Domain-invariant Feature Exploration for Domain Generalization

Anonymous authors
Paper under double-blind review

Abstract

Deep learning has achieved great success in the past few years. However, the performance of deep learning is likely to impede in face of non-IID situations. Domain generalization (DG) enables a model to generalize to an unseen test distribution, i.e., to learn domain-invariant representations. In this paper, we argue that domain-invariant features should be originating from both internal and mutual sides. Internal invariance means that the features can be learned with a single domain and the features capture intrinsic semantics of data, i.e., the property within a domain, which is agnostic to other domains. Mutually invariance means that the features can be learned with multiple domains (cross-domain) and the features contain common information, i.e., the transferable features w.r.t. other domains. We then propose DIFEX for Domain-Invariant Feature EXploration. DIFEX employs a knowledge distillation framework to capture the high-level Fourier phase as the internally-invariant features and learn cross-domain correlation alignment as the mutually-invariant features. We further design an exploration loss to increase the feature diversity for better generalization. Extensive experiments on both time-series and visual benchmarks demonstrate that the proposed DIFEX achieves state-of-the-art performance.

1 Introduction

Over the past years, machine learning, especially deep learning has achieved remarkable success across wide application areas such as computer vision (He et al., 2016) and natural language processing (Vaswani et al., 2017). However, machine learning generally assumes that the training and test datasets are identically and independently distributed (IID) (Esfandiari et al., 2021; Xu et al., 2019), which may not hold in reality. Such non-IID issue is more practical and challenging. For instance, we expect that a model that recognizes the activities of a child can generalize well on the data from an adult, even if their data distributions are different due to different lifestyles and body shapes.

When the target data is available for training (e.g., the adult data is accessible), domain adaptation (DA) (Wilson & Cook, 2020) can be employed to handle the non-IID issue and learn an adapted model for the target domain. However, a more practical situation is when the target domain is *unseen* in training, which makes existing DA approaches not feasible. Domain generalization (DG), or out-of-distribution generalization is one of the most popular research topics that aims to solve such problems (Wang et al., 2021). DG learns a generalized model from multiple training datasets that can generalize well on an unseen dataset. There are several approaches for DG, such as data augmentation (Tobin et al., 2017; Shankar et al., 2018) that increase data diversity and meta-learning (Li et al., 2018a; Balaji et al., 2018) that learns generally transferable knowledge by simulating multiple tasks.

Different from these two approaches, domain-invariant feature learning is a popular DG strategy that aims to learn representations that remain invariant across different domains, thus benefiting cross-domain generalization. For instance, Ganin et al. (2016) designed the domain-adversarial neural network (DANN) using adversarial training, where they tried to confuse the domain classifier such that it could not distinguish which domains the features belonged to, thus achieving domain-invariant learning. Similar to DANN, many

other methods have also been proposed to learn domain-invariant features for DG (Muandet et al., 2013; Li et al., 2018b; Matsuura & Harada, 2020; Sun & Saenko, 2016) and achieved great success.

The popularity and effectiveness of domain-invariant learning naturally motivate us to pursue the rationale behind this kind of approach: what are the domain-invariant features and how to further improve its performance on DG? Previously, for domain adaptation, Zhao et al. (2019) showed that feature alignments across domains are not enough and thereby they paid attention to label functions. However, due to unseen targets in DG, label functions cannot be accessed and Bui et al. (2021) utilized meta-Domain Specific-Domain Invariant to solve the above problem. In this paper, we take a deep analysis of the domain-invariant features for DG. Specifically, we argue that domain-invariant features should be originating from both *internal* and *mutual* sides: the internally-invariant features capture the intrinsic semantic knowledge of the data while the mutually-invariant features learn the cross-domain transferable knowledge. On the one hand, the internally-invariant features are born with the input data that do not change with the existence of other domains. On the other hand, the mutually-invariant features focus on harnessing the cross-domain transferable knowledge that is mined from different distributions. Therefore, the integration of these features can ensure better generalization to unseen domains.

We propose *DIFEX* for Domain-Invariant Feature EXploration of the internally- and mutually-invariant representations. Specifically for internal features, DIFEX employs a knowledge distillation framework to capture the high-level semantics using Fourier transform. For mutual features, DIFEX utilizes correlation alignment to align the feature distributions from any two domains. To allow exploration of these features, DIFEX further adds a regularization to maximize their divergence. We conduct extensive experiments across both image data and time series datasets to comprehensively show the advantage of DIFEX. Results show that our DIFEX outperforms other recent baselines in all datasets.

To sum up, our contributions are mainly three-fold:

1. We propose DIFEX for Domain-Invariant Feature EXploration of both internally- and mutually-invariant representations. We further develop a regularization to maximize their divergence to allow for more feature exploration.[1]
2. Comprehensive experiments on both visual and time series in several different settings demonstrate the superiority and universality of DIFEX.
3. Our method may inspire many similar methods. Exploring more diverse features from both inter- and intra- domain simultaneously may be a promising direction.

2 Related Work

2.1 Domain Generalization

Existing domain generalization approaches (Wang et al., 2021) mainly consist of data augmentation (Shankar et al., 2018; Tobin et al., 2017), domain-invariant feature learning (Muandet et al., 2013; Ganin et al., 2016; Li et al., 2018b), and meta-learning (Li et al., 2018a; Balaji et al., 2018) techniques. Most of existing work pays attention to some specific applications such as computer vision and reinforcement learning. Our work belongs to domain-invariant feature learning-based DG, where we mainly focus on common domain-invariant feature learning methods which can be directly applied to different tasks.

2.2 Domain-invariant Representation Learning

Domain-invariant representation learning (Johansson et al., 2019; Ben-David et al., 2010) has long been a popular solution to addressing domain shift problems such as domain adaptation (Ganin & Lempitsky, 2015; Ganin et al., 2016; Zhao et al., 2019) and domain generalization (Li et al., 2018b; Muandet et al., 2013; Matsuura & Harada, 2020). Specifically for DG, learning domain-invariant representations is critical

[1] To the best of our knowledge, DIFEX is the first work that proposes these two concepts (nternally- and mutually-invariant representations), and utilizes inheritance and exploration simultaneously in DG.

since DG focuses on the invariance across domains. Such domain invariance is achieved can be achieved by explicit feature alignment (Li et al., 2018b; Motiian et al., 2017), adversarial learning (Ganin et al., 2016; Li et al., 2018c; Rahman et al., 2020), or disentanglement (Liu et al., 2020; Peng et al., 2019; Ilse et al., 2020). Recently, some work pointed that only simple alignments maybe not enough for generalization (Bui et al., 2021). Our work shares the same goal with them, but with more possibilities to explore more invariant features that allow more diversities to increase the generalization ability. Our DIFEX may seem similar to the disentanglement framework from many existing efforts since it also learns two kinds of features and then combines them for final prediction. However, it significantly differs from them since we are actually learning two kinds of domain-invariant features, while disentanglement often learns invariant and specific features (Zhao et al., 2019).

