

# Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization

Kei Akuzawa<sup>1</sup>✉, Yusuke Iwasawa<sup>1</sup>, and Yutaka Matsuo<sup>1</sup>

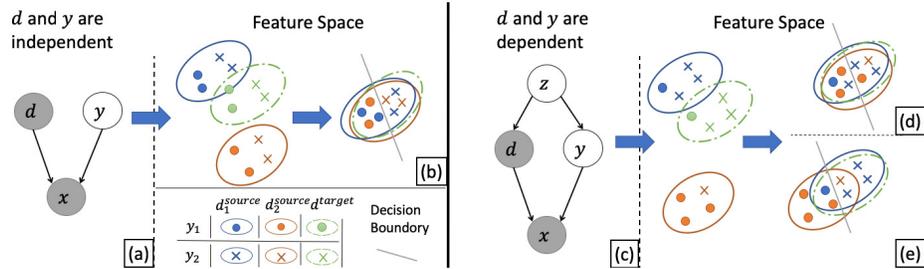
School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. {akuzawa-kei,iwasawa,matsuo}@weblab.t.u-tokyo.ac.jp

**Abstract.** Learning domain-invariant representation is a dominant approach for domain generalization (DG), where we need to build a classifier that is robust toward domain shifts. However, previous domain-invariance-based methods overlooked the underlying dependency of classes on domains, which is responsible for the trade-off between classification accuracy and domain invariance. Because the primary purpose of DG is to classify unseen domains rather than the invariance itself, the improvement of the invariance can negatively affect DG performance under this trade-off. To overcome the problem, this study first expands the analysis of the trade-off by Xie et. al. [33], and provides the notion of *accuracy-constrained domain invariance*, which means the maximum domain invariance within a range that does not interfere with accuracy. We then propose a novel method *adversarial feature learning with accuracy constraint* (AFLAC), which explicitly leads to that invariance on adversarial training. Empirical validations show that the performance of AFLAC is superior to that of domain-invariance-based methods on both synthetic and three real-world datasets, supporting the importance of considering the dependency and the efficacy of the proposed method.

**Keywords:** Invariant Feature Learning · Adversarial Training · Domain Generalization · Transfer Learning

## 1 Introduction

In supervised learning we typically assume that samples are obtained from the same distribution in training and testing; however, because this assumption does not hold in many practical situations it reduces the classification accuracy for the test data [30]. This motivates research into domain adaptation (DA) [9] and domain generalization (DG) [3]. DA methods operate in the setting where we have access to source and (either labeled or unlabeled) target domain data during training, and run some adaptation step to compensate for the domain shift. DG addresses the harder setting, where we have labeled data from several source domains and collectively exploit them such that the trained system generalizes to target domain data without requiring any access to them. Such challenges arise in many applications, e.g., hand-writing recognition (where domain shifts are induced by users, [28]), robust speech recognition (by acoustic conditions, [29]), and wearable sensor data interpretation (by users, [7]).



**Fig. 1.** Explanation of domain-class dependency and the induced trade-off. (a) When the domain and the class are independent, (b) domain invariance and classification accuracy can be optimized at the same time, and the invariance prevents the classifier from overfitting to source domains. (c) When they are dependent, a trade-off exists between these two: (d) optimal classification accuracy cannot be achieved when perfect invariance is achieved, and (e) vice versa. We propose a method to lead explicitly to (e) rather than (d), because the primary purpose for domain generalization is classification, not domain-invariance itself.

This paper considers DG under the situation where domain  $d$  and class labels  $y$  are statistically dependent owing to some common latent factor  $z$  (Figure 1-(c)), which we referred to as *domain-class dependency*. For example, the WISDM Activity Prediction dataset [16], where classes and domains correspond to activities and wearable device users, exhibits this dependency because of the (1) *data characteristics*: some activities (jogging and climbing stairs) are strenuous to the extent that some unathletic subjects avoided them, and (2) *data-collection errors*: other activities were added only after the study began and the initial subjects could not perform them. Note that the dependency is common in real-world datasets and a similar setting has been investigated in DA studies [36, 12], but most prior DG studies overlooked the dependency; moreover, we need to follow a approach separate from DA because DG methods cannot require any access to target data, as we discuss further in Sec. 2.2.

Most prior DG methods utilize invariant feature learning (IFL) [27, 7, 10, 33], which can be negatively affected by the dependency. IFL attempts to learn latent representation  $h$  from input data  $x$  which is invariant to domains  $d$ , or match multiple source domain distributions in feature space. When source and target domains have some common structure (see, [27]), matching multiple source domains leads to match source and target ones and thereby prevent the classifier from overfitting to source domains (Figure 1-(b)). However, under the dependency, merely imposing the perfect domain invariance (which means  $h$  and  $d$  are independent) adversely affects the classification accuracy as pointed out by Xie et al. [33] and illustrated in Figure 1. Intuitively speaking, since  $y$  contains information about  $d$  under the dependency, encoding information about  $d$  into  $h$  helps to predict  $y$ ; however, IFL attempts to remove all domain information from  $h$ , which causes the trade-off. Although that trade-off occurs in source domains (because we use only source data during optimization), it can also negatively

affect the classification performance for target domains. For example, if the target domain has characteristics similar (or same as an extreme case) to those of a certain source domain, giving priority to domain invariance obviously interferes with the DG performance (Figure 1-(d)).

In this paper, considering that prioritizing domain invariance under the trade-off can negatively affect the DG performance, we propose to maximize domain invariance within a range that does not interfere with the classification accuracy (Figure 1-(e)). We first expand the analysis by [33] about domain adversarial nets (DAN), a well-used IFL method, and derive Theorem 1 and 2 which show the conditions under which domain invariance harms the classification accuracy. In Theorem 3 we show that *accuracy-constrained domain invariance*, which we define as the maximum  $H(d|h)$  ( $H$  denotes entropy) value within a range that does not interfere with accuracy, equals  $H(d|y)$ . In other words, when  $H(d|h) = H(d|y)$ , i.e., the learned representation  $h$  contains as much domain information as the class labels, it does not affect the classification performance. After deriving the theorems, we propose a novel method *adversarial feature learning with accuracy constraint (AFLAC)*, which leads to that invariance on adversarial training. Empirical validations show that the performance of AFLAC is superior to that of baseline methods, supporting the importance of considering domain-class dependency and the efficacy of the proposed approach for overcoming the issue.

