supporting the model selection criteria for domain generalization; a Mean Maximum Discrepancy approach Li et al. (2018b) (legend MMD) that aligns the latent representation of all domains (while using adversarial learning to make the aligned distributions match a prior distribution); the CDANN approach Li et al. (2018c), using a conditional invariant adversarial network to learn domain-invariant representations; SagNet Nam et al. (2021), targeting the disentangled representations of style and content; MTL Blanchard et al. (2021a), focusing on the transfer learning of the marginal distributions in the perspective of supervised classification; RSC Huang et al. (2020), proposing an iterative

[1]More precisely, if the benchmark includes 3 training domains, the batch involves 16 samples from each domain.

self-challenging training scheme to enhance the generalization capability of the model on out-of-distribution data. The approach most related to DeconICA is CDANN Li et al. (2018c), which also aims at domain-invariant features using adversarial learning. The difference is rooted in the deconfounding approach. CDANN proceeds by combining an alignment of the domains with Gradient Reversal Layer (GRL) Ganin and Lempitsky (2015) and category-conditioned domain discrimination, while DeconICA extracts mutually independent and domain-independent substitute hidden confounders. Empirically, DeconICA outperforms all baselines on two out of four domains on the VLCS benchmark, with the best average performance (Table 1). **The Office-Home benchmark.** the considered baselines include: FAR Jin et al. (2020), that aligns and repairs the data distribution to ensure a high generalization and discrimination capacity at the same time; CrossGrad Shankar et al. (2018), that trains a label and a domain classifier on examples perturbed by loss gradients of each other's objectives, under various distribution assumptions; JiGen Carlucci et al. (2019), that proposes to solve jigsaw puzzles, to learn the spatial correlation, thus enforce good generalization capacity; CCSA Motiian et al. (2017), using a Siamese architecture to align the different domain distributions; MMD-AAE Saito et al. (2018), aimed to align the domains via considering the discrepancy of domain classifiers.

The proposed DeconICA follows the domain alignment approach, similar to CCSA Motiian et al. (2017) and MMD-AAE Saito et al. (2018), though with a quite different learning criterion. Table 2 shows that DeconICA outperforms the state of the art on all but one domain.

Methods	Caltech101	LabelMe	Sun09	VOC2007	Avg
ResNet-50	98.74	62.18	73.23	73.80	76.99
ICA Brakel and Bengio (2017) Alone	98.42	63.30	72.08	75.02	77.45
GCL Xiu et al. (2021) Alone	98.42	60.22	73.50	70.74	75.72
DeconICAw/o Attention	99.62	65.00	73.74	75.60	78.23
DeconICA(Full Model)	**99.10**	64.12	**74.51**	**79.76**	**79.37**
DeconICA(Kernel = 3)	98.46	64.04	73.83	75.06	77.85
DeconICA(Kernel = 5)	98.66	65.32	73.80	75.45	78.33
DeconICA(Kernel = 7)	**99.10**	64.12	**74.51**	**79.76**	**79.37**
DeconICA($\lambda = 0.5$)	98.94	63.80	73.14	74.63	77.51
DeconICA($\lambda = 2$)	98.82	63.50	73.10	74.14	77.39
DeconICA($\alpha = 1$)	98.58	65.12	74.06	76.93	79.17
DeconICA($\alpha = 0.16$)	**99.10**	64.12	**74.51**	**79.76**	**79.37**
DeconICAw/ $\text{Dim}_V = 0.5 \times \text{Dim}_X$	99.05	**65.36**	74.00	76.73	78.77

Table 5: DeconICA: Ablation Studies and sensitivity w.r.t. hyper-parameters on the VLCS dataset. The full DeconICA scheme has 1dconvolutional kernel of size 7, $\lambda = 1$ and $dim_V = dim_X$.

Lastly, the computational cost per iteration is compared to that of the baselines in Table 6, showing a moderate cost increase compared to ICA, essentially due to the attention mechanism.

Methods	Time(s)/Iteration
ResNet-50	0.20 s
ICA Brakel and Bengio (2017)	0.43 s
GCL Xiu et al. (2021)	0.30 s
DeconICAw/o Attention	0.44 s
DeconICA(Full Model)	0.50 s

Table 6: Training Efficiency on the VLCS Dataset.