2.3 Fourier Features

Recently, some work paid attention to Fourier phase in visual recognition (Yang & Soatto, 2020; Xu et al., 2021). FDA (Yang & Soatto, 2020) introduced the Fourier perspective into domain adaptation. It could generate target-like images for training by simply replacing a small area in the centralized amplitude spectrum of a source image with that of a target image. FACT (Xu et al., 2021) illustrated that Fourier phase information contained high-level semantics and was not easily affected by domain shifts. It utilized the mixup (Zhang et al., 2018) technique to enlarge the diversity of Fourier amplitude without changing labels and thereby learned a model that was insensitive to Fourier amplitude. Another work (Xingchen et al., 2021) utilized Fourier amplitude (style) to calibrate the target domain style on the fly. However, it required some operations when testing. Some other work tried to control Fourier amplitude during training for better generalization (Lin et al., 2022; Zheng et al., 2022). These work are designed for the visual field, although much work (Oppenheim et al., 1979; Oppenheim & Lim, 1981; Hansen & Hess, 2007) have demonstrated that the phase component retains most of the high-level semantics in original signals, which suggests that Fourier phase may be a kind of universal domain-invariant features.

3 Methodology

3.1 Problem Formulation

We follow the problem definition in (Wang et al., 2021) to formulate domain generalization in a C-class classification setting. In domain generalization, multiple labeled source domains, $\mathcal{S} = \{\mathcal{S}^i | i = 1, \cdots, M\}$ are given, where M is the number of sources. $\mathcal{S}^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the ith domain, where n_i denotes the number of instances in $\mathcal{S}^i$. The joint distributions between each pair of domains are different: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq M$. The goal of domain generalization is to learn a robust and generalized predictive function $h : \mathcal{X} \to \mathcal{Y}$ from the M training sources to achieve minimum prediction error on an unseen target domain $\mathcal{S}_{test}$ with unknown joint distribution, i.e., $\min_h \mathbb{E}_{(\mathbf{x},y)\in\mathcal{S}_{test}}[\ell(h(\mathbf{x}), y)]$, where $\mathbb{E}$ is the expectation and $\ell(\cdot,\cdot)$ is the loss function. All domains, including source domains and unseen target domains, have the same input and output spaces, i.e., $\mathcal{X}^1 = \cdots = \mathcal{X}^M = \mathcal{X}^T \in \mathbb{R}^m$, where $\mathcal{X}$ is the m-dimensional instance space, and $\mathcal{Y}^1 = \cdots = \mathcal{Y}^M = \mathcal{Y}^T = \{1, 2, \cdots, C\}$, where $\mathcal{Y}$ is the label space.

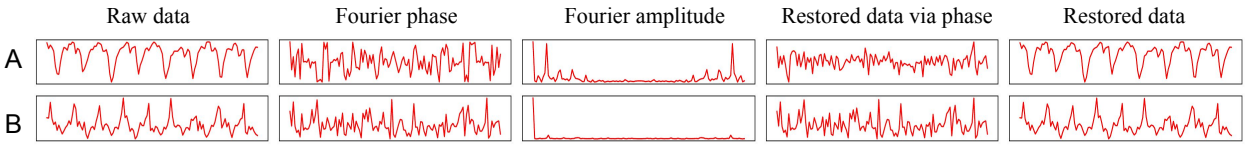


Figure 1: Fourier features as internally-invariant representations.

3.2 Motivation

Existing literature (Yang et al., 2020; Yang & Soatto, 2020; Xu et al., 2021) show that Fourier phase information contains high-level semantics that are not easily affected by domain shifts in the visual field. Many early studies (Oppenheim et al., 1979; Oppenheim & Lim, 1981; Hansen & Hess, 2007) also conclude that in the Fourier spectrum of signals, the phase component retains most of the high-level semantics in the original signals, while the amplitude component mainly contains low-level statistics. Thus, Fourier phase features can act as the internally-invariant features, which, of course, are domain-invariant.

Is Fourier phase sufficient for domain generalization? Figure 1 shows sensor readings of the walking activity collected from two persons in the UCI DSADS dataset (Barshan & Yüksek, 2014).[2] We apply Fourier transform to these two samples and compute their corresponding phase and amplitude. From the results, it is obvious that Fourier phases are similar between two different persons while there exist large gaps in Fourier amplitude. Therefore, Fourier phase information can perfectly act as domain-invariant features to help generalization. We clearly see that the raw data of A and B contains similar periodicity information that could be further utilized for generalization. However, when we restore data with Fourier phases and the sample amplitude, the two restored samples both lose some information compared with raw data such as the periodicity of walking. Thus, Fourier phase along would fail to capture the commonness from multiple training domains since they only focus on their own spectrum without considering the multi-domain statistics. Moreover, some existing work pointed that only simple alignments maybe not enough for generalization (Bui et al., 2021).

3.3 DIFEX

We propose Domain-Invariant Feature EXploration to learn both internally- and mutually-invariant features for domain generalization, short as DIFEX. As shown in Figure 2, DIFEX takes inputs from multiple training domains. Then, after a common feature extractor, we enforce the network to learn the internally- and mutually-invariant features. Subsequently, these features are concatenated to form the invariant features, which can then be used for classification. Furthermore, to allow the network to explore more diversity, we propose the exploration loss to regularize these features by maximizing their divergence.

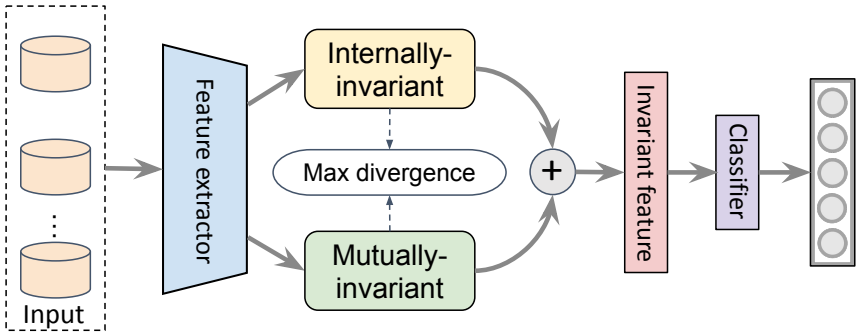


Figure 2: The framework of DIFEX.

3.3.1 A distillation framework to learn internally-invariant features

We show how to obtain internally-invariant features with Fourier phase information. As shown in Figure 3, we utilize a distillation framework to obtain Fourier phase information for classification with raw data. Knowledge distillation is a simple framework to encourage different networks to contain particular characteristics (Hinton et al., 2015; Romero et al., 2014).

Concretely speaking, the teacher network utilizes Fourier phase information and class labels as inputs and outputs, respectively, to obtain Fourier phase information features for classification. According to (Xu et al.,

[2] We do not perform the shift operation, but the schematics have demonstrated that amplitude contains very little information.