The main contributions of this paper can be summarized as follows. Firstly, we show that the implicit assumption of previous IFL methods, i.e., domain and class are statistically independent, is not valid in many real-world datasets, and it degrades the DG performance of them. Secondly, we theoretically show to what extent latent representation can become invariant to domains without interfering with classification accuracy. This is significant because the analysis guides the novel regularization approach that is suitable for our situation. Finally, we propose a novel method which improves domain invariance while maintaining classification performance, and it enjoys higher accuracy than the IFL methods on both synthetic and three real-world datasets.

## 2 Preliminary and Related Work

### 2.1 Problem Statement of Domain Generalization

Denote  $\mathcal{X}$ ,  $\mathcal{Y}$ , and  $\mathcal{D}$  as the input feature, class label, and domain spaces, respectively. With random variables  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$ ,  $d \in \mathcal{D}$ , we can define the probability distribution for each domain as  $p(x, y|d)$ . For simplicity this paper assumes that  $y$  and  $d$  are discrete variables. In domain generalization, we are given a training dataset consisting of  $\{x_i^s, y_i^s\}_{i=1}^{n^s}$  for all  $s \in \{1, 2, \dots, m\}$ , where each  $\{x_i^s, y_i^s\}_{i=1}^{n^s}$  is drawn from the source domain  $p(x, y|d = s)$ . Using the training dataset, we train a classifier  $g : \mathcal{X} \rightarrow \mathcal{Y}$ , and use the classifier to predict labels of samples drawn from unknown target domain  $p(x, y|d = t)$ .

## 2.2 Related Work

DG has been attracting considerable attention in recent years [27, 28]. [18] showed that non-end-to-end DG methods such as DICA [27] and MTAE [11] do not tend to outperform vanilla CNN, thus end-to-end methods are desirable. End-to-end methods based on domain invariant representation can be divided into two categories: adversarial-learning-based methods such as DAN [9, 33] and pre-defined-metric-based methods [10, 20].

In particular, our analysis and proposed method are based on DAN, which measures the invariance by using a domain classifier (also known as a discriminator) parameterized by deep neural networks and imposes regularization by deceiving it. Although DAN was originally invented for DA, [33] demonstrated its efficacy in DG. In addition, they intuitively explained the trade-off between classification accuracy and domain invariance, but did not suggest any solution to the problem except for carefully tuning a weighting parameter. AFLAC also relates to domain confusion loss [31] in that their encoders attempted to minimize Kullback-Leibler divergence (KLD) between the output distribution of the discriminators and some domain distribution ( $p(d|y)$  in AFLAC and uniform distribution in [31]), rather than to deceive the discriminator as DAN.

Several studies that address DG without utilizing IFL have been conducted. For example, CCSA [26], CIDG [21], and CIDDG [22] proposed to make use of semantic alignment, which attempts to make latent representation given class label ( $p(h|y)$ ) identical within source domains. This approach was originally proposed by [12] in the DA context, but its efficacy to overcome the trade-off problem is not obvious. Also, CIDDG, which is the only adversarial-learning-based semantic alignment method so far, needs the same number of domain classification networks as domains whereas ours needs only one. CrossGrad [28], which is one of the recent state-of-the-art DG methods, utilizes data augmentation with adversarial examples. However, because the method relies on the assumption that  $y$  and  $d$  are independent, it might not be directly applicable to our setting. MLDG [19], MetaReg [2], and Feature-Critic [23], other state-of-the-art methods, are inspired by meta-learning. These methods make no assumption about the relation between  $y$  and  $d$ ; hence, they could be combined with our proposed method in principle.

As with our paper, [21, 22] also pointed out the importance of considering the types of distributional shifts that occur, and they address the shift of  $p(y|x)$  across domains caused by the causal structure  $y \rightarrow x$ . However, the causal structure does not cause the trade-off problem as long as  $y$  and  $d$  are independent (Figure 1-(a, b)), thus it is essential to consider and address domain-class dependency problem. They also proposed to correct the domain-class dependency with the class prior-normalized weight, which enforces the prior probability for each class to be the same across domains. Its motivation is different from ours in that it is intended to avoid overfitting whereas we address the trade-off problem.

In DA, [36, 12] address the situation where  $p(y)$  changes across the source and target domains by correcting the change of  $p(y)$  using unlabeled target domain data, which is often accomplished at the cost of classification accuracy for the

source domain. However, this approach is not applicable (or necessary) to DG because we are agnostic on target domains and cannot run such adaptation step in DG. Instead, this paper is concerned with the change of  $p(y)$  within source domain and proposes to maximize the classification accuracy for source domains while improving the domain invariance.

It is worth mentioning that IFL has been used for many other context other than DG, e.g., DA [32, 9], domain transfer [17, 6], and fairness-aware classification [35, 24, 25]. However, adjusting it to each specific task is likely to improve performance. For example, in the fairness-aware classification task [25] proposed to optimize the fairness criterion directly instead of applying invariance to sensitive variables. By analogy, we adapted IFL for DG so as to address the domain-class dependency problem.

### 3 Our approach

#### 3.1 Domain Adversarial Networks

In this section, we provide a brief overview of DAN [9], on which our analysis and proposed method are based. DAN trains a domain discriminator that attempts to predict domains from latent representation encoded by an encoder, while simultaneously training the encoder to remove domain information by deceiving the discriminator.

Formally, we denote  $f_E(x)$ ,  $q_M(y|h)$ , and  $q_D(d|h)$  ( $E$ ,  $M$ , and  $D$  are their parameters) as the deterministic encoder, probabilistic model of the label classifier, and that of domain discriminator, respectively. Then, the objective function of DAN is described as follows:

$$\min_{E, M} \max_D J(E, M, D) = \mathbb{E}_{p(x, d, y)} [-\gamma L_d + L_y], \quad (1)$$

*where*  $L_d := -\log q_D(d|h = f_E(x))$ ,  $L_y := -\log q_M(y|h = f_E(x))$ .