5.1.2 Text Classification

The Multi-Domain Sentiment Dataset the considered baselines include: Roberta-Large Liu et al. (2019), that directly apply the baseline model to extract features and perform classification for domain generalization in the text;a Mean Maximum Discrepancy approach Li et al. (2018b) (legend MMD) that aligns the latent representation of all domains (while using adversarial learning to make the aligned distributions match a prior distribution); the MoE approach Guo et al. (2018a), utilizes an approach of mixture-of-experts for the domain generalization and domain adaptation. Intra Wen et al. (2016b); Ye et al. (2020) proposes a method for domain adaptation for language problems, with a feature adaptation method. The feature adaptation method applies self-distillation to make the pseudo labels of the target domain more robust, thus realizing a sample-level alignment; our baseline model Tan et al. (2022b), which applies self-supervised contrastive learning and a memory block to solve the domain generalization for text classification. Empirically, DeconICA outperforms all baselines on all four domains on the Multi-Domain Sentiment Dataset benchmark, with the best performance (Table 3). **The PHEME Rumour Detection Dataset** The considered baselines include: Roberta-Large Liu et al. (2019), which directly applies the baseline model to extract features and perform classification for domain generalization; a Mean Maximum Discrepancy approach Li et al. (2018b) (legend MMD) that aligns the latent representation of all domains; the MoE approach Guo et al. (2018a), utilizes an approach of mixture-of-experts for the domain generalization and domain adaptation. Intra Wen et al. (2016b); Ye et al. (2020) utilizes the feature adaptation method that applies the self-distillation to make the pseudo labels of the target domain more robust, thus realizing a sample-level alignment; CL, our baseline model and the SCL with memory block for training (SCL + M). Our method, as shown in Table 4, achieves the State-of-the-art performance among all the listed methods.

5.2 Ablation Study on Text classification: Multi-domain Sentiment Dataset

The impact of the different components in DeconICA is assessed using ablation studies on the Multi-Domain Sentiment Dataset. **Impact of the DeconICA Scheme over baseline** The standalone Roberta-Large backbone network trained on all domains yields the bottom performance). The non-linear ICA standalone Brakel and Bengio (2017) (DeconICAw/o Domain Invariance).). **Impact of the Convolutional Kernel size on Text** The impact of the kernel size in the 1d convolutional blocks is displayed in lines 9-10, showing a moderate sensitivity of the approach. Similar to the case of image classification, we find that 7 kernel size best fits the model. **Impact of the Hyper-parameter λ** λ controls the trade-off between the classifier loss and the deconfounder losses, displayed in lines 11-14. **Impact of the Hper-parameter α** The impact of setting the trade-off between the original features and the fused confounder features is shown in lines 2-8, showing a moderate impact. **Impact of the Batch Size** The batch size, on the text classification task, is easier to increment, due to the task's low computing resource requirement than image classification. We find that batch size has a more obvious impact on the training of the proposed scheme, a batch size of 48, has the best performance.

Methods	DVD	Electronics	Kitchen	Book	Avg
Roberta-Large Liu et al. (2019)	90.00	93.95	93.40	92.65	92.50
DeconICAw/ $\alpha = 0.24$	91.00	95.00	93.50	94.75	93.56
DeconICAw/ $\alpha = 0.20$	90.75	95.25	93.75	94.25	93.50
DeconICAw/ $\alpha = 0.16$	**91.75**	**95.00**	**94.50**	**94.75**	**94.00**
DeconICAw/ $\alpha = 0.12$	89.50	93.75	94.75	94.00	93.00
DeconICA(Kernel = 3)	91.25	95.41	**94.50**	93.50	93.67
DeconICA(Kernel = 7)	**91.75**	**95.00**	**94.50**	**94.75**	**94.00**
DeconICAw/ $\lambda = 2.5$	**91.75**	**95.00**	**94.50**	**94.75**	**94.00**
DeconICAw/ $\lambda = 2.0$	91.75	95.00	94.50	94.00	93.81
DeconICAw/ $\lambda = 1.5$	91.75	95.00	94.50	94.00	93.81
DeconICAw/ $\lambda = 1.0$	91.75	95.00	94.50	94.00	93.81
DeconICAw/ bs = 96	90.75	95.50	90.5	91.00	91.94
DeconICAw/ bs = 48	**91.75**	**95.00**	**94.50**	**94.75**	**94.00**
DeconICAw/ bs = 24	88.75	90.25	93.75	93.75	91.63
DeconICAw/ bs = 12	87.75	89.50	89.75	93.00	90.00
DeconICAw/o Domain Invariance	90.25	90.33	94.42	93.25	93.06
DeconICAw/ Domain Invariance	**91.75**	**95.00**	**94.50**	**94.75**	**94.00**

Table 7: Training Efficiency on the VLCS Dataset.

Lastly, the computational cost per iteration is compared to that of the baselines in Table 6, showing a moderate cost increase compared to ICA, essentially due to the attention mechanism.

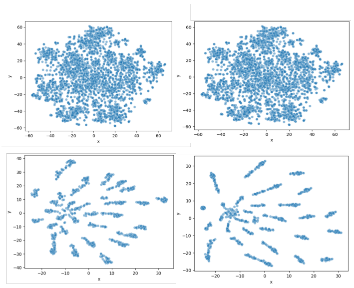

Figure 4: Quality of the **X** representation on the VOC2007 (Left) and LabelMe (Right) domains (VLCS benchmark) using a t-SNE visualization. Top: representation learned by ResNet-50. Bottom: representation learned by DeconICA.