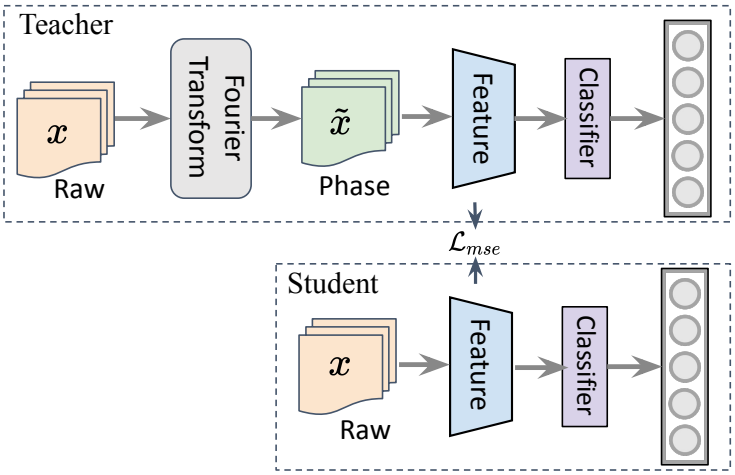


Figure 3: Distillation framework to learn internally-invariant features using Fourier transform.

2021), the Fourier transformation $\mathcal{F}(\mathbf{x})$ for a single-channel two-dimensional data $\mathbf{x}$ is formulated as:

$$\mathcal{F}(\mathbf{x})(u,v) = \sum_{h=1}^{H-1}\sum_{w=0}^{W-1} \mathbf{x}(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)}, \tag{1}$$

where u and v are indices. H and W are the height and the width, respectively. Fourier transformation can be calculated with the FFT algorithm (Nussbaumer, 1981) efficiently. The phase component is then expressed as:

$$\mathcal{P}(x)(u,v) = \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \tag{2}$$

where $R(x)$ and $I(x)$ represent the real and imaginary parts of $\mathcal{F}(x)$, respectively. For data with several channels, the Fourier transformation for each channel is computed independently to obtain the corresponding phase information. We denote Fourier phase of $\mathbf{x}$ as $\tilde{\mathbf{x}}$, then, the teacher network is trained using $(\tilde{\mathbf{x}}, y)$:

$$\min_{\theta_T^f,\theta_T^c} \mathbb{E}_{(\mathbf{x},y)\sim P^{tr}} \mathcal{L}_{cls}(G_T^c(G_T^f(\tilde{\mathbf{x}})), y), \tag{3}$$

where θ_T^f and θ_T^c are the learnable parameters of feature extractor (G_T^f) and classification layer (G_T^c) of the teacher network. P^{tr} is distribution of training data, $\mathbb{E}$ denotes expectation, and $\mathcal{L}_{cls}$ is cross-entropy loss which is a common function for classification.

Once obtaining the teacher network, G_T, we use feature knowledge distillation to guide the student network to learn the Fourier information. Such a distillation is formulated as:

$$\min_{\theta_S^f,\theta_S^c} \mathbb{E}_{(\mathbf{x},y)\sim P^{tr}} \ell_c(G_S^c(G_S^f(\mathbf{x})), y) + \lambda_1 \mathcal{L}_{mse}(G_S^f(\mathbf{x}), G_T^f(\tilde{\mathbf{x}})), \tag{4}$$

where θ_S^f and θ_S^c are the learnable parameters of feature extractor (G_S^f) and classification layer (G_S^c) of the student network. λ_1 is a tradeoff hyperparameter and $\mathcal{L}_{mse}$ is the MSE loss which can make features of the student network close to features of the teacher network.

3.3.2 Explore mutually-invariant features

As discussed early, Fourier phase features alone are insufficient to obtain enough discriminative features for classification. Thus, we explore the mutually-invariant features by leveraging the cross-domain knowledge

contained in multiple training domains.[3] Specifically, given two domains $\mathcal{S}^i, \mathcal{S}^j$, we align their second-order statistics (correlations) using the correlation alignment approach (Sun & Saenko, 2016):

$$\mathcal{L}_{align} = \frac{2}{N \times (N-1)} \sum_{i \neq j}^{N} ||\mathbf{C}^i - \mathbf{C}^j||_F^2, \tag{5}$$

where $\mathbf{C}^i = \frac{1}{n_i - 1}(\mathbf{X}^{i^T}\mathbf{X}^i - \frac{1}{n_i}(\mathbf{1}^T\mathbf{X}^i)^T(\mathbf{1}^T\mathbf{X}^i))$ is the covariance matrix and $||\cdot||_F$ denotes the matrix Frobenius norm.

Since there may exist duplication and redundancy between the internally-invariant features and mutually-invariant features, we expect that two parts can extract different invariant features as much as possible. This allows more diversities in features that are beneficial to generalization. Towards this goal, we regularize the distance between the internally-invariant ($\mathbf{z}_1$) and mutually-invariant ($\mathbf{z}_1$) features by maximizing their divergence, which we call the *exploration loss*:

$$\mathcal{L}_{exp}(\mathbf{z}_1, \mathbf{z}_2) = -d(\mathbf{z}_1, \mathbf{z}_2), \tag{6}$$

where $d(\cdot, \cdot)$ is a distance function and we simply utilize the $L2$ distance for simplicity: $\mathcal{L}_{exp} = -||\mathbf{z}_1 - \mathbf{z}_2||_2^2$.

3.3.3 Summary of DIFEX

To sum up, our method is split into two steps. First, we optimize Eq. 1. Second, we optimize the following objective:

$$\min_{\theta_f, \theta_c} \mathbb{E}_{(\mathbf{x},y)\sim P^{tr}} \mathcal{L}_{cls}(G_c(G_f(\mathbf{x})), y) + \lambda_1 \mathcal{L}_{mse}(\mathbf{z}_1, G_T^f(\tilde{\mathbf{x}})) + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{exp}(\mathbf{z}_1, \mathbf{z}_2), \tag{7}$$

where G_c and G_f are the classification layer and the feature net respectively while θ_c and θ_f are corresponding parameters. λ_1, λ_2, and λ_3 are hyperparameters. Eq. 7 contains four objectives: classification, internally-invariant feature learning, mutually-invariant feature learning, and exploration of diverse features. The first and the third terms are two common objectives in invariant representation learning for domain generalization, and they are not enough according to the existing work Zhao et al. (2019); Bui et al. (2021). With the help of exploration of internally-invariant features and diverse features (the second and the last terms), we can alleviate the above problems to achieve better performance, which is proved in later experiments. For the current implementation, we mainly tune the hyperparameters, which balances these four terms. In the future, we plan to design more heuristic ways to automatically determine the hyperparameters.

As for inference, when a data sample $\mathbf{x}$ comes, we can predict its label by a simple forward-pass:

$$y = \arg\min G_c(G_f(\mathbf{x})). \tag{8}$$

3.4 Discussions

An alternative to learning Fourier features is to directly use the teacher network, and then use another feature extractor to extract mutually-invariant features. However, this would seriously increase the model size for inference as it requires two feature extractors. In addition, it also requires performing FFT for each sample as another input which is time-consuming for inference. Although our distillation framework would introduce more parameters in the training phase, it will not increase the model size and additional inputs for inference, which we believe is more useful in a real deployment.

One may argue that DIFEX may fail if there is only one training domain thus no mutually-invariant features. We provide a positive answer to this situation with virtual domain labels, i.e., domain alignment training on one domain by giving random domain labels. This simple trick seems effective in real experiments.

[3] Mutually-invariant features focus on harnessing the cross-domain transferable knowledge that is mined from different distributions. We think alignments among sources can bring common knowledge of sources.