Here, the second term in Eq. 1 simply maximizes the log likelihood of  $q_M$  and  $f_E$  as well as in standard classification problems. On the other hand, the first term corresponds to a minimax game between the encoder and discriminator, where the discriminator  $q_D(d|h)$  tries to predict  $d$  from  $h$  and the encoder  $f_E(x)$  tries to fool  $q_D(d|h)$ .

As [33] originally showed, the minimax game ensures that the learned representation has no or little domain information, i.e., the representation becomes domain-invariant. This invariance ensures that the prediction from  $h$  to  $y$  is independent from  $d$ , and therefore hopefully facilitates the construction of a classifier capable of correctly handling samples drawn from unknown domains (Figure 1-(b)). Below is a brief explanation.

Because  $h$  is a deterministic mapping of  $x$ , the joint probability distribution  $p(h, d, y)$  can be defined as follows:

$$\begin{aligned} p(h, d, y) &= \int_x p(x, d, h, y) dx = \int_x p(x, d, y) p(h|x) dx \\ &= \int_x p(x, d, y) \delta(f_E(x) = h) dx, \end{aligned} \quad (2)$$

and in the rest of the paper, we denote  $p(h, d, y)$  as  $\tilde{p}_E(h, d, y)$  because it depends on the encoder's parameter  $E$ . Using Eq. 2, Eq. 1 can be replaced as follows:

$$\min_{E, M} \max_D J(E, M, D) = \mathbb{E}_{\tilde{p}_E(h, d, y)} [\gamma \log q_D(d|h) - \log q_M(y|h)]. \quad (3)$$

Assuming  $E$  is fixed, the solutions  $M^*$  and  $D^*$  to Eq. 3 satisfy  $q_{M^*}(y|h) = \tilde{p}_E(y|h)$  and  $q_{D^*}(d|h) = \tilde{p}_E(d|h)$ . Substituting  $q_{M^*}$  and  $q_{D^*}$  into Eq. 3 enable us to obtain the following optimization problem depending only on  $E$ :

$$\min_E J(E) = -\gamma H_{\tilde{p}_E}(d|h) + H_{\tilde{p}_E}(y|h). \quad (4)$$

Solving Eq. 4 allows us to obtain the solutions  $M^*$ ,  $D^*$ , and  $E^*$ , which are in Nash equilibrium. Here,  $H_{\tilde{p}_E}(d|h)$  means conditional entropy with the joint probability distribution  $\tilde{p}_E(d, h)$ . Thus, minimizing the second term in Eq. 4 intuitively means learning (the mapping function  $f_E$  to) the latent representation  $h$  which contains as much information about  $y$  as possible. On the other hand, the first term can be regarded as a regularizer that attempts to learn  $h$  that is invariant to  $d$ .

### 3.2 Trade-off Caused by Domain-Class Dependency

Here we show that the performance of DAN is impeded by the existence of domain-class dependency. Concretely, we show that the dependency causes the trade-off between classification accuracy and domain invariance: when  $d$  and  $y$  are statistically dependent, no values of  $E$  would be able to optimize the first and second term in Eq. 4 at the same time. Note that the following analysis also suggests that most IFL methods are negatively influenced by the dependency.

To begin with, we consider only the first term in Eq. 4 and address the optimization problem:

$$\min_E J_1(E) = -\gamma H_{\tilde{p}_E}(d|h). \quad (5)$$

Using the property of entropy,  $H_{\tilde{p}_E}(d|h)$  is bounded:

$$H_{\tilde{p}_E}(d|h) \leq H(d). \quad (6)$$

Thus, Eq. 5 has the solution  $E_1^*$  which satisfies the following condition:

$$H_{\tilde{p}_{E_1^*}}(d|h) = H(d). \quad (7)$$

Eq. 7 suggests that the regularizer in DAN is intended to remove all information about domains from latent representation  $h$ , thereby ensuring the independence of domains and latent representation.

Next, we consider only the second term in Eq. 4, thereby addressing the following optimization problem:

$$\min_E J_2(E) = H_{\bar{p}_E}(y|h). \quad (8)$$

Considering  $h$  is the mapping of  $x$ , i.e.,  $h = f_E(x)$ , the solution  $E_2^*$  to Eq. 8 satisfies the following equation:

$$H_{\bar{p}_{E_2^*}}(y|h) = H(y|x). \quad (9)$$

Here we obtain  $E_1^*$  and  $E_2^*$ , which can achieve perfect invariance and optimal classification accuracy, respectively. Using them, we can obtain the following theorem, which shows the existence of the trade-off between invariance and accuracy: perfect invariance ( $E_1^*$ ) and optimal classification accuracy ( $E_2^*$ ) cannot be achieved at the same time.

**Theorem 1** *When  $H(y|x) = 0$ , i.e., there is no labeling error, and  $H(d) > H(d|y)$ , i.e., the domain and class are statistically dependent,  $E_1^* \neq E_2^*$  holds.*

**Proof 1** *Assume  $E_1^* = E_2^* = E^*$ . Using the properties of entropy, we can obtain the following:*

$$H_{\bar{p}_{E^*}}(d|h) \leq H_{\bar{p}_{E^*}}(d, y|h) = H_{\bar{p}_{E^*}}(d|h, y) + H_{\bar{p}_{E^*}}(y|h) \leq H_{\bar{p}_{E^*}}(d|y) + H_{\bar{p}_{E^*}}(y|h). \quad (10)$$

*Substituting  $H_{\bar{p}_{E^*}}(y|h) = H(y|x)$  and  $H_{\bar{p}_{E^*}}(d|h) = H(d)$  into Eq. 10, we can obtain the following condition:*

$$H(d) - H(d|y) \leq H(y|x). \quad (11)$$

*Because the domain and class are dependent on each other, the following condition holds:*

$$0 < H(d) - H(d|y) \leq H(y|x), \quad (12)$$

*but Eq. 12 contradicts with  $H(y|x) = 0$ . Thus,  $E_1^* \neq E_2^*$ .*

Theorem 1 shows that the domain-class dependency causes the trade-off problem. Although it assumes  $H(y|x) = 0$  for simplicity, we cannot know the true value of  $H(y|x)$  and there are many cases in which little or no labeling errors occur and thus  $H(y|x)$  is close to 0.