5.3 Qualitative Evaluation: inspecting the SHCs

We investigate the DeconICA factor representation learned by DeconICA, compared to the baseline representation learned by ResNet-50. Considering the VOC2007 (Left) and LabelMe (Right) domains of the VLCS benchmark.

We apply t-SNE Van der Maaten and Hinton (2008) on the **X** representation of the data, available in both ResNet-50 and DeconICA. As shown on Fig. 4, the DeconICA scheme induces well-separated clusters of points (Fig. 4, bottom), as opposed to ResNet-50 standalone (Fig. 4, top).

6 Conclusion

Domain generalization (DG) aims to improve the generalization ability of a machine learning model in an unknown domain. This paper solves the DG task from a causal perspective, in which the poor generalization ability is considered from the hidden confounder. We model the 'dataset bias', containing the background and domain bias as the hidden confounder. Informed by the Deconfounder theory, we choose a non-linear ICA method to factorize the causes, representing the substitute confounder. These factors are subsequently trained to be domain-invariant via adversarial learning, forcing their identifiability. The proposed causal DG framework is theoretically solid. The empirical results on various classification tasks validate its effectiveness.

Limitations

There are two limitations to this research: First, more empirical results on large-scale datasets will be included in the future. Second, an improvement over the identifiability of the proposed non-linear ICA method will be studied. The current method relies on auxiliary labels, i.e., the domain labels, to achieve identifiability. This weakness raises a question of the identifiability of this method for, e.g., the single-domain domain generalization task, where there are no diverse domain labels.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1):151–175.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. *Advances in neural information processing systems (NIPS)*.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. 2021a. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55.

Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. 2021b. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100.

Philemon Brakel and Yoshua Bengio. 2017. Maximizing independence with gans for non-linear ica. In *The International Conference on Machine Learning Conference Workshops (ICMLW)*.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alexander D'Amour. 2019. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics Online (NAACL-HLT)*.

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

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4019–4028.

Kyra Gan, Andrew Li, Zachary Lipton, and Sridhar Tayur. 2021. Causal inference with selectively deconfounded data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning (ICML)*.

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

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

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7830–7838.

Jiang Guo, Darsh J Shah, and Regina Barzilay. 2018a. Multi-source domain adaptation with mixture of experts. *arXiv preprint arXiv:1809.02256*.

Jiang Guo, Darsh J Shah, and Regina Barzilay. 2018b. Multi-source domain adaptation with mixture of experts. *arXiv preprint arXiv:1809.02256*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems (NIPS)*.

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

Aapo Hyvarinen and Hiroshi Morioka. 2016. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems (NIPS)*.

Aapo Hyvarinen and Hiroshi Morioka. 2017. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. 2019. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kosuke Imai and Zhichao Jiang. 2019. Comment: The challenges of multiple causes. *Journal of the American Statistical Association*, 114(528):1605–1610.

Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2020. Feature alignment and restoration for domain generalization and adaptation. *arXiv preprint arXiv:2006.12009*.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)*.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*.

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

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018a. Learning to generalize: Meta-learning for domain generalization. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018b. Domain generalization with adversarial feature learning. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018c. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning (ICML)*.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *The IEEE international Conference on Computer Vision (ICCV)*.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *The European conference on computer vision (ECCV)*.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Domain generalization for text classification with memory-based supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6916–6926.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022b. Domain generalization for text classification with memory-based supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6916–6926.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

Ziqi Wang, Marco Loog, and Jan van Gemert. 2021. Respecting domain relations: Hypothesis invariance for domain generalization. In *The International Conference on Pattern Recognition (ICPR)*.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016a. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016b. A discriminative feature learning approach for deep face recognition. In *ECCV*.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. *arXiv preprint arXiv:2009.07806*.

Zidi Xiu, Junya Chen, Ricardo Henao, Benjamin Goldstein, Lawrence Carin, and Chenyang Tao. 2021. Supercharging imbalanced data learning with energy-based contrastive representation transfer. In *The Conference on Neural Information Processing Systems (NeurIPS)*.

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

Pei Yang and Wei Gao. 2013. Multi-view discriminant transfer learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2020. Deconfounded image captioning: A causal retrospect. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal attention for vision-language tasks. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. 2022. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. *arXiv preprint arXiv:2009.11538*.

Chris Yoon, Ghassan Hamarneh, and Rafeef Garbi. 2019. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. In *The International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. 2015. Multi-source domain adaptation: A causal view. In *The AAAI conference on artificial intelligence (AAAI)*.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. 2021. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systemsm (NIPS)*.

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2023. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*.