4 Experiments

We extensively evaluate our method in both visual and time-series domains including image classification, sensor-based human activity recognition, EMG recognition, and single-domain generalization. The training data are randomly split into two parts: 80% for training and 20% for validation. The best model on the validation split is selected to evaluate the target domain. For fairness, we re-implement seven recent strong comparison methods following DomainBed (Gulrajani & Lopez-Paz, 2021): ERM, DANN (Ganin et al., 2016), CORAL (Sun & Saenko, 2016), Mixup (Zhang et al., 2018), GroupDRO (Sagawa et al., 2020), RSC (Huang et al., 2020), ANDMask (Parascandolo et al., 2021), and SWAD (Cha et al., 2021). In addition, for specific applications, we also add some latest comparison methods to compare. Please note that although our method contains two parts of features, we maintain the same architecture to predict for fairness, which means that both the dimension of internally-invariant features and the dimension of the mutually-invariant features are only half of the dimension of the other methods.

4.1 Image classification

Table 1: Accuracy on Digits-DG. The **bold** and underline items are the best and the second-best results.

Source	Target	ERM	DANN	CORAL	Mixup	GroupDRO	RSC	ANDMask	SWAD	DIFEX
MM,SV,SY	M	97.55	97.77	97.62	97.5	97.48	97.78	96.85	97.37	**97.82**
M,SV,SY	MM	55.52	55.62	57.68	**57.95**	53.47	56.27	56	55.77	57.9
M,MM,,SY	SV	59.98	61.85	57.82	54.75	55.63	62.38	59.47	59.00	**64.3**
M,MM,SV	SY	89.25	89.37	90.12	89.8	**92.15**	89.25	88.17	87.92	89.98
AVG	-	75.58	76.15	75.81	75	74.68	76.42	75.12	75.02	**77.5**

Table 2: Accuracy on PACS. The **bold** and underline items are the best and the second-best results.

Source	Target	ERM	DANN	CORAL	Mixup	GroupDRO	RSC	ANDMask	SWAD	DIFEX
C,P,S	A	77	78.71	77.78	79.1	76.03	79.74	76.22	**81.93**	80.86
A,P,S	C	74.53	75.3	77.05	73.46	76.07	76.11	73.81	76.45	**77.6**
A,C,S	P	95.51	94.01	92.63	94.49	91.2	**95.57**	91.56	94.67	**95.57**
A,C,P	S	77.86	77.83	**80.55**	76.71	79.05	76.64	78.06	76.79	79.49
AVG	-	81.22	81.46	82	80.94	80.59	82.01	79.91	82.46	**83.38**

Table 3: Accuracy on VLCS. The **bold** and underline items are the best and the second-best results.

Source	Target	ERM	DANN	CORAL	Mixup	GroupDRO	RSC	ANDMask	SWAD	DIFEX
L,S,V	C	93.64	94.49	96.33	96.18	**97.81**	96.89	93.50	94.56	96.61
C,S,V	L	60.05	64.34	64.42	63.55	62.24	62.91	64.83	61.78	**67.21**
C,L,V	S	68.46	67.15	68.65	68.86	69.23	69.07	63.28	67.82	**74.31**
C,L,S	V	74.02	72.69	70.73	71.95	70.73	70.38	68.87	67.89	**75.24**
AVG	-	74.04	74.67	75.03	75.13	75.00	74.81	72.62	73.01	**78.34**

Datasets and implementation details We adopt three conventional DG benchmarks. (1) **Digits-DG** (Zhou et al., 2020) contains four digit datasets: MNIST (LeCun et al., 1998), MNIST-M (Ganin & Lempitsky, 2015), SVHN (Netzer et al., 2011), SYN (LeCun et al., 1998). The four datasets differ in font style, background, and image quality. Following (Zhou et al., 2020), we utilize their selected 600 images per class per dataset. (2) **PACS** (Li et al., 2017) is an object classification benchmark with four domains (photo, art-painting, cartoon, sketch). There exist large discrepancies in image styles among different domains. Each domain contains seven classes and there are 9,991 images in total. (3) **VLCS** Fang et al. (2013) comprises photographic domains (Caltech101, LabelMe, SUN09, VOC2007). It contains 10,729 examples of 5 classes.

For each dataset, we simply leave one domain as the test domain which is unseen in training while the others as the training domains.

For the architecture, we use ResNet-18 for PACS and VLCS, and Digits-DG dataset uses DTN as the backbone following Liang et al. (2020). Since splits, learning epochs, and some other factors all have influences on results, we mainly compare with the methods extended by ourselves for fairness. The maximum training epoch is set to 120. The initial learning rate for Digits-DG is 0.01 while 0.005 for the other two datasets. The learning rate is decayed by 0.1 twice at the 70% and 90% of the max epoch respectively.

Results The results on three image benchmarks are shown in Table 1, 2, and 3, respectively. Overall, our method has the average improvements of 1.08%, 0.92%, and 3% average accuracy than the second best method on three datasets. Visual classification for domain generalization is a challenging task, and it is difficult to have an improvement over 1%. As can be seen in these tables, the second-best method only has a slight improvement compared to the third one. This demonstrates the great performance of our approach in these datasets. Moreover, we see that alignments across domains and exploiting characteristics of data own can both bring remarkable improvements. In addition, some latest methods such as ANDMask even perform worse than ERM on some benchmarks, indicating that generalizing to unseen domains is really challenging.

4.2 Sensor-based human activity recognition

Datasets and implementation details We also evaluate our method on a cross-dataset DG benchmark with four different datasets on human activity recognition (HAR). The four datasets are: UCI daily and sports dataset (**DSADS**) (Barshan & Yüksek, 2014) consists of 19 activities collected from 8 subjects wearing body-worn sensors on 5 body parts. And it contains about 1140000 samples with three sensors. **USC-HAD** (Zhang & Sawchuk, 2012). USC-SIPI human activity dataset (USC-HAD) composes of 14 subjects (7 male, 7 female, aged from 21 to 49) executing 12 activities with a sensor tied on the front right hip. And it contains 5441000 samples with two sensors. **UCI-HAR** (Anguita et al., 2012). UCI human activity recognition using smartphones data set (UCI-HAR) is collected by 30 subjects performing 6 daily living activities with a waist-mounted smartphone. And it contains 1310000 samples with two sensors. **PAMAP2** (Reiss & Stricker, 2012). PAMAP2 physical activity monitoring dataset (PAMAP2) contains data of 18 different physical activities, performed by 9 subjects wearing 3 sensors. And it has 3850505 samples with three sensors.

To evaluate our method thoroughly for time-series data, we design three settings: Cross-Person, Cross-Position, and Cross-Dataset generalization with these four datasets.:

- **Cross-Person Generalzation.** We split DSADS, USC-HAD, and PAMAP [4] For each dataset, we split data into four parts according to the persons and utilize 0,1,2, and 3 to denote four domains. For example, 14 persons in USC-HAD are divided into four groups, $[1, 11, 2, 0], [6, 3, 9, 5], [7, 13, 8, 10], [4, 12]$ where different numbers denote different persons. We try our best to make each domain have a similar number of data.
- **Cross-Position Generalization.** We choose DSADS since it contains sensors worn on five different positions. Therefore, we split DSADS into five domains according to sensor positions.
- **Cross-Dataset Generalization.** To perform Cross-dataset generalization for HAR, we need to unify inputs and labels of each dataset first. Each dataset corresponds to a different domain. Six common classes are selected. Two sensors from each dataset that belong to the same position are selected and data is down-sampled to ensure the dimension of data same.