In addition, we can omit the assumption and obtain a more general result:

**Theorem 2** *When  $I(d; y) := H(d) - H(d|y) > H(y|x)$ ,  $E_1^* \neq E_2^*$  holds.*

**Proof 2** *Similar to Proof 1, we assume that  $E_1^* = E_2^*$  and thus Eq. 11 is obtained. Obviously, Eq. 11 does not hold when  $H(d) - H(d|y) > H(y|x)$ .*

Theorem 2 shows that when the mutual information of the domain and class  $I(d; y)$  is greater than the labeling error  $H(y|x)$ , the trade-off between invariance and accuracy occurs. Then, although we cannot know the true value of  $H(y|x)$ , the performance of DAN and other IFL methods are likely to decrease when  $I(d; y)$  has large value.

### 3.3 Accuracy-Constrained Domain Invariance

If we cannot avoid the trade-off, the next question is to decide how to accommodate it, i.e., to what extent the representation should become domain-invariant for DG tasks. Here we provide the notion of accuracy-constrained domain invariance, which is the maximum domain invariance within a range that does not interfere with the classification accuracy. The reason for the constraint is that the primary purpose of DG is the classification for unseen domains rather than the invariance itself, and the improvement of the invariance could detrimentally affect the performance. For example, in WISDM, if we know the target activity was performed by a young rather than an old man, we might predict the activity to be jogging with a higher probability; thus, we would have to avoid removing such domain information that may be useful in the classification task.

**Theorem 3** *Define accuracy-constrained domain invariance as the maximum  $H_{\bar{p}_E}(d|h)$  value under the constraint that  $H(y|x) = 0$ , i.e., there is no labeling error, and classification accuracy is maximized, i.e.,  $H_{\bar{p}_E}(y|h) = H(y|x)$ . Then, accuracy-constrained domain invariance equals  $H(d|y)$ .*

**Proof 3** *Using Eq. 10 and  $H_{\bar{p}_E}(y|h) = H(y|x)$ , the following inequation holds:*

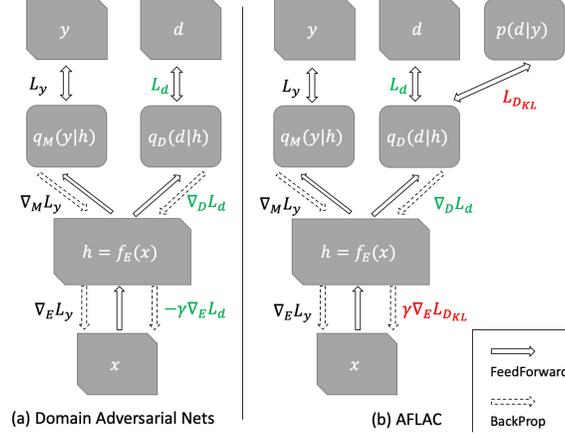
$$H_{\bar{p}_E}(d|h) \leq H(y|x) + H(d|y). \quad (13)$$

*Substituting  $H(y|x) = 0$  into Eq. 13, the following inequation holds:*

$$H_{\bar{p}_E}(d|h) \leq H(d|y). \quad (14)$$

*Thus, the maximum  $H_{\bar{p}_E}(d|h)$  value under the optimal classification accuracy constraint is  $H(d|y)$ .*

Note that we could improve the invariance more when  $H(y|x) > 0$  (that is obvious considering Eq. 13), but we cannot know the true value of  $H(y|x)$  as we discussed in Sec. 3.2. Thus, accuracy-constrained domain invariance can be viewed as the worst-case guarantee.



**Fig. 2.** Comparative illustration of DAN and AFLAC. (a) The classifier and discriminator try to minimize  $L_y$  and  $L_d$ , respectively. The encoder tries to minimize  $L_y$  and maximize  $L_d$  (fool the discriminator). (b) The discriminator tries to approximate true  $\tilde{p}_E(d|h)$  by minimizing  $L_d$ . The encoder tries to minimize divergence between  $\tilde{p}_E(d|h)$  and  $p(d|y)$  by minimizing  $L_{D_{KL}}$ .

### 3.4 Proposed Method

Based on the above analysis, the remaining challenge is to determine how to achieve accuracy-constrained domain invariance, i.e., imposing regularization such that makes  $H_{\tilde{p}_E}(d|h) = H(d|y)$  holds. Although DAN might be able to achieve this condition by carefully tuning the strength of the regularizer ( $\gamma$  in Eq. 1), such tuning is time-consuming and impractical, as suggested by our experiments. Alternatively, we propose a novel method named AFLAC by modifying the regularization term of DAN: whereas the encoder of DAN attempts to fool the discriminator, that of AFLAC attempts to directly minimize the KLD between  $p(d|y)$  and  $q_D(d|h)$ . Formally, AFLAC solves the following joint optimization problem by alternating gradient descent.

$$\min_D W(E, D) = \mathbb{E}_{p(x,d)}[L_d] \quad (15)$$

$$\min_{E,M} V(E, M) = \mathbb{E}_{p(x,d,y)}[\gamma L_{D_{KL}} + L_y], \quad (16)$$

$$\text{where } L_{D_{KL}} := D_{KL}[p(d|y)|q_D(d|h = f_E(x))].$$

The minimization of  $L_y$  and  $L_d$ , respectively, means maximization of the log-likelihood of  $q_M$  and  $q_D$  as well as in DAN. However, the minimization of  $L_{D_{KL}}$  differs from the regularizer of DAN in that it is intended to satisfy  $q_D(d|h) = p(d|y)$ . And if  $q_D(d|h)$  well approximates  $\tilde{p}_E(d|h)$  by the minimization of  $L_d$  in Eq. 15, the minimization of  $L_{D_{KL}}$  leads to  $\tilde{p}_E(d|h) = p(d|y)$ . Figure 2-(b) outlines the training of AFLAC.

Here we formally show that AFLAC is intended to achieve  $H_{\tilde{p}_E}(d|h) = H(d|y)$  (accuracy-constrained domain invariance) by a Nash equilibrium analysis similar to [13, 33]. As well as in Section 3.1,  $D^*$  and  $M^*$ , which are the solutions to Eqs. 15, 16 with fixed  $E$ , satisfy  $q_D^* = \tilde{p}_E(d|h)$  and  $q_M^* = \tilde{p}_E(y|h)$ , respectively. Thus,  $V$  in Eq. 16 can be written as follows:

$$V(E) = \mathbb{E}[\gamma D_{KL}[p(d|y)|\tilde{p}_E(d|h)]] + H_{\tilde{p}_E}(y|h). \quad (17)$$

$E^*$ , which is the solution to Eq. 17 and in Nash equilibrium, satisfies not only  $H_{\tilde{p}_{E^*}}(y|h) = H(y|x)$  (optimal classification accuracy) but also  $\mathbb{E}_{h,y \sim \tilde{p}_{E^*}(h,y)}[D_{KL}[p(d|y)|\tilde{p}_{E^*}(d|h)]] = 0$ , which is a sufficient condition for  $H_{\tilde{p}_{E^*}}(d|h) = H(d|y)$  by the definition of the conditional entropy.

In training,  $p(x, d, y)$  in the objectives (Eqs. 15, 16) is approximated by empirical distribution composed of the training data obtained from  $m$  source domains, i.e.,  $\{x_i^{(1)}, y_i^{(1)}, d = 1\}_{i=1}^{n^{(1)}}, \dots, \{x_i^{(m)}, y_i^{(m)}, d = m\}_{i=1}^{n^{(m)}}$ . Also,  $p(d|y)$  used in Eq. 16 can be replaced by the maximum likelihood or maximum a posteriori estimator of it. Note that, we could use some distances other than  $D_{KL}[p(d|y)|q_D(d|h)]$  in Eq. 16, e.g.,  $D_{KL}[q_D(d|h)|p(d|y)]$ , but in doing so, we could not observe performance gain, hence we discontinued testing them.

## 4 Experiments

### 4.1 Datasets

Here we provide a brief overview of one synthetic and three real-world datasets (PACS, WISDM, IEMOCAP) used for the performance evaluation. Although WISDM and IEMOCAP have not been widely used in DG studies, previous human activity recognition and speech emotion recognition studies (e.g., [1, 8, 5]) used them in the domain generalization setting (i.e., source and target domains are disjoint), so they can be regarded as the practical use case of domain generalization. The concrete sample sizes for each  $d$  and  $y$ , and the network architectures for each dataset are shown in supplementary.<sup>1</sup>

**BMNISTR** We created the Biased and Rotated MNIST dataset (BMNISTR) by modifying the sample size of the popular benchmark dataset MNISTR [11], such that the class distribution differed among the domains. In MNISTR, each class is represented by 10 digits. Each domain was created by rotating images by 15 degree increments: 0, 15, 30, 45, 60, and 75 (referred to as M0, ..., M75). Each image was cropped to  $16 \times 16$  in accordance with [11]. We created three variants of MNISTR that have different types of domain-class dependency, referred to as BMNISTR-1 through BMNISTR-3. As shown in Table 1-left, BMNISTR-1, -2 have similar trends but different degrees of dependency, whereas BMNISTR-1 and BMNISTR-3 differ in terms of their trends.

**PACS** The PACS dataset [18] contains 9991 images across 7 categories (dog, elephant, giraffe, guitar, house, horse, and person) and 4 domains comprising

<sup>1</sup> Code and Supplementary are available at <https://github.com/akuzeee/AFLAC>

different stylistic depictions (Photo, Art painting, Cartoon, and Sketch). It has domain-class dependency probably owing to the data characteristics. For example,  $p(y = \text{person} | d = \text{Phot})$  is much higher than  $p(y = \text{person} | d = \text{Sketch})$ , indicating that photos of a person are easier to obtain than those of animals, but sketches of persons are more difficult to obtain than those of animals in the wild. For training, we used the ImageNet pre-trained AlexNet CNN [15] as the base network, following previous studies [18, 19]. The two-FC-layer discriminator was connected to the last FC layer, following [9].

**WISDM** The WISDM Activity Prediction dataset contains sensor data of accelerometers of six human activities (walking, jogging, upstairs, downstairs, sitting, and standing) performed by 36 users (domains). WISDM has the dependency for the reason noted in Sec. 1. In data preprocessing, we use the sliding-window procedure with 60 frames (=3 seconds) referring to [1], and the total number of samples was 18210. We parameterized the encoder using three 1-D convolution layers followed by one FC layer and the classifier by logistic regression, following previous studies [34, 14].

**IEMOCAP** The IEMOCAP dataset [4] is the popular benchmark dataset for speech emotion recognition (SER), which aims at recognizing the correct emotional state of the speaker from speech signals. It contains a total of 10039 utterances pronounced by ten actors (domains, referred to as Ses01F, Ses01M through Ses05F, Ses05M) with emotional categories, and we only consider the four emotional categories (happy, angry, sad, and neutral) referring to [5, 8]. Also, we referred to [5] about data preprocessing: we split the speech signal into equal-length segments of 3s, and extracted 40-dimensional log Mel-spectrogram, its deltas, and delta-deltas. We parameterized the encoder using three 2-D convolution layers followed by one FC layer and the classifier by logistic regression.