The HAR model contains two blocks. Each has one convolution layer, one pool layer, and one batch normalization layer. A single-fully-connected layers layer serves as the classifier. All methods are implemented with PyTorch (Paszke et al., 2019). The maximum training epoch is set to 150. The Adam optimizer with weight decay 5×10^{-4} is used. The learning rate for the rest methods is 10^{-2}. We tune hyperparameters for each method. Moreover, we add another two latest methods, GILE (Qian et al., 2021) and AdaRNN (Du et al., 2021), for comparison in Cross-Person generalization setting.

[4]The baselines for UCI-HAR are good enough, and thereby we do not perform experiments on it.

Results The results are shown in Table 4, 5, and 6 for three settings, respectively. Our method achieves the best average performance compared to the other state-of-the-art methods in all three settings. In the Cross-Person setting, DIFEX has about 2.8%, 5.5%, 1.5% improvement compared to the second-best method for DSADS, USC-HAD, and PAMAP respectively. DIFEX has an improvement with about 2.3% in the Cross-Position setting while it has an improvement with about 7.6% in the Cross-Dataset setting. The results demonstrate DIFEX has a good generalization capability for time-series classification.

Table 4: Classification accuracy on HAR in Cross-Peroson setting. The **bold** items are the best results.

Dataset	DSADS					USC					PAMAP				
Target	0	1	2	3	AVG	0	1	2	3	AVG	0	1	2	3	AVG
ERM	83.11	79.30	87.85	70.96	80.30	80.98	57.75	74.03	65.86	69.66	89.98	78.08	55.77	84.44	77.07
DANN	89.12	84.17	85.92	83.38	85.65	81.22	57.88	76.69	70.72	71.63	82.18	78.08	55.39	87.26	75.73
CORAL	90.96	85.83	86.62	78.16	85.39	78.82	58.93	75.02	53.72	66.62	86.16	77.85	49.00	87.81	75.20
Mixup	89.56	82.19	89.17	**86.89**	86.95	79.98	64.14	74.32	61.28	69.93	89.44	80.30	58.45	87.68	78.97
GroupDRO	91.75	85.92	87.59	78.25	85.88	80.12	55.51	74.69	59.97	67.57	85.22	77.69	56.19	84.95	76.01
RSC	84.91	82.28	86.75	77.72	82.92	81.88	57.94	73.39	65.13	69.59	87.11	76.92	**60.26**	87.84	78.03
ANDMask	85.04	75.79	87.02	77.59	81.36	79.88	55.32	74.47	65.04	68.68	86.74	76.44	43.61	85.56	73.09
SWAD	89.69	76.8	86.93	80.13	83.39	82.4	56.56	71.07	69.08	69.78	85.18	74.93	54.06	81.78	73.99
GILE	81.00	75.00	77.00	66.00	74.75	78.00	62.00	**77.00**	63.00	70.00	83.00	68.00	42.00	76.00	67.50
AdaRNN	80.92	75.48	90.18	75.48	80.52	78.62	55.28	66.89	73.68	68.62	81.64	71.75	45.42	82.71	70.38
DIFEX	**94.30**	**87.24**	**92.15**	85.35	**89.76**	**83.52**	**69.16**	76.45	**79.42**	**77.14**	**90.65**	**82.35**	59.90	**89.25**	**80.54**

Table 5: Classification accuracy on DSADS in Cross-Position setting. The **bold** items are the best results.

Target	0	1	2	3	4	AVG
ERM	41.52	26.73	35.81	21.45	27.28	30.56
DANN	45.45	25.36	38.06	28.89	25.05	32.56
CORAL	33.22	25.18	25.81	22.32	20.64	25.43
Mixup	48.77	**34.19**	37.49	29.50	29.95	35.98
GroupDRO	27.12	26.66	24.34	18.39	24.82	24.27
RSC	46.56	27.37	35.93	27.04	29.82	33.34
ANDMask	47.51	31.06	39.17	30.22	29.90	35.57
SWAD	43.83	30.18	39.35	25.38	27.58	33.26
DIFEX	**49.55**	32.73	**41.75**	**33.43**	**34.20**	**38.33**

Table 6: Accuracy on HAR. The **bold** and underline items are the best and the second-best results.

Source	Target	ERM	DANN	CORAL	Mixup	GroupDRO	RSC	ANDMask	SWAD	DIFEX
USC-HAD,UCI-HAR,PAMAP	DSADS	26.35	29.73	39.46	37.35	**51.41**	33.1	41.66	28.30	46.9
DSADS,UCI-HAR,PAMAP	USC-HAD	29.58	45.33	41.82	47.39	36.74	39.7	33.83	40.99	**49.28**
DSADS,USC-HAD,PAMAP	UCI-HAR	44.44	46.06	39.1	40.24	33.2	45.28	43.22	43.94	**46.44**
DSADS,USC-HAD,UCI-HAR	PAMAP	32.93	43.84	36.61	23.12	33.8	45.94	40.17	16.11	**52.72**
AVG	-	33.32	41.24	39.25	37.03	38.79	41.01	39.72	32.33	**48.83**

Moreover, we have some more insightful findings.

1. *Can the simple alignments always bring benefits?* Obviously, the answer is no. Although CORAL can bring improvements compared to ERM for DSADS in the Cross-Person setting and in the Cross-Dataset setting, it performs worse than ERM for the other benchmarks.
2. *Different alignments bring different effects.* It seems that DANN performs better than CORAL, which inspires us that we can be able to replace CORAL with DANN to capture mutually-invariant features in the future work.
3. *Some methods proposed for visual classification initially also work for time series.* From Table 4, 5, and 6, we can see that RSC brings improvements in most circumstances although it was initially designed for visual classification, which demonstrates there may exist commonness between these

two different modalities. However, compared to remarkable improvements for visual classification, the improvements for time-series data are not significant anymore. It illustrates that these methods may be not general enough.

4. *The same method for different DG benchmarks may have completely different performances.* ERM, the baseline without any generalized techniques, sometimes performs better than other state-of-the-art methods, e.g., for USC in the Cross-Person setting while it performs worse than others sometimes. This phenomenon illustrates that domain generalization is a hard task and we need to design different methods for various situations. In this condition, our method that can achieve the best performance almost on all tasks is inspiring.
5. *For data with fewer raw channels and for more difficult benchmarks, learning more diverse features brings better improvements.* In the Cross-Person setting, USC-HAD is the most difficult benchmark containing only 6 channels. In this situation, how to exploit more diversity and useful features is more important. Thus, our method with both internally-invariant and mutually-invariant features can bring much larger improvements. A similar phenomenon exists in the Cross-Dataset setting.