## 4.2 Baselines

To demonstrate the efficacy of the proposed method AFLAC, we compared it with vanilla CNN and adversarial-learning-based methods. Specifically, **(1) CNN** is a vanilla convolutional networks trained on the aggregation of data from all source domains. Although CNN has no special treatments for DG, [18] reported that it outperforms many traditional DG methods. **(2) DAN** [33] is expected to generalize across domains utilizing domain-invariant representation, but it can be affected by the trade-off between domain invariance and accuracy as explained in Section 3.2. **(3) CIDDG** is our re-implementation of the method proposed in [22], which is designed to achieve semantic alignment on adversarial training. Additionally, we used **(4) AFLAC-Abl**, which is a version of AFLAC modified for ablation studies. AFLAC-Abl replaces  $D_{KL}[p(d|y)|q_D(d|h)]$  in Eq. 16 of  $D_{KL}[p(d)|q_D(d|h)]$ , thus it attempts to learn the representation that is perfectly invariant to domains or make  $H(d|h) = H(d)$  hold as well as DAN. Comparing AFLAC and AFLAC-Abl, we measured the genuine effect of taking domain-class dependency into account. When training AFLAC and AFLAC-Abl, we cannot obtain true  $p(d|y)$  and  $p(d)$ , hence we used their maximum likelihood estimators for calculating the KLD terms.

**Table 1.** Left: Sample sizes for each domain-class pair in BMNISTR. Those for the classes 0~4 are variable across domains, whereas the classes 5~9 have identical sample sizes across domains. Right: Mean F-measures for the classes 0~4 and classes 5~9 with the target domain M0. RI denotes relative improvement of AFLAC to AFLAC-Abl

Dataset	Class	M0	M15	M30	M45	M60	M75	Dataset	Class	CNN	DAN	CIDDG	AFLAC	AFLAC	RI
													-Abl		
BMNISTR-1	0~4	100	85	70	55	40	25	BMNISTR-1	0~4	83.86	84.54	87.50	87.46	<b>90.62</b>	3.6%
	5~9	100	100	100	100	100	100		5~9	83.90	85.24	87.46	86.46	<b>88.10</b>	1.9%
BMNISTR-2	0~4	100	90	80	70	60	50	BMNISTR-2	0~4	82.54	85.30	87.64	88.60	<b>89.64</b>	1.2%
	5~9	100	100	100	100	100	100		5~9	82.18	85.80	86.74	87.60	<b>89.04</b>	1.6%
BMNISTR-3	0~4	100	25	100	25	100	25	BMNISTR-3	0~4	71.26	79.22	76.76	76.56	<b>80.02</b>	4.5%
	5~9	100	100	100	100	100	100		5~9	78.62	<b>83.14</b>	82.64	82.94	82.80	-0.2%

### 4.3 Experimental Settings

For all the datasets and methods, we used RMSprop for optimization. Further, we set the learning rate, batch size, and the number of iterations as 5e-4, 128, and 10k for BMNISTR; 5e-5, 64, and 10k for PACS; 1e-4, 64, and 10k for IEMOCAP; 5e-4 (with exponential decay with decay step 18k and 24k, and decay rate 0.1), 128, and 30k for WISDM, respectively. Also, we used the annealing of weighting parameter  $\gamma$  proposed in [9], and unless otherwise mentioned chose  $\gamma$  from  $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$  for DAN, CIDDG, AFLAC-Abl, and AFLAC. Specifically, on BMNISTR and PACS, we employed a leave-one-domain-out setting [11], i.e., we chose one domain as target and used the remaining domains as source data. Then we split the source data into 80% of training data and 20% of validation data, assuming that target data are not absolutely available in the training phase. On IEMOCAP, we chose the best  $\gamma$  from  $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$  using disjoint validation domain, referring to [8, 5]. On WISDM, we randomly selected  $\langle 20 / 16 \rangle$  users as  $\langle \text{source} / \text{target} \rangle$  domains, and split the source data into training and validation data because one-domain-leave-out evaluation is computationally expensive. Then, we conducted experiments multiple times with different random weight initialization; we trained the models on 10, 20, and 20 seeds in BMNISTR, WISDM, and IEMOCAP, chose the best hyperparameter that achieved the highest validation accuracies measured in each epoch, and reported the mean scores (accuracies and F-measures) for the hyperparameter. On PACS, because it requires a long time to train on, we chose the best  $\gamma$  from  $\{0.0001, 0.001, 0.01, 0.1\}$  after three experiments, and reported the mean scores in experiments with 15 seeds.

### 4.4 Results

We first investigated the extent to which domain-class dependency affects the performance of the IFL methods. In Table 1-right, we compared the mean F-measures for the classes 0 through 4 and classes 5 through 9 in BMNISTR with the target domain M0. Recall that the sample sizes for the classes 0~4 are variable

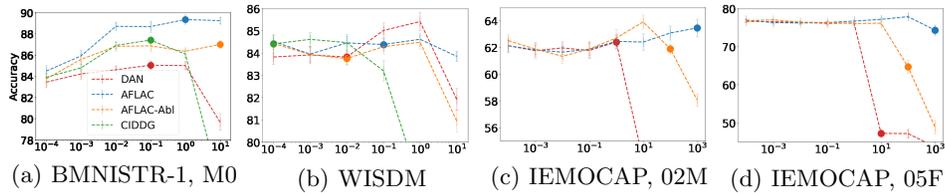
**Table 2.** Accuracies for each dataset and target domain. The  $I(d; y)$  column is estimated from source domain data, which indicates the domain-class dependency.

Dataset	Target	$I(d; y)$	CNN	DAN	CIDDG	AFLAC-Abl	AFLAC
BMNISTR-1	M0	0.026	83.9 ± 0.4	85.0 ± 0.4	87.4 ± 0.3	87.0 ± 0.4	<b>89.3 ± 0.4</b>
	M15	0.034	98.5 ± 0.2	98.5 ± 0.1	98.3 ± 0.2	98.3 ± 0.2	<b>98.8 ± 0.1</b>
	M30	0.037	97.5 ± 0.1	97.4 ± 0.1	97.4 ± 0.2	97.6 ± 0.1	<b>98.3 ± 0.2</b>
	M45	0.036	89.9 ± 0.9	90.2 ± 0.6	89.8 ± 0.5	92.8 ± 0.5	<b>93.3 ± 0.6</b>
	M60	0.030	96.7 ± 0.3	97.0 ± 0.2	97.2 ± 0.1	96.6 ± 0.2	<b>97.4 ± 0.2</b>
	M75	0.017	87.1 ± 0.5	87.3 ± 0.4	<b>88.2 ± 0.3</b>	87.7 ± 0.5	88.1 ± 0.4
	Avg		92.3	92.6	93.1	93.3	<b>94.2</b>
BMNISTR-2	Avg		92.2	92.7	93.1	94.0	<b>94.5</b>
BMNISTR-3	Avg		90.6	91.7	91.4	91.6	<b>92.9</b>
PACS	photo	0.102	82.2 ± 0.4	81.8 ± 0.4	-	82.5 ± 0.4	<b>83.5 ± 0.3</b>
	art_painting	0.117	61.0 ± 0.5	60.9 ± 0.5	-	62.6 ± 0.4	<b>63.3 ± 0.3</b>
	cartoon	0.131	<b>64.9 ± 0.5</b>	<b>64.9 ± 0.6</b>	-	64.2 ± 0.3	<b>64.9 ± 0.3</b>
	sketch	0.023	<b>61.4 ± 0.5</b>	<b>61.4 ± 0.5</b>	-	59.6 ± 0.7	60.1 ± 0.7
	Avg		67.4	67.2	-	67.2	<b>68.0</b>
WISDM	16 users	0.181	84.0 ± 0.4	83.8 ± 0.3	<b>84.4 ± 0.4</b>	83.7 ± 0.3	<b>84.4 ± 0.3</b>
IEMOCAP	Ses01F	0.005	56.0 ± 0.7	60.1 ± 0.7	-	<b>62.9 ± 0.5</b>	60.4 ± 0.9
	Ses01M		61.0 ± 0.3	63.5 ± 0.5	-	<b>68.0 ± 0.5</b>	66.1 ± 0.3
	Ses02F	0.045	61.2 ± 0.5	60.4 ± 0.5	-	<b>65.8 ± 0.5</b>	64.2 ± 0.4
	Ses02M		<b>76.6 ± 0.4</b>	47.2 ± 0.7	-	64.7 ± 1.7	74.3 ± 1.3
	Ses03F	0.037	69.2 ± 0.9	<b>71.9 ± 0.4</b>	-	70.0 ± 0.6	70.1 ± 0.4
	Ses03M		56.9 ± 0.4	<b>57.3 ± 0.5</b>	-	56.2 ± 0.4	56.8 ± 0.4
	Ses04F	0.120	75.5 ± 0.5	75.5 ± 0.6	-	75.4 ± 0.6	<b>75.7 ± 0.6</b>
	Ses04M		58.5 ± 0.5	57.4 ± 0.5	-	58.7 ± 0.5	<b>59.2 ± 0.5</b>
	Ses05F	0.063	61.8 ± 0.4	62.4 ± 0.5	-	61.9 ± 0.3	<b>63.4 ± 0.7</b>
	Ses05M		47.6 ± 0.3	46.9 ± 0.4	-	49.6 ± 0.4	<b>49.9 ± 0.4</b>
	Avg		62.4	60.3	-	63.3	<b>64.0</b>

across domains, whereas the classes 5~9 have identical sample sizes across domains (Table 1-left). The F-measures show that AFLAC outperformed baselines in most dataset-class pairs, which supports that domain-class dependency reduces the performance of domain-invariance-based methods and that AFLAC can mitigate the problem. Further, the relative improvement of AFLAC to AFLAC-Abl is more significant for the classes 0~4 than for 5~9 in BMNISTR-1 and BMNISTR-3, suggesting that AFLAC tends to increase performance more significantly for classes in which the domain-class dependency occurs. Moreover, the improvement is more significant in BMNISTR-1 than in BMNISTR-2, suggesting that the stronger the domain-class dependency is, the lower the performance of domain-invariance-based methods becomes. This result is consistent with Theorem 2, which shows that the trade-off is likely to occur when  $I(d; y)$  is large. Finally, although the dependencies of BMNISTR-1 and BMNISTR-3 have different trends, AFLAC improved the F-measures in both datasets.

Next we compared the mean accuracies (with standard errors) in both synthetic (BMNISTR) and real-world (PACS, WISDM, and IEMOCAP) datasets (Table 2). Note that the performance of our baseline CNN on PACS, WISDM, and IEMOCAP is similar but partly different from that reported in previous studies ([22], [1], and [8], respectively) probably because the DG performance strongly depends on validation methods and other implementation details as reported in many recent studies [1, 8, 2, 23]. Also, we trained CIDDG only on BMNISTR and WISDM due to computational resource constraint. This table enables us to make the following observations. **(1)** Domain-class dependency in real-world datasets negatively affects the DG performance of IFL methods. The results obtained on PACS (Avg) and WISDM showed that the vanilla CNN outperformed the IFL methods (DAN and AFLAC-Abl). Additionally, the results on IEMOCAP shows that AFLAC tended to outperform AFLAC-Abl when  $I(d; y)$  had large values (in Ses04 and Ses05), which is again consistent with Theorem 2. These results support the importance of considering domain-class dependency in real-world datasets. **(2)** AFLAC performed better than the baselines on all the datasets in average, except for CIDDG on WISDM. Note that AFLAC is more parameter efficient than CIDDG as we noted in Sec. 2.2. These results supports the efficacy of the proposed model to overcome the trade-off problem.