4.3 EMG gesture recognition

To further prove the superiority of our methods on time-series data, we also validate our method on a more challenging benchmark, EMG for gestures Data Set Lobov et al. (2018) which contains raw EMG data recorded by MYO Thalmic bracelet[5]. Electromyography (EMG) is based on bioelectric signals and is a common type of time-series data used in many fields, such as healthcare and entertainment. For the dataset, the bracelet is equipped with eight sensors equally spaced around the forearm and eight sensors simultaneously acquire myographic signals when 36 subjects performed series of static hand gestures. The dataset contains $40,000-50,000$ recordings in each column with 7 classes and we select 6 common classes for our experiments. We randomly divide 36 subjects into four domains without overlapping and each domain contains data of 9 persons. The baseline methods and implementation details are similar to those for HAR. Experiments are done in three random trials, thus mitigating influences of unfair splits. EMG data comes from bioelectric signals, and thereby it is affected by many factors, such as environments and devices, which means the same person may generate different sensor data even when performing the same activity with the same device at a different time or with the different devices at the same time. Therefore, this benchmark is more challenging.

Table 7: Results on EMG in Cross-Person setting.

Target	0	1	2	3	AVG
ERM	64.08	58.05	52.51	57.29	57.98
DANN	60.68	70.4	64.06	52.39	61.88
CORAL	62.97	73.7	71.53	70.17	69.59
Mixup	59.21	66.54	61.54	64.62	62.98
GroupDRO	63.62	67.64	72.19	67.99	67.86
RSC	65.02	75.19	69.86	61.25	67.83
ANDMask	59.33	66.15	71.83	65.09	65.6
SWAD	42.02	57.06	53.89	48.73	50.42
DIFEX	**71.83**	**82.8**	**76.91**	**75.84**	**76.84**

Table 7 shows that our method achieves the best average performance and is **7.25**% better than the second-best method. Just similar to results mentioned above, alignments among domains can bring improvements in most circumstances while only traditional alignments are not enough, which proves the need of both internally-invariant features and mutually-invariant features. Moreover, from the first task in Table 7, we see that sometimes simple domain alignments may deteriorate performance. But ours performs the best on all tasks.

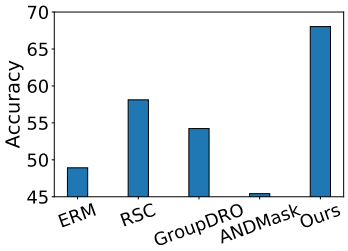


Figure 4: Results on USC-HAD for single-person DG.

4.4 Single domain generalization

In this section, we perform single domain generalization to demonstrate superiority of our method.

USC-HAD We randomly choose two subjects from USC-HAD and each subject serves as one domain. As shown in Figure 4, DIFEX achieves the best performance with over 9% improvements compared to four other state-of-the-art methods. It demonstrates DIFEX can make full use of both internally and mutually domain-invariant features. To sum up, DIFEX is effective in both image and time-series benchmarks with multiple and single domains, indicating that it is a general approach for domain generalization.

[5]For more details about EMG, please refer to `https://archive.ics.uci.edu/ml/datasets/EMG+data+for+gestures`.

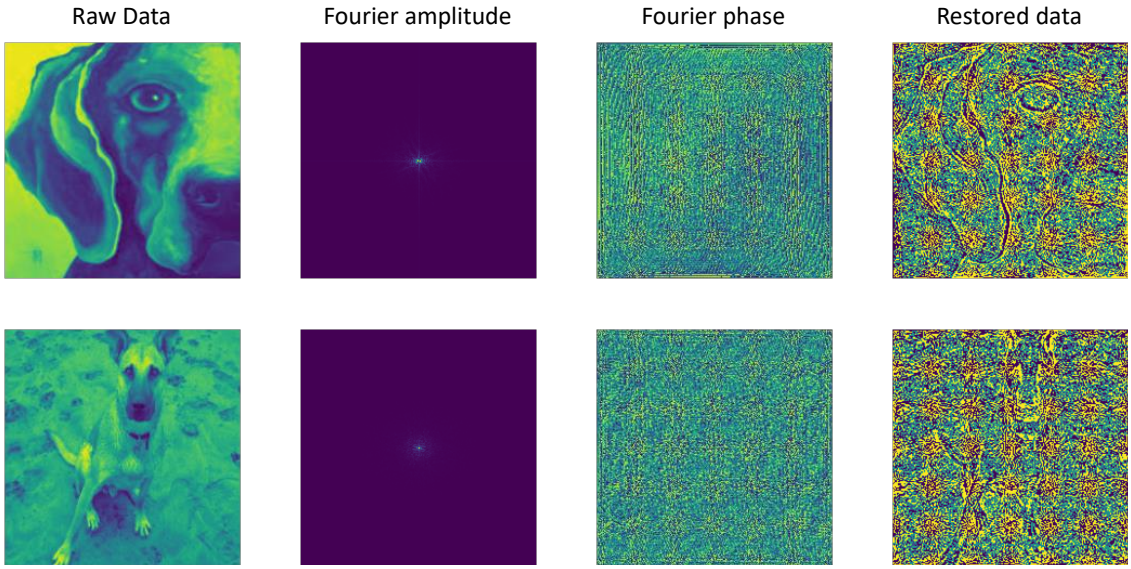


Figure 5: Fourier features as internally-invariant representations for images.

Table 8: Accuracy on PACS. The **bold** and underline items are the best and the second-best results.

Source	Target	ERM	Mixup	GroupDRO	RSC	ANDMask	DIFEX
S	A	35.5	36.67	38.53	36.57	35.5	**46.68**
A	C	57.25	58.49	59.94	62.5	57.25	**64.46**
C	P	81.86	82.87	84.55	84.49	81.86	**86.17**
P	S	31.48	32.25	34.11	33.16	31.48	**56.81**
AVG	-	51.52	52.57	54.28	54.18	51.52	**63.53**

PACS We choose two datasets from PACS each time. We treat one selected dataset as the source domain and the other as the target. As shown in Table 8, our method achieves the best performance with over 9% improvements on average compared to five latest state-of-the-art methods. In this situation, GroupDRO shows its ability of generalization since it is designed for out-of-distribution without many source domains while RSC also shows an acceptable performance since it is a general method. The results demonstrate our method can perform well on single domain generalization for image classification.

4.5 Analysis

Ablation Study We perform ablation study in this section.

Does it work if using internally-invariant or mutually-invariant features alone? As it is shown in Figure 6, directly utilizing mutually-invariant features brings a slight improvement compared to ERM while only utilizing internally-invariant features distilled from a teacher net has the same performance as ERM. Combining both two kind of features bring a much better performance. This demonstrates that only internally-invariant or mutually-invariant features may be not discriminative enough.