Finally, we investigated the relationship between the strength of regularization and performance. In DG, it is difficult to choose appropriate hyperparameters because we cannot use target domain data at validation step (since they are not available during training); therefore, hyperparameter insensitivity is significant in DG. Figure 3 shows the hyperparameter sensitivity of the classification accuracies for DAN, CIDDG, AFLAC-Abl, and AFLAC. These figures suggest that DAN and AFLAC-Abl sometimes outperformed AFLAC with appropriate  $\gamma$  values, but there is no guarantee that such  $\gamma$  values will be chosen by validation whereas AFLAC is robust toward hyperparameter choice. Specifically, as shown in Figures 3-(b, d), DAN and AFLAC-Abl outperformed AFLAC with  $\gamma = 1$  and 10, respectively. One possible explanation of those results is that accuracy for target domain sometimes improves by giving priority to domain invariance at the cost of the accuracies for source domains, but AFLAC improves domain invariance only within a range that does not interfere with accuracy for source domains. However, as shown in Figure 3, the performance of DAN and AFLAC-Abl are sensitive to hyperparameter choice. For example, although they got high scores with  $\gamma = 1$  in Figure 3-(b), the scores dropped rapidly when  $\gamma$  increases to 10 or decreases to 0.01. Also, the scores of DAN and AFLAC-Abl in Figure 3-(c) dropped significantly with  $\gamma > 10$ , and such large  $\gamma$  was indeed chosen by overfitting to validation domain. On the other hand, Figures 3-(a, b, c, d) show that the accuracy gaps of AFLAC-Abl and AFLAC increase with strong regularization (such as when  $\gamma = 10$  or 100). These results suggest that AFLAC, as it was designed, does not tend to reduce the classification accuracy with strong regularizer, and such robustness of AFLAC might have yielded the best performance shown in Table 2.



**Fig. 3.** Classification Accuracy with various  $\gamma$ . Each caption shows dataset name and target domain. The round markers correspond to  $\gamma$  values chosen by validation. The error bars correspond to standard errors.

## 5 Conclusion

In this paper, we addressed domain generalization under domain-class dependency, which was overlooked by most prior DG methods relying on IFL. We theoretically showed the importance of considering the dependency and the way to overcome the problem by expanding the analysis of [33]. We then proposed a novel method AFLAC, which maximizes domain invariance within a range that does not interfere with classification accuracy on adversarial training. Empirical validations show the superior performance of AFLAC to the baseline methods, supporting the importance of the domain-class dependency in DG tasks and the efficacy of the proposed method to overcome the issue.

## References

1. Andrey, I.: Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing* (2017)
2. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: Metareg: Towards domain generalization using meta-regularization. In: *Advances in Neural Information Processing Systems* 31 (2018)
3. Blanchard, G., Lee, G., Scott, C.: Generalizing from several related classification tasks to a new unlabeled sample. In: *Proc. of the 24th International Conference on Neural Information Processing Systems* (2011)
4. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* **42**(4), 335 (Nov 2008)
5. Chen, M., He, X., Yang, J., Zhang, H.: 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters* **25**, 1–1 (07 2018)
6. Chou, J.C., chieh Yeh, C., yi Lee, H., shan Lee, L.: Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In: *Proc. Interspeech* (2018)
7. Erfani, S., Baktashmotlagh, M., Moshtaghi, M., Nguyen, V., Leckie, C., Bailey, J., Kotagiri, R.: Robust domain generalisation by enforcing distribution invariance. In: *25th International Joint Conference on Artificial Intelligence* (2016)

8. Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., Schmauch, B.: Speech emotion recognition with data augmentation and layer-wise learning rate adjustment. CoRR **abs/1802.05630** (2018), <http://arxiv.org/abs/1802.05630>
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* (2016)
10. Ghifary, M., Balduzzi, D., Kleijn, W.B., Zhang, M.: Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)
11. Ghifary, M., Bastiaan Kleijn, W., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2015)
12. Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., Schölkopf, B.: Domain adaptation with conditional transferable components. In: Proc. of the 33rd International Conference on International Conference on Machine Learning (2016)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. of the 27th International Conference on Neural Information Processing Systems (2014)
14. Iwasawa, Y., Nakayama, K., Yairi, I., Matsuo, Y.: Privacy issues regarding the application of dnns to activity-recognition using wearables and its countermeasures by use of adversarial training. In: Proc. of the 26th International Joint Conference on Artificial Intelligence. pp. 1930–1936 (2017)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc. of the 25th International Conference on Neural Information Processing Systems. pp. 1097–1105 (2012)
16. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.* (2011)
17. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: manipulating images by sliding attributes. In: Proc. of the 30th Neural Information Processing Systems (2017)
18. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2017)
19. Li, D., Yang, Y., Song, Y., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. In: Proc. of the 32nd AAAI Conference on Artificial Intelligence (2018)
20. Li, H., Jialin Pan, S., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
21. Li, Y., Gong, M., Tian, X., Liu, T., Tao, D.: Domain generalization via conditional invariant representations. In: Proc. of the 32nd AAAI Conference on Artificial Intelligence (2018)
22. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: The European Conference on Computer Vision (ECCV) (September 2018)
23. Li, Y., Yang, Y., Zhou, W., Hospedales, T.M.: Feature-critic networks for heterogeneous domain generalization. CoRR **abs/1901.11448** (2019), <http://arxiv.org/abs/1901.11448>
24. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.S.: The variational fair autoencoder. In: Proc. International Conference on Representation Learning (2016)

25. Madras, D., Creager, E., Pitassi, T., Zemel, R.S.: Learning adversarially fair and transferable representations. In: Proc. of the 35th International Conference on Machine Learning (2018)
26. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2017)
27. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: Proc. of the 30th International Conference on Machine Learning (2013)
28. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. In: Proc. International Conference on Learning Representations (2018)
29. Sriram, A., Jun, H., Gaur, Y., Satheesh, S.: Robust speech recognition using generative adversarial networks. In: The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
30. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (2011)
31. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2015)
32. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. CoRR **abs/1412.3474** (2014), <http://arxiv.org/abs/1412.3474>
33. Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: Proc. of the 30th International Conference on Neural Information Processing Systems (2017)
34. Yang, J., Nguyen, M.N., San, P.P., Li, X., Krishnaswamy, S.: Deep convolutional neural networks on multichannel time series for human activity recognition. In: Proc. of the 24th International Joint Conference on Artificial Intelligence (2015)
35. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Proc. of the 30th International Conference on Machine Learning (2013)
36. Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain adaptation under target and conditional shift. In: Proc. of the 30th International Conference on Machine Learning (2013)