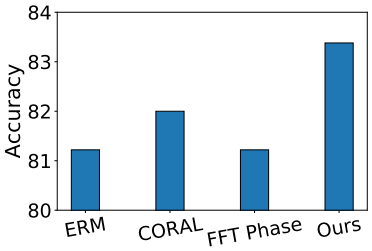
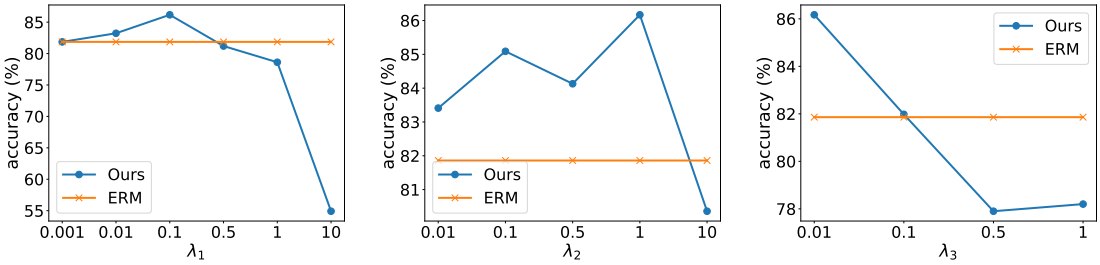


Figure 6: Ablation study on the PACS dataset.

Do three parts of Eq. 7 all work? As shown in Table 9, without one part still has an improvement compared to ERM. Compared to mutually-invariant features and exploration, internally-invariant features play a slightly less important role. And combining three parts can bring the best performance, showing that each part is essential for generalization.

Table 9: Ablation study.

Benchmark	PACS with multiple sources					HAR in Cross-Dataset					PACS with single source				
Target	A	C	P	S	AVG	D	U	H	P	AVG	A	C	P	S	AVG
ERM	77	74.53	95.51	77.86	81.22	26.35	29.58	44.44	32.93	33.32	35.5	57.25	81.86	31.48	51.52
w./o Intern.	80.66	75.51	95.33	78.19	82.42	**46.9**	47.26	45.76	51.35	47.82	41.75	63.52	84.91	50.42	60.15
w./o Mutual.	80.47	75.6	95.33	78.06	82.36	42.61	44.12	45.12	46.49	44.58	43.16	63.01	84.91	41.77	58.21
w./o Exp.	**80.86**	75.85	95.33	78.14	82.55	38.21	48.63	45.88	51.25	45.99	43.02	63.99	85.75	44.9	59.42
DIFEX	**80.86**	**77.6**	**95.57**	**79.49**	**83.38**	**46.9**	**49.28**	**46.44**	**52.72**	**48.83**	**46.68**	**64.46**	**86.17**	**56.81**	**63.53**



(a) λ_1 (b) λ_2 (c) λ_3

Figure 7: Parameter sensitivity (PACS for single domain generalization).

Motivation Examples for Visual Images We have shown an example that some Fourier features can be viewed as internally-invariant representations for time-series data, and we give an example for visual data. As shown in Figure 5, Fourier features for images are similar to features for sensors. Two figures in the second column demonstrate that Fourier amplitude may be useless for classification. Two figures in the last column are data restored with only Fourier phase. They illustrate that Fourier phase contains most of classification information since we can vaguely see dogs while it still loses some information compared to the original data since the dogs are not clear compared to original raw data (two figures in the first column). [6]

Parameter Sensitivity There are mainly three hyperparameters in our method: λ_1 for distillation knowledge from a teacher net with Fourier phase information, λ_2 for the CORAL alignment loss, and λ_3 for the exploration loss. We evaluate the parameter sensitivity of our method in Figure 7 where we change one parameter and fix the other to record the results. From these results, we can see that our method achieves better performance near the optimal point, demonstrating that our method is insensitive to some hyperparameter choices to some extent. We also note that λ_3 for exploration loss is a bit sensitive which may need attention in real applications.[7]

5 Conclusion

In this paper, we propose DIFEX to learn both internally-invariant and mutually-invariant features for domain generalization. DIFEX utilizes an inheritance and exploration framework to combine internally-invariant features distilled from a teacher net with Fourier phase information and mutually-invariant features obtained from domain alignments. Extensive experiments on both image and time-series classification demonstrate the superiority of DIFEX.

In the future, we plan to combine more internally-invariant features and mutually-invariant features for better generalization. Moreover, we may utilize normalization techniques or some other distance to guide explorations for more stable optimization.

[6] Although restored data in Fig.6 have many checkboard artifacts, they have little influence on performance. We first utilize a teacher network to learn useful information for classification. And then we extract internally-invariant representations with the help of the teacher net which has filtered noisy information. In this step, we also utilize the classification loss to reduce the influence of useless information.

[7] When λ_3 is smaller, e.g. $\lambda = 0$, the performance will drop, as shown in Table 9.

References

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International workshop on ambient assisted living*, pp. 216–223. Springer, 2012.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31:998–1008, 2018.

Billur Barshan and Murat Cihan Yüksek. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11):1649–1667, 2014.

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

Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, and Chongjun Wang. Adarnn: Adaptive learning and forecasting of time series. In *ACM International Conference on Knowledge and Information Management (CIKM)*, 2021.

Yasaman Esfandiari, Sin Yong Tan, Zhanhong Jiang, Aditya Balu, Ethan Herron, Chinmay Hegde, and Soumik Sarkar. Cross-gradient aggregation for decentralized learning from non-iid data. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3036–3046. PMLR, 2021.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

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

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

Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24(7):1873–1885, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 124–140. Springer, 2020.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348. PMLR, 2020.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536. PMLR, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

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

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018c.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, et al. Deep frequency filtering for domain generalization. *arXiv preprint arXiv:2203.12198*, 2022.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *arXiv preprint arXiv:2011.01681*, 2020.

Sergey Lobov, Nadia Krilova, Innokentiy Kastalskiy, Victor Kazantsev, and Valeri A Makarov. Latent factors limiting the performance of semg-interfaces. *Sensors*, 18(4):1122, 2018.

Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11749–11756, 2020.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pp. 80–111. Springer, 1981.

A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pp. 632–637. IEEE, 1979.

Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5): 529–541, 1981.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *ICLR*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. volume 32, pp. 8026–8037, 2019.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pp. 5102–5112. PMLR, 2019.

Hangwei Qian, Sinno Jialin Pan, Chunyan Miao, H Qian, SJ Pan, and C Miao. Latent independent excitation for generalizable sensor-based cross-person activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11921–11929, 2021.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124, 2020.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pp. 108–109. IEEE, 2012.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

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

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

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Zhao Xingchen, Liu Chang, Sicilia Anthony, Jae Hwang Seong, and Fu Yun. Test-time fourier style calibration for domain generalization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Hongzuo Xu, Yongjun Wang, Zhiyue Wu, and Yijie Wang. Embedding-based complex feature value coupling learning for detecting outliers in non-iid categorical data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5541–5548, 2019.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14383–14392, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4085–4095, 2020.

Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9011–9020, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Mi Zhang and Alexander A Sawchuk. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 1036–1043, 2012.

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

Kecheng Zheng, Yang Cao, Kai Zhu, Ruijing Zhao, and Zheng-Jun Zha. Famlp: A frequency-aware mlp-like architecture for domain generalization. *arXiv preprint arXiv:2203.12893*, 2022.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13025–13032, 2020.

A More details about DIFEX

A.1 Compare to other Fourier-based methods, e.g. FACT

There are several existing approaches exploiting Fourier-based features for DG, as discussed in related work section. More specifically, although FACT Xu et al. (2021) and ours share similar ideas, we are totally different in methodology:

- FACT utilizes amplitude Mixup to perform data augmentation and its teacher-student framework is for consistency alignment. It can be seen as an adaptation of the data augmentation technique and the model optimization technique.
- Although our DIFEX also tries to utilize Fourier phase information, we focus on the feature extraction. And our teacher network directly extracts the Fourier phase information for the student network to learn.
- Some other methods utilizing Fourier for domain generalization are mentioned in Section 2.3 of the main paper. However, none of them directly utilizes the Fourier phase to guide the part of feature learning (internally invariant features).

A.2 How the mutually-invariant and the internally-invariant features are extracted.

As shown in Figure 2, this operation is done by directly splitting the last layer of the feature network into two parts: one for mutually-invariant and the other for internally-invariant features, respectively. This can be easily implemented by the neural network design. Then, we utilize the learning goals (i.e. distillation, alignment, and exploration) to guide the feature extraction process. The feature extractor network, except for the last layer, is shared by these two parts (the blue block in the figure). The summation operation (i.e., the circled '+' sign) in the figure denotes concatenating two parts. We believe that utilizing independent networks for these two parts respectively can bring better performance. However, it may be unfair since it utilizes a larger feature network.

B Implementation details

We try our best to ensure fair comparisons. As mentioned in the main paper, we utilize the same backbone, the same optimizer, the same splits, and some other same tricks, e.g. data transformation (The epoch is so large that each method can reach convergence.). We tune hyperparameters for each method and select the best to compare (we even tune some base hyperparameters, e.g. lr, for some methods but we report the mainstream of the base hyperparameters.).

Moreover, note that DomainBed is a popular benchmark but not the only one. Domainbed utilizes Adam and totally random parameters, which are not suitable for some methods. Therefore, lots of methods, e.g. FACT, do not utilize it. In addition, utilizing the results reported in the original paper are also not suitable. Different tricks, different data splits, and even different initial states can have large influences on the final results (As shown in DomainBed, in the totally same and random environment, some methods decline seriously.). We try to get a balance between totally random and good performance. We believe ours can achieve better performances with more careful tuning. But we focus on methods and fairness.

C Limitations

Table 10: Classification accuracy with different exploration distance computation ways.

Dataset	Target	A	C	P	S	AVG	
PACS	$L2$	80.86	77.60	95.57	79.49	83.38	
	norm-$L1$	81.30	78.41	96.17	80.02	83.98	
Dataset	Target	0	1	2	3	AVG	
PAMAP	$L2$	90.65	82.35	59.90	89.25	80.54	
	norm-$L1$	90.85	82.51	63.71	89.32	81.60	
Dataset	Target	DSADS	USC-HAD	UCI-HAR	PAMAP	AVG	
Cross-Dataset	$L2$	46.90	49.28	46.44	52.72	48.83	
	norm-$L1$	56.25	49.36	46.35	56.43	52.10	
Dataset	Target	0	1	2	3	4	AVG
Cross-Position	$L2$	49.55	32.73	41.75	33.43	34.20	38.33
	norm-$L1$	50.87	33.85	43.49	34.71	33.46	39.28

We have mentioned in the main paper that exploring with $L2$ distance may bring a difficulty to optimize ($L2$ can be infinity). Therefore, we try another computation way, a normalization distance (norm-$L1$),

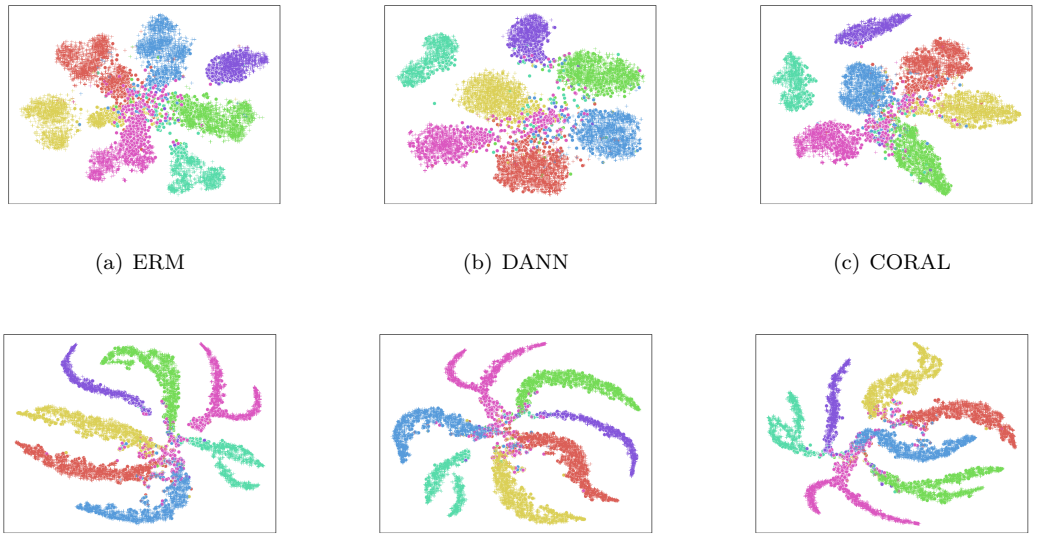
$$L_{exp1}\mathbf{z_1}, \mathbf{z_2} = -||\frac{\mathbf{z_1}}{||\mathbf{z_1}||_2} - \frac{\mathbf{z_2}}{||\mathbf{z_2}||_2}||_1. \tag{9}$$

The results are shown in Table 10 and we can see that norm-$L1$ can bring improvements, especially for difficult problems, e.g. Cross-Dataset[8]. In the future, we may try some other distance, such as cosine.

D Visualization study

We present some visualizations to show the reasons behind the rationales of our method. As shown in Figure 8, we perform t-SNE to show the learned features of training data and target data directly. Figure 8(a) demonstrates that ERM without any generalization operations leads that the learned features of test data have a different distribution from the learned features of training data and target data are hard to classify.

[8]Please note that our method with $L2$ has achieved best results compared to other methods and normalization techniques are mainly for better performance and easier optimization



(a) ERM (b) DANN (c) CORAL

(d) Ours (e) Internally-invariant features (f) Mutually-invariant features

Figure 8: Visualization of the t-SNE embeddings of learned feature spaces for PACS with different methods. Different colors correspond to different classes and different shapes correspond to different domains. Circle and plus correspond to test data and training data respectively. *Best viewed in color and zoom in.*

Figure 8(b) and Figure 8(c) illustrate that learning with alignments can alleviate the above problem but some classes still suffer from the same issue, especially for CORAL. Since mutually-invariant features are learned with the same technique as CORAL, a similar phenomenon exists in Figure 8(f), which again emphasizes the necessity of the other techniques. And Figure 8(e) proves that internally-invariant features may make sense for generalization. Figure 8(d) demonstrates that our methods can further alleviate distribution shifts compared to traditional methods with domain alignments, and thereby it can bring better performance. Moreover, with the last fully-connect layer, we can assign different weights to internally-invariant features and mutually-invariant features for better performance in different situations.